



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN
PHILOSOPHY, SCIENCE, COGNITION, AND SEMIOTICS (PSCS)

Ciclo 38

Settore Concorsuale: 01/B1 - INFORMATICA

Settore Scientifico Disciplinare: INF/01 - INFORMATICA

INTERWORLDS: STUDYING METAPHOR USING NEUROSymbOLIC AI

Presentata da: Anna Sofia Lippolis

Coordinatore Dottorato

Luisa Lugli

Supervisore

Andrea Giovanni Nuzzolese

Co-supervisore

Aldo Gangemi

Esame finale anno 2026

Abstract

This thesis advocates for a pattern-centric account of meaning, where shared knowledge patterns are used to guide and constrain large language models (LLMs), and evaluates it on the case of metaphor. In fact, metaphor is a core mechanism of human meaning-making across expert and lay discourse. At the same time, computational systems have long struggled with it, making it an appropriate domain of study, especially considering the peculiar aspects of ambiguity and multimodality.

To establish design principles for this pattern-centric approach to metaphor, we first investigate ontology engineering as a more controlled setting for understanding if, when, and how symbolic methods require LLM flexibility. Therefore, we investigate to what extent can LLM-based methods assist ontology engineering tasks, in a way that keeps human oversight, formal rigor, and reproducibility at the center. Concretely, we introduce a suite of resources and methods for human-in-the-loop ontology engineering, including the Ontogenia generation framework and studies on automated ontology evaluation. To ensure robustness, we also propose CoLLM, a three-test framework for LLM consistency. Taken together, these studies show that heterogeneous, task-specific datasets are essential and that LLM support is helpful yet unstable: results drift across versions, and automated suggestions can bias experts.

These outcomes argue for human-in-the-loop oversight, explicit reliability tracking, and pattern-guided constraints, which motivates our shift to computational metaphor processing, where these design choices are crucial. For this second strand, we introduce two resources, the Medical Metaphor Corpus (MMC), capturing graded metaphoricity across expert and lay medical discourse, and the Balanced Conceptual Metaphor Testing Dataset (BCMTD), a controlled dataset with generic, scientific, and literal contrasts, and we formalize Conceptual Blending mechanisms for metaphorical combination in a Blending Ontology. Building on this ontology, we implement a Logic Augmented Generation (LAG) neurosymbolic pipeline that combines graph-level constraints with LLM flexibility to represent, recognize, and analyze metaphors. Empirically, LLM-only pipelines perform strongly on clear cases yet systematically struggle on borderline metaphors; the proposed LAG-based approach outperforms baselines, improves robustness and interpretability, and preserves the dynamism of generative methods. This work has implications for clinical communication systems that must generate and interpret patient metaphors, educational tools that scaffold conceptual understanding, and any AI application where capturing or producing graded, context-sensitive meaning is essential. Overall, the thesis contributes to: (i) methods for LLM-assisted, pattern-centric ontology engineering, and (ii) methods for the computational operationalization of metaphor, which demonstrate how knowledge patterns can align cognitive and neural capabilities with symbolic structure for meaning-aware AI.

Keywords: Metaphor, Knowledge Engineering, Conceptual Blending Theory, Large Language Models

Contents

Abstract	iii
1 Introduction	1
1.1 Aim of the work	3
1.2 Relevance of the work	4
1.3 Research Questions	5
1.4 Contributions	5
1.4.1 Theoretical contributions.	5
1.4.2 Resources.	6
1.4.3 Methods.	6
1.5 Thesis outline	6
1.6 Summary of publications	7
1.6.1 Peer reviewed papers	7
1.6.2 Papers under review	9
2 Foundations	11
2.1 Semantic Web foundations and knowledge engineering	12
2.1.1 Ontologies and semantic web resources	12
2.1.2 Frames and frame semantics	15
2.1.3 Knowledge patterns in ontology engineering	17
2.2 The nature of metaphor	25
2.2.1 Conceptual and grounded foundations of metaphor	25
2.2.2 Theoretical frameworks	27
2.2.3 Processing metaphorical meaning in natural cognition	41
2.3 The prevalence and necessity of metaphor in science	47
2.3.1 Constitutive and non-constitutive metaphors	48
2.3.2 The relations of scientific metaphors	50
2.3.3 The performative consequences of metaphorical choice	51
2.4 The ultimate test case for computational meaning	56
3 Computational solutions for metaphor processing	59
3.0.1 Metaphor datasets	60
3.0.2 Computational approaches to Conceptual Blending Theory	62
3.0.3 Knowledge representation of metaphor	64
3.0.4 AI for computational metaphor processing	66
3.0.5 Visual metaphor processing	72
3.0.6 Hybrid approaches for computational metaphor processing	73
3.1 The metaphor crisis	75
3.1.1 Methodological issues	76
3.1.2 The metaphors of crisis	84
3.2 AI for metaphor, metaphor for AI	89

4	Methodologies for metaphor pattern extraction and automated knowledge engineering	93
4.1	Traditional methodologies	95
4.1.1	Top-down approaches	96
4.1.2	Bottom-up data-driven approaches	104
4.1.3	Towards the AI era	105
4.2	AI-assisted extraction methodologies	105
4.2.1	Neural methods	105
4.2.2	LLM-enhanced ontology engineering	107
4.2.3	Neuro-Symbolic methods	112
4.2.4	The missing link: systematic LLM integration in ontology engineering	114
4.3	Ontogenia: empirical investigations into LLM-based ontology generation	114
4.3.1	The Ontogenia methodology: a preliminary study	115
4.3.2	A comparative analysis of context engineering techniques	119
4.3.3	Ontology generation using Large Language Models	130
4.4	Reasoning-driven ontology generation for specific domains	131
4.4.1	Domain-specific ontology generation	132
4.4.2	RAG-Based Ontology Extension with ODPs	135
4.4.3	Establishing design principles for pattern-based, domain-specific ontology generation	137
4.5	Ontology evaluation with LLMs	138
4.5.1	Early approaches to ontology evaluation	138
4.5.2	Terminological coverage evaluation	140
4.5.3	Inference and modeling evaluation	142
4.5.4	Assisted ontology evaluation	146
4.6	Defining reproducibility in the era of LLM-aided Knowledge Engineering	156
4.6.1	A matter of terminology	158
4.6.2	The CoLLM Framework	160
4.6.3	Evaluation setup	163
4.6.4	Results	167
4.6.5	Discussion	171
4.6.6	Establishing consistency standards for LLM-aided knowledge engineering	172
4.7	Ethical considerations	173
4.8	Reframing ontology engineering in the LLM era	175
5	Operationalizing metaphor: from phenomenon to method	179
5.1	The need for domain-specific datasets	181
5.1.1	The Medical Metaphor Corpus (MMC)	181
5.1.2	The Balanced Conceptual Metaphor Testing Dataset (BCMTD)	192
5.2	Representing, recognising, and producing metaphor	194
5.2.1	Representing: the Blending Ontology	194
5.3	Computational implementation of the blending property	201
5.3.1	Problem formulation	201
5.3.2	Frame-based extraction of the blending property	202
5.3.3	Embedding-based extraction of the blending property	203
5.3.4	LLM-based extraction of the blending property	209
5.4	Recognising: a LAG approach to computational metaphor processing	212

5.4.1	Motivation	212
5.4.2	Methodology	213
5.4.3	Application and evaluation	216
5.4.4	Toward metaphor generation	226
5.5	Implications for Computational Metaphor Treatment	226
6	Conclusions and future work	229
6.1	Future Work	234
A	Appendix A: ontology generation studies details	237
A.1	Ontogenia methodology details	239
A.2	Domain ontology generation prompt	240
B	Appendix B: CoLLM study details	245
B.1	Summary and details of the CoLLM studies	245
B.2	Results of studies 1, 3, 4, 5	245
B.3	Overall discussion on the studies	245
C	Appendix C: Details of Chapter 5	251
C.1	Prompt of experiment of Section 5.3.4	251
C.2	LAG Method and implementation details	251
C.2.1	Text2AMR2FRED Usage	251
C.2.2	Dataset examples and statistics	251
C.2.3	Prompt designs and information extraction	252
C.3	Evaluation details	252
C.3.1	Inter Annotator Agreement	252
C.3.2	Quality of the enhanced graphs	252
C.3.3	Error analysis	252
	Bibliography	255
	Acknowledgements	293

List of Figures

2.1	Participation ODP.	18
2.2	Conceptual blending depiction of the sentence “This surgeon is a butcher”	31
2.3	Conceptual blending in <i>PopBlends</i>	32
2.4	ECMT analysis.	34
2.5	Apple-core lesion.	45
4.1	The WHOW ontology network.	102
4.2	<i>pArtLink</i> architecture diagram.	108
4.3	Metacognitive prompting for knowledge engineering.	116
4.4	Illustration of Memoryless CQbyCQ (top part) and Ontogenia (bottom part).	122
4.5	Ontology generation settings illustration.	124
4.6	Analysis of an ontology for a given CQ.	125
4.7	Scores for SemanticWebCourse.	127
4.8	Overview of OE-Assist.	149
4.9	Comparison of assisted and unassisted settings in the interface.	151
4.10	Results overview.	153
4.11	The CoLLM framework.	161
4.12	Bar chart for STD (x-axis), Performance Drift (PD) (y-axis), and Correlation of them. All Studies and all experiments combined as one.	170
5.1	Example of KL divergence between source frames and the target frame CRIME	207
5.2	Comparison of metaphorical scores for two metaphorical domains analyzed in the study	210
5.3	Four-layer evaluation framework for metaphor processing.	214
5.4	Proposed LAG-based pipeline for the sentence “Crime has infected communities everywhere”.	215
5.5	Example prompts and extracts of correct and wrong outputs according to manual evaluation for visual metaphor understanding.	217

List of Tables

2.1	Pattern card for the Participation content ontology design pattern. .	19
2.2	Bleakley’s Nine Functions of Medical Metaphors.	54
3.1	Categorization of computational metaphor approaches by cognitive complexity and robustness axes.	92
4.1	User Story Example: Music Ontology for Bell Usage.	100
4.3	Ontometrics results for Ontogenia.	118
4.4	OWL axioms in Ontogenia.	118
4.5	OOPS! Pitfalls for Incremental Ontology generation.	126
4.6	Comparison of the three prompting techniques.	128
4.7	Adequate CQ modelling (%) by Llama and o1 models for the selected stories.	128
4.8	Domain Statistics for Competency Questions.	133
4.9	Count of errors in modelling for each LLM, divided into Easy and Hard competency questions grouped by ontology domain.	134
4.10	Automatic Pattern Retrieval via RAG.	137
4.11	Measurement Results.	141
4.13	Location Contradiction Detection subtest data.	143
4.14	OntoEval and OntoEval-small datasets statistics.	148
4.15	Performance of LLMs in the automatic ontology evaluation.	152
4.17	Examples of LLMs across distinct models and different versions. . . .	162
4.18	Summary of the CoLLM results across five studies	167
4.19	Results of Study 2.	168
4.2	CQ generation systems: overview.	177
4.16	Summary of key terms related to reproducibility and consistency, with definitions and supporting sources.	178
5.1	Primary sources for MMC.	183
5.2	LLM Performance on scientific metaphor detection without weights .	190
5.3	LLM Performance on scientific metaphor detection with weights. . . .	190
5.4	Ambiguity / error typology observed in <i>tie</i> items.	191
5.5	Summary of BCMTD dataset statistics by type, including instances, percentage of metaphorical sentences, samples, examples, and metaphors.	193
5.6	Summary of datasets statistics, including instances, percentage of metaphorical sentences, and samples.	218
5.7	Performance comparison of various methods for metaphor detection on MOH-X and TroFi	220
5.8	Performance metrics for the BCTMD Dataset for metaphor detection. .	220
5.9	Performance metrics for the MCC Dataset for metaphor detection. . . .	220
5.10	Comparison of results on LAG features.	221
5.11	Ablation study on the visual metaphor dataset.	221
5.12	Comparison of LAG for metaphor detection.	222

5.13	Ablation study results for MOH-X, TroFi, BCMTD, and MMC.	222
A.1	Mapping between the Ontogenia methodology and the MP procedure.	238
B.1	Summary of the selected studies used in the evaluation of CoLLM	246
B.2	Summary of the experimental setup for all research studies used in the evaluation of CoLLM.	247
B.3	Results of Study 1	248
B.4	Results of Study 3.	248
B.5	Results of Study 4.	248
B.6	Results of Study 5.	249
B.7	Discussion over observations, evidence, impacts, and recommenda- tions across studies.	250
C.1	Evaluation Form for Graph Representation.	253
C.2	Examples of sentence types with corresponding metaphors used in experiments.	253
C.3	Table about inter-annotator agreement.	253

List of Abbreviations

AI	Artificial Intelligence
AMR	Abstract Meaning Representation
CMT	Conceptual Metaphor Theory
CBT	Conceptual Blending Theory
COT	Chain Of Thought
CQ	Competency Question
DnS	Descriptions and Situations
DMT	Deliberate Metaphor Theory
DMIP	Deliberate Metaphor Identification Procedure
ECMT	Extended Conceptual Metaphor Theory
FE	Frame Element
IRI	Internationalized Resource Identifier
KG	Knowledge Graph
KE	Knowledge Engineering
KP	Knowledge Pattern
LLM	Large Language Model
LOD	Linked Open Data
MIP	Metaphor Identification Procedure
MP	Metacognitive Prompting
MSE	Mean Squared Error
OEM	Ontology Engineering Methodology
OIE	Open Information Extraction
OWL	Ontology Web Language
ODP	Ontology Design Pattern
PI	Philosophical Investigations
RDF	Resource Description Framework
RT	Relevance Theory
SAT	Speech Acts Theory
SKG	Semantic Knowledge Graph
SMT	Structure Mapping Theory
SPV	Selectional Preference Violation
SUS	System Usability Scale
URI	Uniform Resource Identifier
WES	Wikidata Entity Search
WHOW	Water Health Open knowLedge
WHOW-KG	Water Health Open knowLedge-Knowledge Graph
XD	eXtreme Design
XKG	eXtended Knowledge Graph

Chapter 1

Introduction

“I wish to concentrate on the metaphors for I believe they are important and have been very misleading.”

M. MacDonald, *The Philosopher’s Use of Analogy* (1938)

“We speak of one thing being like some other thing, when what we are really craving to do is to describe something that is like nothing on earth.”

V. Nabokov, *Bend Sinister* (1947)

Metaphors shape everyday reasoning; yet, the same power to explain concepts can also mislead. This thesis addresses that dual nature by developing computational methods to operationalize metaphor: to make explicit what it reveals and what it obscures. Indeed, metaphor, as an analogical mapping between partially aligned domains, presents inherent complexities that current research in computer science has not yet captured adequately.

Two methods have so far dominated: symbolic systems and neural systems. The former offer explicit, human-authored semantics; they are interpretable but narrow. Neural systems are trained on large-scale data and achieve broad coverage. Yet they often rely on surface associations and can be unreliable, including state-of-the-art large language models (LLMs). Despite progress in computational approaches to metaphor, limitations persist: performance on metaphor processing tasks remains modest, domain-specific resources are scarce, formal representational support is poor, and the dialogue between theory and implementable models is generally absent.

The central gap is therefore a computational method that combines broad coverage with verifiable structure so that metaphor can be detected, interpreted, and used for inference in different kinds of unlabeled data.

At a fundamental level, the problem of metaphor is a problem of mediating between worlds. This mediating space, which we call the *interworld*, is where metaphorical meaning emerges. This thesis develops computational methods to operationalize that space and extract implicit meaning.

These challenges matter because, as R. Hoffmann observed in 1995, metaphorical reasoning sits at the center of AI, touching knowledge acquisition, meaning representation, and creative inference (Hoffman, 1995). Metaphor’s monstrous nature stems precisely from its incompatibility with finite representational schemes: it requires matching fundamentally different concepts in ways that produce emergent conceptual content that neither domain possessed independently. Until the scope of computational systems is broadened—until AI can effectively accommodate structured incompatibilities under explicit constraints—it might not fully deal with the fundamental mysteries of cognition central to robust definitions of AI (*ibid.*). U. Eco, too, claimed no algorithm could exist for metaphor, as metaphor effectiveness cannot depend on fixed procedures (Eco, 1986).

By resisting rule-bound formalization, metaphor is also a source of knowledge.

It functions as an epistemic device: it projects structure from a source domain onto a target so that new relations become salient while others are hidden, creating an entirely new concept. The generative power of metaphor violates the finite vocabulary assumption underlying traditional symbolic methods. On the other hand, generativity alone is insufficient: such mappings must remain systematically constrained to be useful. The openness to generation within structural constraints is what current computational approaches fail to capture. LLMs make this challenge unavoidable: their scale and adaptivity force recognition that meaning is inherently context-sensitive and interpretative.

This act of conceptual creation echoes the very etymology of *metaphorá*, the transport or transference of meaning from one place to another (Aristotle, 1920, 21, 1457b6–9): note indeed that even in contemporary Greek, transport trucks have *metaphorái* written on. The metaphorical transfer of sense happens in the interworld, which we borrow from phenomenology and anthropology and from which the dissertation takes its title. The interworld is a threshold or third space that is neither wholly one’s familiar world nor wholly the foreign other’s, but a meeting ground of both, the place of encounter between the familiar and the foreign (Husserl, 1970; Steinbock, 1995). In this view, the two spheres are interdependent: encountering an alien world is the only way to recognize one’s own homeworld as one perspective among many. While we each inhabit our own perspective, our perceptions intersect in an interworld where they overlap (so that the individual *I* becomes *we* through social existence: see Merleau-Ponty, 1968; Linell, 2009). In this space, as anthropologist V. Turner described, norms are suspended and new possibilities can be explored (Turner, 1969).

Theories of conceptual blending, metaphoric frames, and conceptual interaction all implicitly require such an interworld: a space where incompatible structures can be brought into systematic correspondence and generate emergent meaning. To study this space computationally, in this thesis, Knowledge Engineering (KE) provides the means: a representational layer that supports relational mappings and constraint checks while remaining based on shared conceptualization. We implement this layer through Knowledge Patterns (KPs): reusable formal modules capturing recurrent situations.

In this context, both current issues and possible benefits of metaphor are exemplified by scientific metaphors, on which we will focus as a privileged object of study. In fact, scientific discourse depends on productive metaphors to build concepts and explanations, yet the same mappings can mislead when unintended entailments leak, which makes science the right domain to test computationally. Scientific metaphors document the lifecycle of metaphorical meaning from generative novelty to conventionalized usage, and let us test whether systems merely recognize frozen mappings or can participate in the generative process.

1.1 Aim of the work

The aim of this thesis is to develop a computational framework for metaphor processing that combines the broad coverage of LLMs with the interpretability and consensus-based yet modularized approach of KE approaches.

As a result, this study takes the form of a transdisciplinary investigation into metaphor. It treats metaphor at the intersection of philosophy, cognitive linguistics, knowledge representation, and neurosymbolic AI, resulting in a thesis in the field

of computational philosophy. In that intersection, the thesis contributes to three communities.

For cognitive linguistics and philosophy of language, it offers an operational notion of an intermediate layer that preserves roles and relations while keeping meaning usage sensitive, and it treats metaphor as an epistemic device that can be observed at scale and computationally. For KE, it shows how frames and KPs can serve as a philosophically-motivated bridge between lexicons and ontologies, and how these structures can work with LLMs while being aware of limits and benefits of this collaboration, especially for the tasks of ontology generation, evaluation, and reflections on the reproducibility of LLM-based findings. For AI, it delivers scientific metaphor datasets and a reproducible neurosymbolic workflow in which LLMs propose and graphs verify constraints and type compatibility, together with a four layer, consensus aware evaluation for scientific and multimodal metaphors.

1.2 Relevance of the work

By treating metaphor, especially in scientific discourse, as a structured test of understanding, the work advances AI, KE, and scientific communication at large. Scientific metaphors are ideal testbeds: they must remain computationally tractable while retaining the cognitive richness that shapes discovery and practice, and they carry performative consequences in clinical, explanatory, and policy settings. This dual requirement exposes limits in today's models and motivates methods that join symbolic structure with adaptive generation.

More broadly, the relevance of this thesis shows that since metaphor is a pervasive mechanism for thinking and explaining, especially in science, where it shapes concepts and guides inquiry, treating it computationally is paramount to mirror human abilities. For AI, metaphor forces models to connect abstract patterns to embodied and cultural knowledge, exposing the limits of correlation-based systems. This complexity motivates neurosymbolic designs that use explicit KPs to guide and verify generation; which is how it is possible to move from surface similarity to analogical and creative reasoning. In such a context, scientific metaphors are particularly interesting, because they must remain computationally tractable while retaining the cognitive richness that drives discovery, and because their life cycle from live to conventionalized usage is documented. As a result, it lets us study how frames evolve and when mappings succeed. Understanding these metaphors well is also the path to generating new, context appropriate metaphors that support explanation, education, and invention in science.

The generative dimension matters beyond metaphor. It exposes a fundamental requirement: semantic representations must support productive conceptual combination under structural constraints. Traditional ontologies optimize for stable, pre-defined content; neural models generate freely but unreliably. Scientific practice requires both: the capacity to construct novel conceptual configurations (how else could science advance?) while maintaining systematic coherence (how else could findings be shared and tested?). This work therefore aims to advance toward semantic technologies capable of participating in knowledge creation.

The scope of the work is English generic and scientific prose with selected visual cases. The work does not aim to exhaust all metaphor typologies. Limitations include domain specificity and reliance on available frame and graph coverage. These are addressed through consensus-aware evaluation, error analysis, and an explicit

plan to extend datasets, frames, and patterns across domains and languages in future work.

1.3 Research Questions

The work aims to answer the following Research Questions (RQs):

- **RQ1:** *What is the role of KPs in developing systems that can move beyond fixed definitions to process the flexible and dynamic nature of meaning as found in human thought and language, and metaphors in particular?*
- **RQ2:** *What are the computational requirements to study the phenomenon of metaphors, either generic or scientific?*
- **RQ3:** *How has automation in knowledge engineering evolved with LLMs, and how can LLMs be systematically integrated into a reproducible, human-in-the-loop pipeline for KE automation of metaphor processing?*
- **RQ4:** *How can metaphor be operationalized?*
 - RQ4.1: *Data*. What characteristics of metaphor usage emerge in scientific discourse, and how do current LLMs handle graded metaphoricity under human uncertainty?
 - RQ4.2: *Representation*. Can a formal ontology capture the non-conservative nature of metaphor while aligning with perspectivization in context?
 - RQ4.3: *Operational analysis*. What computational pipeline can reliably recover the blending property across metaphorical text and images?

By addressing these research questions, this thesis aims to provide a principled and reproducible account of how metaphor can be represented, computed, and evaluated in general and scientific discourse.

While the research does not claim to exhaustively address all possible inquiries within these domains, its primary objective is to demonstrate the interconnectedness of representation, computation, and evaluation for metaphor in the LLM era.

1.4 Contributions

This thesis introduces a Blending Ontology for conceptual integration, two resources for evaluation, and a neurosymbolic Logic Augmented Generation (LAG) pipeline adapted to metaphor processing with a consensus-aware, four-layer evaluation; it also develops and studies LLM-assisted knowledge engineering methods including pattern-guided ontology generation, ontology evaluation, and a framework for measuring LLM consistency in KE.

1.4.1 Theoretical contributions.

- Review of metaphor theory, identifying points of agreement and tension.
- Development of a review on the role of LLMs in current knowledge engineering tasks.

1.4.2 Resources.

- **Medical Metaphor Corpus (MMC).** 792 sentences with source provenance across nine channels spanning scholarly articles, specialist handbooks, news, patient-oriented social platforms, and crowdsourcing. Each item has crowd-validated perceived metaphoricity on a 0–7 scale plus a binarized label. Coverage includes 82 metaphor types, 24 target domains, and 38 source domains. The MMC is also available as a graph on Framester (Gangemi et al., 2016).
- **Balanced Conceptual Metaphor Testing Dataset (BCMTD).** 147 sentences, balanced as it contains one-third general conceptual metaphors, one-third scientific conceptual metaphors derived from the MMC, and one-third literal controls, designed to minimize lexical artifacts and to probe domain transfer.
- The **Blending Ontology** that formalizes conceptual integration over frames, with axioms for blendable frames, blending spaces, emergent blended entities, and alignment with Cognitive Perspectivization to support role mapping and constraint checking.

1.4.3 Methods.

- **Logic Augmented Generation for metaphor processing.** Application of a neurosymbolic pipeline to metaphor processing, where LLMs propose candidate extensions of base graphs. Knowledge graphs ensure interpretability and constraint checks while using LLMs as reactors.
- **Metaphor evaluation.** A four-layer protocol for metaphor processing that measures detection, understanding of source and target domains, reasoning via recovery of the blending property, and cross-domain and cross-modal generalization. Item difficulty is modulated by human agreement to reflect practical reliability.
- **LLM-assisted knowledge engineering.**
 - *Ontogenia* for pattern-guided ontology generation using metacognitive prompting and ontology design patterns, reaching quality comparable to novice engineers under CQ verification and expert review.
 - *Ontology evaluation methods with LLMs* for human in the loop ontology evaluation with competency question verification, but also introducing measures for terminological coverage, ontological coverage, and CQ satisfaction; quantifying gains when suggestions are correct and automation bias when they are wrong.
 - *CoLLM* as a framework to assess consistency in LLM-aided KE, with empirical baselines on repeatability, sensitivity to model updates, and model replaceability, informing versioning and reporting standards.

1.5 Thesis outline

After this introduction, the thesis proceeds as follows.

Chapter 2. This chapter motivates knowledge patterns as the bridge between formal semantics and cognitively grounded meaning, using metaphor as the test case. It traces fundamental concepts for the Semantic Web, from ontologies to frame-based

modeling, showing how they capture recurring situations grounded in experience. It describes metaphor as central for meaning because it requires mapping relations across domains and because scientific discourse relies on such mappings. It closes by establishing the philosophical stakes and preparing the reader for later chapters where these patterns are used to analyze and operationalize metaphor.

Chapter 3. This chapter surveys metaphor processing methods and resources and argues for a neurosymbolic solution for metaphor, extracting design requirements that the implementation must meet. It examines computational strands from blending theory to modern LLM approaches and frames the bidirectional link between AI and metaphor, where metaphor both challenges and improves our computational systems. It positions KP-driven neurosymbolic designs as the path forward and specifies how subsequent chapters will meet the requirements in practice. It therefore functions as the methodological hinge between foundations and implementation, defining what must be built and how it will be tested.

Chapter 4. This chapter develops and evaluates LLM-assisted knowledge engineering practices, focusing on pattern-guided generation, human-in-the-loop evaluation, and consistency analysis. It reviews top-down and bottom-up Ontology Engineering (OE) methodologies and then examines AI-assisted pipelines, with a focus on Ontogenia as a method for pattern-guided generation and OE-Assist for human-in-the-loop evaluation. It investigates reliability and reproducibility of LLM-based KE pipelines through the CoLLM framework, reporting empirical findings on repeatability and update robustness, and documenting automation bias in user studies. The chapter consolidates design principles for integrating LLMs as reactive spaces within reproducible ontology workflows.

Chapter 5. This chapter instantiates the requirement set in the previous chapters with data, representation, and computation; then evaluates a neurosymbolic pipeline for metaphor processing. It introduces two resources, the Medical Metaphor Corpus and the Balanced Conceptual Metaphor Testing Dataset, and shows, via confidence-weighted evaluation, that current LLMs handle clear cases but struggle on borderline metaphors. It treats the formal representation of metaphor as a Blending Ontology, inspired from Conceptual Blending Theory, then studies extraction of the blending property through frame-based, embedding-based, and LLM-based methods, motivating a neurosymbolic integration. Building on the results, it presents a Logic Augmented Generation pipeline, and evaluates detection, understanding, reasoning, and generation.

The Conclusion synthesizes contributions, limitations, and implications of knowledge patterns for neurosymbolic AI, and outlines extensions across domains, modalities, and languages.

1.6 Summary of publications

We list the following publications generated during the PhD, which are presented in chronological order and grouped as peer reviewed works (subdivided in: journal, conference, and workshop papers) and manuscripts currently under review.

1.6.1 Peer reviewed papers

1.6.1.1 Journal papers.

- A. S. Lippolis, G. Lodi, and A. G. Nuzzolese. The water health open knowledge graph. *Scientific Data*, 12(1):274, 2025.

- **Contribution:** The author’s contribution to the work consists in the participation in the ontology modeling, the LLM-based evaluation setup, in this thesis both presented in Sections 4.1.1.3 and 4.5.2, and participation in the writing of the paper.

1.6.1.2 Conference papers.

- A. S. Lippolis, M. J. Saeedizade, R. Keskiärrkkä, S. Zuppiroli, M. Ceriani, A. Gangemi, E. Blomqvist, and A. G. Nuzzolese. Ontology generation using large language models. In *European Semantic Web Conference*, pages 321–341. Springer, 2025.
 - **Contribution:** The author’s contribution to the work consists in the participation to the methodology and design of experiments, the execution of experiments on the Ontogenia method, alongside the refinement of the technique. Furthermore, the author participated in the evaluation setup and writing of the paper. This work is presented in this thesis in Section 4.3.2.
- A. S. Lippolis, M. J. Saeedizade, R. Keskiärrkkä, S. Zuppiroli, M. Ceriani, A. Gangemi, E. Blomqvist, and A. G. Nuzzolese. Large Language Models Assisting Ontology Evaluation. In *International Semantic Web Conference*, Springer, 2025.
 - **Contribution:** The author’s contribution to the work consists in the participation to the methodology, design and coordination of experiments, along with the prompt and interface design. Furthermore, the author participated in the evaluation setup and writing of the paper. This work is presented in this thesis in Section 4.5.4.

1.6.1.3 Workshop papers.

- A. S. Lippolis. Discovering and Modeling of Tropes in Scientific Texts. In *Proceedings of the Doctoral Consortium at ISWC 2023, co-located with 22nd International Semantic Web Conference (ISWC 2023)*, vol. 3678, CEUR-WS.org, 2023.
 - **Contribution:** This paper, presented and discussed at the ISWC 2023 Doctoral Consortium, describes an early version of this PhD project, including initial method choices.
- A. S. Lippolis, G. Lodi, and A. G. Nuzzolese. Pattern-based design for modelling an ontology network in the water and health domains. In *Proceedings of the 14th Workshop on Ontology Design and Patterns (WOP) 2023*, pages 19–30. 2023.
 - **Contribution:** The paper presents the application of a pattern-based design for the ontology network constituting the backbone of the WHOW-KG, presented in the WHOW project, and also presented in Chapter 4.1.1.3. The author’s contribution to the work consists in the participation in practical modelling and the writing of the paper.
- A. S. Lippolis, A. Klironomos, D. F. Milon-Flores, H. Zheng, A. Jouglar, E. Norouzi, and A. Hogan. Enhancing entity alignment between Wikidata and ArtGraph using LLMs. In *Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH)*, volume 3540, pages 1–12. 2023.
 - **Contribution:** In this article, a methodology that leverages LLMs for entity linking between a knowledge graph and Wikidata has been outlined, also presented

in Section 4.2.2.1. The author's contribution lies in designing and running experiments on the WES methodology presented in the work, participating in the annotation and evaluation setup, and writing of the paper.

- A. S. Lippolis, M. Ceriani, S. Zuppiroli, and A. G. Nuzzolese. Ontogenia: Ontology generation with metacognitive prompting in large language models. In *European Semantic Web Conference*, pages 259–265. Springer, 2024.
 - **Contribution:** This is the only poster among the contributions. It describes the preliminary study in which we test the Ontogenia method for the task of ontology generation, explained in Section 4.3.1. The author's contribution to the work consists in designing the methodology, running the experimental setup and contributing to the writing of the paper.
- A. S. Lippolis, M. J. Saeedizade, R. Keski-särkkä, S. Zuppiroli, M. Ceriani, A. Gangemi, E. Blomqvist, and A. G. Nuzzolese. Assessing the Capability of Large Language Models for Domain-Specific Ontology Generation. In *Proceedings of the Second Workshop on Evaluation of Language Models in Knowledge Engineering (ELMKE 2025)*, co-located with the 22nd Extended Semantic Web Conference (ESWC 2025). CEUR Workshop Proceedings, CEUR-WS.org, 2025.
 - **Contribution:** The author's contribution to the work consists in the participation to the methodology and design of experiments, the prompt design, and its execution on the DeepSeek LLM. Furthermore, the author participated in the evaluation setup and writing of the paper. This work is presented in this thesis in Section 4.4.1.
- S. Cappa, A. S. Lippolis, S. Zoia. Meanings are like Onions: a Layered Approach to Metaphor Processing. In *Explicit and Implicit Knowledge Extraction Workshop, co-located with FOIS*, 2025.
 - **Contribution:** The author's contribution to the work lies in the coordination and design of the framework, particularly the conceptual combination part, and the writing of the paper. Reflections from this paper are particularly presented in Chapter 2 of this work.

1.6.2 Papers under review

- M. J. Saeedizade, R. Alharbi, H. B. Giglou, A. S. Lippolis, E. Blomqvist, V. Tamma, F. Grasso, T. R. Payne, J. D'Souza, S. Auer, *et al.* A Framework for Assessing LLM Consistency in Knowledge Engineering. In *Semantic Web Journal*, under review. <https://www.semantic-web-journal.net/system/files/swj3865.pdf>
 - **Contribution:** The contribution of the author to this work consists in the organization of the related works section, the methodology definition and the experimental running on the Case Study 2, along with the general participation to the writing of the paper and the writing of Case Study 2-related sections and discussion.
- A. S. Lippolis, A. G. Nuzzolese, A. Gangemi. Enhancing Multimodal Analogical Reasoning with Logic Augmented Generation. In *Information Fusion*, under review. arXiv:2504.11190 (2025)
 - **Contribution:** The author's contribution to the work lies in the definition of the methodology, the creation of the BCTMD dataset, the experiments running, and

the writing of the paper. This work is presented in this thesis in Section 5.4, and the dataset is specifically presented in Section 5.1.2.

- A. S. Lippolis, A. G. Nuzzolese, A. Gangemi. The Medical Metaphors Corpus (MMC). In *Scientific Data*, under review. arXiv:2508.07993 (2025)
 - **Contribution:** The contribution of the author to the work consists in the gathering, annotation design and setup and evaluation of the dataset, along with writing of the paper. This dataset is presented in Section 5.1.1.
- S. Zuppiroli, C. F. Longo, A.S. Lippolis, R. Paolillo, L. Giammei, M. Ceriani, F. Poggi, A. Zinilli, A. G. Nuzzolese. The Belief-Desire-Intention Ontology Design Pattern for modelling mental reality and agency. In *Journal of Web Semantics*, under review (2025).
 - **Contribution:** The contribution of the author to the work consists in the Applications section of the paper, where we demonstrated how the ontology could provide the symbolic backbone necessary to augment a Large Language Model with respect to inference and modeling abilities . This work is presented in Section 4.5.

Chapter 2

Foundations

How can computational systems truly understand meaning when even our most basic expressions resist formal definition? This chapter establishes the foundations for addressing this gap through knowledge patterns (KPs), namely reusable ontological structures that capture both the formal semantics required for computation and the cognitive grounding essential for genuine understanding. The concepts, technologies and research in the field will specifically focus on topics that are relevant to the work presented in this dissertation, and aim to answer the following overarching RQ: *What is the role of KPs in developing systems that can move beyond fixed definitions to process the flexible and dynamic nature of meaning as found in human thought and language, and metaphors in particular?*

The path to meaningful comprehension requires two interconnected developments whose foundations will be examined in detail in this chapter: first, in Section 2.1, we explore how semantic web technologies and knowledge engineering have evolved to represent meaning formally, tracing the development from ontologies to frame-based approaches that mirror human conceptualization. We show how KPs emerged as a solution to the rigidity of traditional ontologies, providing flexible yet structured ways to capture recurrent situations and their conceptual relationships.

Second, we investigate the nature of metaphor as a cognitive phenomenon and then further discuss the relevance of metaphor in science in Section 2.3.

Throughout this chapter, we argue that KPs serve as the crucial, philosophically-grounded mediating structure between formal computational requirements and human cognition. This foundation sets the stage for our subsequent exploration of how these patterns can be exploited to create systems capable of genuine metaphorical understanding.

2.1 Semantic Web foundations and knowledge engineering

This section establishes the theoretical and practical foundations necessary for understanding how meaning is formally represented and computationally manipulated in semantic web technologies. We begin by examining the definition of ontology, tracing the evolution of the concept from philosophical inquiry to definitions in the field of knowledge representation. Building on this foundation, we explore how frame semantics provides a cognitive bridge between natural language understanding and formal ontological structures.

The chapter then demonstrates how these theoretical insights translate into knowledge engineering methodologies through KPs. This progression from abstract conceptualization to concrete implementation illustrates a central theme: the semantic web's success depends on maintaining fidelity to human cognitive processes while enabling computational tractability.

These foundations represent a fundamental approach to knowledge representation that recognizes the inherently social, contextual, and experiential nature of human understanding, thus setting the stage for the more specialized discussions that follow in subsequent chapters.

2.1.1 Ontologies and semantic web resources

Ontology, as the etymology suggests, is a long-standing term in philosophy, referring to the study of being and categories of existence. For the philosopher W.V.O.

Quine, the question: “What exists?” is answered by: “Everything”, namely the entities our best scientific theories must refer to, viewing ontology as a collection of ontological commitments of a theory; specifically, what entities must exist for the theory to be true (Quine et al., 1948).

The study of ontologies in computer science metaphorically borrows the philosophical term to denote an artifact that represents in a formal way what exists in a domain. In this vision, computer scientists do not consider much the ontological commitments but rather the logical forms that can be used for machine-readable knowledge representation. Thus, representation is a condition *sine-qua-non* to exist as an information object (Guarino, Oberle, and Staab, 2009). Indeed, Gruber (1993) defined ontologies as the “explicit specification of a conceptualization”.

Gruber (1993) provides a conceptual refinement of ontology by addressing a fundamental limitation in how exactly conceptualization was previously understood. In the work, it is argued that a conceptualization should not change when the world changes: the meaning of concepts should be captured independently of particular world states, thus they have to be established within a community. This refinement strengthens the theoretical foundation of ontologies by ensuring conceptual stability, which further enables the collective aspect that has been emphasized as essential for interoperability.

In fact, later, Borst (1997) defined an ontology as a “formal specification of a *shared* conceptualization”. The addition of the *shared* aspect in ontology development with respect of the first definitions is crucial, showing the view that consensus between parties is fundamental to solve interoperability issues and grant wide adoption of ontological artifacts¹. Here, it is thus apparent that ontologies can serve as abstract conceptual schemas for data integration, sharing, and discovery.

In this work, the matter of computational shared categorization, that is, how it is created from mirroring human cognitive processes, re-engineered over time, and potentially automated, acquires a prominent role. Here, we adopt Guarino, Oberle, and Staab (2009)’s definition of ontology as a semantic data model that is shared among stakeholders, represents a specific domain or ontological commitment, and serves as a formal transposition of a conceptual theory into machine-readable semantics.

Since ontologies fundamentally concern the articulation of conceptual relationships, they are a fundamental part of the Semantic Web project, which its founder T. Berners-Lee defined as “an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in co-operation” (Berners-Lee, Hendler, and Lassila, 2001): in other words, a new way to create encyclopedic knowledge. It differs fundamentally from traditional encyclopedic projects in several key respects: while classical encyclopedias like the *Encyclopædia Britannica* primarily organize knowledge for human consumption through discrete, static entries, the Semantic Web envisions encyclopedic knowledge as a hypertextual web of machine-readable relationships. This shift emphasizes the change from knowledge as a collection of isolated articles to knowledge as a computational

¹Meaning negotiation as a constitutive concept of a semantic web ontology finds its counterpart in a large body of cognitive science research, and in particular in the field of social metacognition. Recently, Borghi, Mazzuca, and Tummolini (2025) have shown that during abstract language use, for people to reach mutual understanding and common ground with others, they indeed have to negotiate meaning together. This negotiation process is distinct from mere socialness: in social metacognition, the contribution of others is critical to learning concepts and determine the kind of meaning that is undetermined.

resource where facts can be automatically discovered, verified, and combined across domains.

Berners-Lee recognized that realizing this ambitious Web vision demanded the creation of knowledge representation systems that were both decentralized and open. Rather than relying on any single repository to contain complete knowledge, these systems needed to enable the integration of information from multiple sources, with each contributing its own partial understanding of reality. Achieving this required establishing common standards for interoperability. Building upon the foundational RDF data model, the W3C Consortium developed an ecosystem of complementary standards.

The Resource Description Framework (RDF) standard² represents knowledge as triples: statements consisting of subject, predicate, and object. Each element of the triple is identified by an Internationalized Resource Identifier (IRI), a globally unique identifier that extends the Uniform Resource Identifier (URI) protocol concept to allow non-ASCII characters. These triples form a directed labeled graph structure that connects resources in a machine-readable way. Building on this foundation, the Web Ontology Language (OWL), specifically OWL 2³, structures knowledge through three primary components: classes (which represent categories of entities), object properties (which define relationships between entities), and data properties (which link entities to literal values like strings or numbers). It thus allows us to define ontologies with description logics-based semantics for reasoning, enabling the expression of complex class relationships, property characteristics (where domain and range specifications serve as axioms for inference under the Open-World Assumption rather than rigid database constraints), and logical constraints that support automated inference (Hitzler et al., 2009).

Knowledge graphs (KGs) represent the practical manifestation of these semantic web standards in real-world applications. A KG is a structured representation of knowledge that consists of entities (nodes) and their relationships (edges), where entities are connected through semantic relationships that can be processed by machines. Unlike traditional databases that store data in tables and adopt a siloed approach, KGs explicitly model the meaning and context of information through their graph structure. For that, they use abstract ontological specifications for data instantiation. KGs are often built following the principles of Linked Open Data (LOD), which promote the use of shared identifiers, standardized vocabularies, and open publication formats. This allows datasets from different domains to interlink, fostering interoperability and reuse across knowledge systems. Well-known examples include Google’s Knowledge Graph, which powers search results with structured information about entities and their relationships (Singhal, 2012), Wikidata, which provides structured data underlying Wikipedia (Erxleben et al., 2014), and various enterprise knowledge graphs used by organizations to integrate and reason over their internal data sources.

The power of KGs lies in their ability to combine the formal semantics of ontologies with the scalability needed for real-world data integration. They enable what Hogan et al. (2021) describe as “open-world reasoning”, where new facts can be inferred from existing knowledge, and where the graph can continuously evolve as new information becomes available. These capabilities make KGs particularly suited for applications requiring semantic search, question answering, recommendation systems, and automated knowledge discovery. However, the construction and

²<https://www.w3.org/TR/rdf-schema/>

³<https://www.w3.org/TR/owl2-overview/>

maintenance of knowledge graphs present significant challenges. Among these, the knowledge boundary problem, i.e., determining what constitutes meaningful, coherent units of knowledge, becomes particularly relevant when dealing with large-scale, heterogeneous data sources that cannot be closely analyzed by human beings. This is where the connection between formal ontologies and cognitive approaches becomes crucial: for the field of the semantic web, frames have been proposed as the fundamental units of meaning (Gangemi and Presutti, 2010). In the next chapter, we explore their importance and function in this context.

2.1.2 Frames and frame semantics

Frame semantics originates from the need to structure how meaning depends on the knowledge we use to classify experience. In fact, C. Fillmore, the founder of frame semantics, noted how his work could be likened to ethnographic fieldwork: like an anthropologist studying an unfamiliar culture, the researcher asks which categories of experience a speech community encodes through its habitual linguistic choices Fillmore (2006, p.373). Frame semantics is thus the idea, derived from cognitive linguistics, that word meanings are understood via structured scenarios (also called 'schemas', 'scripts', 'ideational scaffolding', 'cognitive models', etc.) that organize knowledge and context. In this way, frames bridge the encyclopedic nature of language and its linguistic expression through prototypical features. As Fillmore (2006, p. 373) says,

"By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available".

This conceptual apparatus has proved attractive well beyond linguistics. The notions of "frame" and "framing" have been used in a range of different fields, including artificial intelligence (Minsky, 1974), semantics (Fillmore, 1975; Fillmore, 2006), and knowledge engineering (Clark, 2008). Indeed, frames did not appear *ex nihilo*; rather, they crystallise a decades-long line of research on structured knowledge representations. F. Bartlett (1932)'s notion of a *schema* captured the Gestalt organisation of memory, emphasising how past experience supplies an active framework that shapes recall. M. Minsky re-engineered Bartlett's insight for machine reasoning: a frame in Artificial Intelligence (AI) is a hierarchically organised slot-filler structure whose slots hold default values that can be overridden in specific contexts, allowing rapid inheritance-based inference (Minsky, 1974). R. Schank and R. Abelson named frames as *scripts*, i.e. ordered sequences of scene frames that model stereotypical everyday events such as "dining out", encoding not only what entities exist but what typically happens next (Schank and Abelson, 1977). T. Winograd's SHRDLU system then showed that procedural frames (procedures bound to slots) could drive real-time language understanding and action planning in a microworld (Winograd, 1977). Cognitive linguistics also reclaimed the frame concept: R. Langacker treats grammatical constructions as pointers into rich encyclopaedic frames, so that syntax always evokes background scenes (Langacker, 1987).

Despite the differences between frame theories, in the context of the Semantic Web, the frame remains "the primary research object over the semantic web" (Gangemi and Presutti, 2010, p. 3), and can be considered as a "primitive" (Nuzzolese et al.,

2011), a portion of background knowledge that connects language analysis with concepts and knowledge. It allows to have fixed but flexible enough structures to enable standardization, reusability, and adaptability. Frames are thus considered units of meaning that necessarily need to be related to a context model involving information, situational, social and formal aspects of the research object.

Frames specify a set of roles, or frame elements (FEs), that identify the participants, props, and relationships making up the prototypical scene; some FEs are core (indispensable for recognizing the frame) while others are peripheral (optional but informative). For instance, in the sentence “Elena bought a bicycle from Luca for €250”, the verb *bought* triggers the *Commercial_Transaction* (Buying) frame: Elena fills the Buyer role, Luca the Seller, the bicycle the Goods, and €250 the Money FE that quantifies the exchange.

As shown in the example, frame evocation in language involves mapping linguistic elements onto structured conceptual roles. However, this surface alignment often hides the deeper connections between explicit linguistic cues and tacit background knowledge. For instance, while “Elena bought a bicycle from Luca for €250” clearly expresses the *Commercial_Transaction* frame, consider a variation like “Elena finally got her dream bike”. Here, the same frame may be implicitly evoked: we infer a Buyer, Goods, and an exchange, yet none of the frame elements are overtly stated.

Moreover, when we consider figurative or metaphorical expressions, the situation becomes even more complex. A sentence like “Elena invested her energy into that friendship” evokes a *Commercial_Transaction* frame in a metaphorical way: emotional commitment is framed as a purchase or investment, with energy as currency and friendship as goods. Frame evocation indeed goes beyond surface syntax and requires interpretation grounded in context, conceptual mapping, and world knowledge.

The centrality of frames for our thought processes has been further confirmed by empirical work. For instance, behavioral studies show that people activate rich, frame-like simulations when verifying simple word properties, e.g. whether a peach has fuzz, supporting the claim that concepts are stored in the brain as structured situational frames rather than as feature lists (Barsalou, 2012). Neuro-imaging studies confirm this: activity patterns in medial prefrontal and posterior-medial networks re-instantiate event schemas while participants encode and recall realistic stories, showing that frames scaffold both comprehension and memory retrieval (Masís-Obando, Norman, and Baldassano, 2022). Indeed, in cognitive science, frames are widely regarded as mental entities that “provide the fundamental representation of knowledge in human cognition” (Barsalou, 2012, p. 21).

Because of the central role of cognition, frame semantics can be used “as a cognitive model for representing the formal meaning of ontology entities” (Asprino et al., 2017, p. 4905). This method abstracts away from surface mismatches in how ontologies are modeled. For example, two ontologies might both contain the concept of *buying*, one as a class *Purchase* event, another as a property *boughtBy* linking buyer and product, which traditional alignment might miss. But if both evoke the *Commercial_Transaction* frame with Buyer, Seller, Goods, Payment roles, we can recognize a correspondence at the frame level. Furthermore, as seen by this and the previous examples, different types of frame evocation (such as literal, implicit, and metaphorical), even when they are present in the same sentence, require distinct strategies for formalization and computational modeling.

In the context of the semantic web, this means frame-based ontologies can link lexical knowledge to formal knowledge. Projects like FRED (Gangemi et al., 2017) follow exactly this aim by extracting frames from text and anchoring them in OWL

ontologies. This machine reader integrates the output of existing NLP tools and LOD resources to expand tacit knowledge. Other online resources that support frame-based analysis are FrameNet, the first frame-based resource present online (Baker, Fillmore, and Lowe, 1998), and Framester (Gangemi et al., 2016), a semantic web resource that building on FrameNet, integrates cognitive and computational semantics. While FrameNet itself is a lexicon, Framester transforms and interlinks FrameNet with ontologies and knowledge graphs while providing a rigorous formal treatment for Fillmore's frame semantic.

Framester is in fact described as a frame-centric resource that acts as a hub between several other linguistic resources that are already part of the LOD cloud (Gangemi et al., 2016). In practice, Framester links the FrameNet lexicon of frames to ontological counterparts, integrating resources like WordNet (Fellbaum, 2010), DBpedia (Auer et al., 2007), and the DOLCE foundational ontology (Gangemi et al., 2002). By providing formal semantics to frames in a curated linked data version of multiple linguistic resources, Framester enables full OWL reasoning over frame representations of events and their participants. This approach significantly extends traditional ontology definitions, which might not inherently contain such rich event representations or link to natural language meaning.

In practical ontology engineering, the notion of a frame is generalized and re-emerges as an Ontology Design Pattern (ODP) or, more generally, a KP (Gangemi, 2020), a topic we assess in the next chapter.

2.1.3 Knowledge patterns in ontology engineering

Design patterns entered knowledge representation along the same intellectual path that led from C. Alexander (1977)'s "pattern language" in architecture to Minsky (1974)'s "frames" in Artificial Intelligence and, finally, to ODPs, which can be considered specialized types of KPs (Gangemi and Presutti, 2009). Extending ideas by the W3C Semantic Web Best Practices and Deployment Working Group, Gangemi (2005) and Blomqvist and Sandkuhl (2005) introduced ODPs as recurrent and stable but flexible shared elements of experience, in a way similar to frames, but applicable to ontology engineering in a typically top-down manner (See Section 4.1.1).

ODPs are thus defined as reusable, small-scale ontologies that capture a recurring modelling solution. The intent is not only that the use of such ODPs, and appropriate support tooling, will guide ontology engineers (whether experienced or novices) and support them in developing high quality ontologies with greater ease and confidence, but also that treating what are in other fields called "frames" as "patterns" lets ontology engineers work with cognitively grounded building blocks.

More broadly, KPs consist in all the types of knowledge, including ODPs, that can be reorganized as a pattern and reused in top-down and bottom-up approaches (see Section 4.1 for a more detailed account of such approaches). Nuzzolese et al. (2011) define a KP as "a small, connected unit of meaning" (p. 520): a formalised frame consisting of a handful of key classes, relations, and axioms that together describe a prototypical situation.

Building on this "pattern science" vision, the semantic web community converges on three requirements. KPs need to be:

1. **Task-based (competency-question (CQ) oriented).** A KP must answer the competency questions (natural language questions outlining and constraining the scope of an ontology) that motivated its creation, e.g. Which researcher authored which paper? (Grüninger and Fox, 1995).

2. **Well-grounded (data- and text-anchored).** The pattern should align with popular linked-data vocabularies and be traceable to textual or data evidence, ensuring that it scales and remains verifiable.
3. **Cognitively sound.** Patterns that mirror common-sense conceptualisations are easier to learn, explain, and reuse, therefore they should preserve the cognitive roots of frames.

As shown from these requirements, a KP exposes several façades: vocabulary layer, formal axioms, inference rules, SPARQL templates, usage examples, and so on, so that users can adopt only what they need (Gangemi and Presutti, 2009). Within the ODP portal⁴, patterns are grouped into six families (Structural, Correspondence, Content, Reasoning, Presentation, Lexico-syntactic); KPs typically belong to the Content family but often combine Structural and Reasoning motifs for completeness. Content ODPs address modeling challenges related to ontology content, whether in general-purpose or domain-specific contexts. They offer reusable solutions to problems that are either difficult to model accurately or commonly encountered, where a shared conceptual structure can enhance consistency and interoperability. Throughout this work, when talking about ODPs, we will refer to the Content type.

Among Content ODPs, classic examples include the AgentRole pattern (linking an agent, the role it plays, and the time interval)⁵ and the Participation pattern (connecting events, participants, and their functions)⁶. For instance, the Participation ODP as shown in Figure 2.1 and whose information is represented in Table 2.1 models that an object takes part in an event. Apart from the name and aim, the ODP answers to specific requirements, such as competency questions and user stories (where one represents a usage scenario as requirement for the ODP), and can be related to other ODPs. Additional information that a published ODP could have are reviews, modeling issues and references.

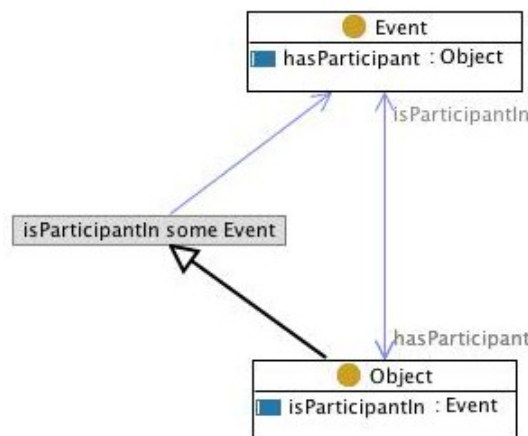


FIGURE 2.1: Participation ODP.

A foundational ODP is Descriptions & Situations (D&S)⁷. Introduced by Gangemi and Mika (2003a), it separates (i) a Description (the conceptualization: norms, plans,

⁴<http://www.ontologydesignpatterns.org/> and <https://github.com/odpa/patterns-repository/tree/master>

⁵<https://github.com/odpa/patterns-repository/tree/master/AgentRole>

⁶<https://github.com/odpa/patterns-repository/tree/master/Participation>

⁷<https://github.com/odpa/patterns-repository/tree/master/DescriptionAndSituation>.

Name	Participation
Intent	Model that an Object participates in an Event.
Competency questions	Which Objects participate in a given Event? Which Events does a given Object participate in?
Solution description	A minimal pattern capturing a simple binary relation between Object and Event via the properties <code>hasParticipant</code> and its inverse <code>isParticipantIn</code> .
Core elements	Classes: Event, Object. Object properties: <code>hasParticipant</code> , <code>isParticipantIn</code> .
Reusable OWL building block	http://ontologydesignpatterns.org/cp/owl/participation.owl
Consequences	Suitable for generic object–event participation. Cardinalities can be restricted in application-specific ontologies. This pattern is non-temporal; for time-scoped relations, use the Time-indexed Participation pattern.
Scenario	<i>Aldo Gangemi participated in the premiere of La Dolce Vita.</i>
Related CPs	Time-indexed Participation

TABLE 2.1: Pattern card for the Participation content ontology design pattern.

diagnoses, frames) from (ii) a Situation (a relational context of concrete entities that satisfies that description). More specifically, it addresses limitations of plain RDF/OWL triples in representing contextualized, n -ary facts and context-dependent roles. Rather than asserting timeless types (e.g., Alice is a Patient), D&S lets us say that an entity plays a role in a specific context. This separation enables (i) principled reification of n -ary relations (who–did–what–when–where–according to which rule), (ii) time-indexed and revisable classifications, and (iii) multi-perspective modeling, where the same constellation of entities can satisfy different Descriptions (e.g., clinical vs. administrative views) without conflation.

For example, a Diagnosis Description defines `PatientRole`, `ClinicianRole`, and parameters such as `Condition` and `Evidence`. A concrete outpatient visit is a Situation that `dul:satisfies` that Description; it `dul:isSettingFor` Dr. Rossi (playing `ClinicianRole`), Alice (playing `PatientRole`), the X-ray (`Evidence`), and the ICD-10 code (`Condition`). The role assignment is expressed by `dul:classifies` (optionally via a `dul:Classification` sub-situation to attach time), so that “Alice is a patient” holds *in that visit*, not globally.

2.1.3.1 Benefits of pattern-based knowledge engineering

Gangemi and Presutti (2009) list several benefits of this pattern-based approach for knowledge engineering practices: simplified ontology design, advanced data exploration and extraction, lens-based views over data, pattern-driven querying, and deep linguistic grounding. In fact, an experiment showed that modellers who

reused ODPs in their ontology development methodology produced correct ontologies significantly faster than a control group (Hammar, 2014). In a separate industrial case-study, refactoring an ontology around ODPs aligned it with common Linked-Data vocabularies and greatly increased semantic interoperability (Hammar, Lin, and Tarasov, 2010). Beyond manual reuse, patterns can also drive knowledge-engineering automation: Blomqvist (2009b) describes a semi-automatic pipeline that selects and instantiates ODPs to generate large parts of an ontology, cutting handwritten axiom effort and improving design quality. Beyond these empirical studies, the efficacy of KPs is evident from the fact that they have been used for more than a decade in ontology engineering (see, among the most recent examples, how various are the domains of application: from robotics to cultural heritage, from law to water and health: Lippolis, Lodi, and Nuzzolese, 2025; Nuzzolese et al., 2025; Gayathri et al., 2025). The use of patterns from this large body of research shows that by reusing well-tested patterns, the knowledge engineer imports a plug-and-play mini-theory with built-in competency questions, reasoning support, and lexical alignments, all of which have been shown to accelerate design, improve interoperability, and improve knowledge engineering automation outcomes.

KP engineering can proceed along two complementary routes. Pattern construction (often realised through ontology engineering methodologies such as the *eXtreme Design* methodology: Presutti et al. (2009), see Section 4.1.1.3) starts with explicit requirements, selects one or more appropriate ODPs from existing libraries, then renames, specialises, and composes the patterns until the questions are answerable, including formal axioms, SPARQL templates, and provenance metadata (Gangemi and Presutti, 2010). Repositories to draw from, for example, can be FrameNet (Baker, Fillmore, and Lowe, 1998), the ODP portal (Gangemi, 2005), and similar resources that provide expert-crafted patterns. To achieve the goal of answering the requirements, KPs retrieved through a top-down strategy usually require re-engineering, e.g. the transformation of FrameNet into Linked Open Data and a library of OWL 2 patterns (Presutti et al., 2016). For a more detailed account of this process, see Chapter 4.

By contrast, pattern induction is a bottom-up, data-driven workflow: recurrent graph motifs are mined from large linked-data corpora (e.g. DBpedia, Wikidata), generalised into concise type–property skeletons, “lifted” to domain-independent roles, and pruned against quality metrics to yield so-called Empirical ODPs (Nuzzolese, 2014). This construction process is needed when having to manage vast heterogeneous datasets. In practice, projects can blend the top-down and bottom-up approaches: using induced patterns to sketch the data landscape, then refining or replacing them with ontologies as competency questions arise. For this acquisition methodology, linked-data corpora and semi-structured artefacts can be mined automatically; the main challenge is filtering noise and identifying patterns into clean OWL templates (Presutti et al., 2016).

Pattern-centric tools such as FRED (Gangemi et al., 2017), Tipalo (Nuzzolese et al., 2013), and Aemoo (Nuzzolese et al., 2016) show that KPs can drive automatic text-to-OWL conversion, exploratory search, and question answering. Typical evaluation axes in this field include boundary precision to measure how well the pattern isolates the intended concept, reasoning gain, and usability scores.

Thanks to the multiple benefits they bring about, KPs are thus still very much used today in ontology engineering. However, it is important to note KPs are not just a computational resource; as a bridge between data and human concepts, they are philosophically motivated. In the next section, we explore how bridging structure and meaning has been at the center of philosophical investigation.

2.1.3.2 Meaning in-use: A Wittgensteinian Perspective

Traditional knowledge representation approaches often assume that concepts possess fixed, definable essences that can be captured independently of their contexts of use (see §2.1.1). Under this view, the meaning of a concept like “payment” or “document” should be specifiable through necessary and sufficient conditions, creating stable foundations for formal ontologies. However, this essentialist approach faces a fundamental challenge: concepts systematically resist such fixed definitions. When new needs emerge, e.g., with credit cards as novel forms of payment or digital files as new types of documents, our conceptual systems adapt fluidly, suggesting that meaning is not anchored in static properties but emerges through dynamic patterns of use.

A concrete example of this phenomenon is shown by Piantadosi and Hill (2022): people call something a “postage stamp” if they pay for it and then attach it to a letter to facilitate delivery. The meaning of “postage stamp” is intrinsically intertwined with other concepts like “payment”, “letter”, and “delivery”. This interrelation, as scholars like Deacon (1998) and Santoro et al. (2021) have noted, is fundamental because when associated terms shift in meaning even slightly (for instance, when credit cards emerged as a new form of “payment”), the meaning of “postage stamp” adjusts accordingly (we understand immediately that stamps can be purchased with credit cards). Such dynamic relationships between concepts constitute the essential, defining aspects of meaning: possessing the appropriate relational network allows one to determine reference even as contexts evolve.

Even in domains that might seem to require more stable conceptual foundations, such as the physical sciences, there is a recognized need for organizing structures such as KPs: in the physical sciences, for example (Peijnenburg and Atkinson, 2003; Peijnenburg and Atkinson, 2007). In fact, when there is data available, the problem facing a scientist is to organize and structure it in different ways. These organizations and reorganizations, upon empirical verification, can yield insights into the nature of reality that would not otherwise be possible.

L. Wittgenstein’s later philosophy precedes this discussion and offers a framework for understanding how meaning emerges through the flexible interconnection that can be formally represented into knowledge patterns.

In fact, according to Wittgenstein, the meaning of words is a socially embedded activity that arises from their use within specific contexts and practices, rather than from a fixed relationship between words and objects or abstract definitions, a phenomenon he called “language-game” (Wittgenstein, 2009, §§7, 23).

In *Philosophical Investigations* (PI), the philosopher proposes that we should understand language by examining how words are used in practice; that is, by observing the associated behaviors and contexts that give them significance (Wittgenstein, 2009, §43).

The challenge of representing this interrelational nature of meaning is further illuminated by PI, §115: “A picture held us captive. And we could not get outside it, for it lay in our language and language seemed to repeat it to us inexorably”, where the picture is our own language that quietly dictates how we look for essences. Wittgenstein advises that when philosophers use a word, such as ‘knowledge’, ‘being’, ‘object’, ‘I’, ‘proposition’, ‘name’, and try to grasp the essence of the thing, one must always ask oneself whether those words are ever used that way within the language-game that is their original home (§116).

As a consequence, Wittgenstein ties meaning to aspect-seeing (*seeing as*). That is,

the same perceptual input can be taken under different aspects, and understanding consists in grasping which aspect is in play within a practice (Wittgenstein, 2009, Part II, xi). Symbol use constitutes a specific class of behaviors that leverage the knowledge that symbol meaning is conventional and practice-bound, and that this meaning can—and indeed must—be continuously adapted within the language games.

As Wittgenstein writes in PI (§242): “If language is to be a means of communication there must be agreement not only in definitions but also in judgments”. Shared judgments about when and how concepts apply constitute an implicit background that are often left unconsidered.

As Halpin (2013) note, language-games are not restricted to natural language alone: the Semantic Web itself can be interpreted as a novel language-game, where URIs function not as fixed referents but as flexible tools for coordination among agents⁸. The authors argue that the act of using a URI in context, across tagging, search, or other agent behavior, constitutes its very sense.

In this way, we go back to the ontology definitions of the beginning of the chapter, as conceptualizations shaped by collective activity.

An apparent tension is then raised: if Wittgenstein is right that meaning resists fixed definitions, how can formal ontologies, which seem to require some structural stability, capture this fluid, use-based semantics? Rather than seeing this as a fundamental contradiction, we can view knowledge patterns as offering a middle path⁹.

KPs preserve certain structural invariants (roles, relations, constraints) that enable computational processing and interoperability, while allowing the specific meanings of concepts to emerge through their deployment in particular contexts and communities of use.

Later, in *Remarks on Colour* (III, §302), Wittgenstein adds: “Would it be correct to say our concepts reflect our life? They stand in the middle of it”. Concepts, though generally collective mirrors of reality, are mediators of our engagement with the world. From the standpoint of social semantics, the meaning of a concept or identifier on the Semantic Web is not determined solely by its logical entailments or the intentions of its creator, but by how communities of agents determine it, use it, and

⁸A Baudrillardian reading would stress that Semantic Web formalisms render sign-to-sign relations explicit (RDF/ontologies) without recovering a privileged tie to the real; on this view, knowledge patterns continue to yield operational meaning rather than referential truth conditions. For a contemporary discussion of Baudrillard’s view applied to AI/hyperreality, see Sommerer (2025), and for the classic thesis, see Baudrillard (1994).

⁹One might note the irony in leveraging the profoundly anti-systematic philosophy of the later Wittgenstein, with his repeated warnings against the craving for generality, to ground a project in formal representation. This tension, however, is productive: several influential strands in AI and information studies explicitly return to Wittgenstein to reframe formalization. P. Agre explicitly invokes Wittgenstein’s later work as part of a practice-oriented reorientation of AI; for example, he argues that philosophical ideas formalize the ways we experience our lives and that an analogous dynamic holds for computational models (Agre, 1997, p. 9), and characterizes formalization itself as a discursive operation that reorganizes a domain of practice (Agre, 1992, p. 25). Collins, in his analysis of tacit knowledge, makes direct use of Wittgenstein’s rule, following considerations to show that rule application presupposes background capacities that resist complete codification (Collins, 2010, pp. 16–17). Even when not offering exegesis of Wittgenstein, design-oriented critiques of symbolic AI draw on a Wittgensteinian problematic: rule-following, background understanding, language as use; for example, Winograd and Flores set out to dissolve rather than solve the classic representational questions (Winograd and Flores, 1986, p. xiii), and develop a background-sensitive view of language and action consistent with these Wittgensteinian themes. On this reading, Wittgenstein functions as a stress-test for naive formalism: the remedy is not to abandon form, but to engineer use-sensitive formalisms whose roles, relations, and constraints enable coordination without pretending to fix meanings in advance.

understand it across time; update it, yet still using fixed reference scaffolding structures to enable information sharing on- and off-line.

We now turn to why this discourse matters. If meaning is used within language-games (*PI* §43; §§115–116), we need representations that let us see how words and concepts are actually used. In Wittgenstein’s terms, a “perspicuous representation” (§122)—a form of representation that produces understanding by “seeing connections”—is a layout that makes connections surveyable. Wittgenstein emphasized “the importance of finding and inventing intermediate cases” (§122) to illuminate these connections. In the context of our work, these intermediate cases correspond to knowledge patterns.

A perspective arises from the dual nature of understanding: the same situation can be organised by different patterns, and each framing makes a distinct situation available to imagination. According to the philosopher (*PI* §531), sameness lies both in what we are about (a shared scene that supports paraphrase across frames) and in our being about it (a grasp tied to “these very words in this very order” or, more generally, to the specific configuration that cues a frame). Some features of a frame support substitution across descriptions; others are non-substitutable, constructional or stylistic cues that anchor uptake. When alternative understandings are on the table, shared grasp requires tacit knowledge: perspective taking to bridge schemas and, when needed, to invent an intermediate pattern that reconciles them. The same patterning holds in language, visuals, and other media, so frames and knowledge patterns must be modelled as multimodal/supramodal structures.

Knowledge patterns, in fact, arrange roles, constraints, and recurrent cases so that one can see how things hang together across domains, while leaving meaning to be fixed by use in particular situations. The same material can be seen under different aspects depending on which description is in play (compare *PI*, Part II, §xi on aspect-seeing). KPs are thus exactly the perspicuous layouts (cf. §122) that make aspects explicit and reusable across domains, and that are therefore desirable since it is due to our lack of a surveyable overview that our failures arise.

Halpin (2013) similarly observe that in the Semantic Web, ambiguity in reference is not a failure but a feature, resolved not (only) by explicit referential mapping but through shared social conventions and usage patterns.

From the point of view of knowledge representation, this coincides with the aim of formal ontologies to somehow capture the fluid and shared interrelational contexts in which concepts operate¹⁰.

Following this line of thought, in Hookway’s terms, KPs are not devices but relations: “the interface” is “more properly a form of relating to technology”, a threshold constituted by the “combined activities of human and machine”, whose

¹⁰E. Husserl and F. Brentano can be read as formal scaffolding for a pattern-based view of meaning. Husserl’s formal ontology, especially the theory of parts, wholes, and moments in the Third Logical Investigation lets us model many frame roles as dependent components of an episode: a payer, amount, or instrument role exists only *in* a payment event and is constrained by typed dependencies among bearer and role. Brentano’s notion of intentionality complements this view by marking certain relations as intrinsically *about* something: an invoice is about a debt, a diagnosis about a condition, a query about an information need. Together they underwrite knowledge patterns that preserve structural invariants while leaving meaning to be fixed by use in Wittgenstein’s sense. This also clarifies *PI*, §531: some understandings are replaceable because two descriptions preserve the same dependence-and-aboutness pattern over a situation; others are not, because uptake hinges on the very configuration that cues a frame (constructional phrasing, genre, layout). Specified this way, KPs are naturally supramodal: the same pattern can be realized in wording, image, or interaction, and coordination comes from shared structure (Husserl, 2001; Husserl, 1983; Simons, 1987; Mulligan, Simons, and Smith, 1984; Brentano, 1995; Jacquette, 2004).

activity “is defined in its coupling of the processes of holding apart and drawing together” (Hookway, 2014). Read this way, a KP maintains separation among heterogeneous actors and representations while selectively permitting passage, co-producing a unified condition of use. This is why, the interface serves as both a constraining gateway—a necessary passage point that shapes and limits how we engage with technology—and a generative space where meaningful interaction actively unfolds within technological use itself.

Following naturally from their functions as perspicuous representations, KPs also force us to analogize. To apply a KP to a new domain requires seeing structural similarities across different contexts, essentially an analogical operation that maps roles and relations from familiar to unfamiliar territory. For instance, KPs can be transferred from source cases to target problems (Goel and Bhatta, 2004) or, in general, be reused by analogy (Gangemi and Presutti, 2009).

To this regard, one analogy-driven ubiquitous and complex, if not the most complex language phenomenon, structural to our thinking, is metaphor. Just as a KP captures the minimal participant/role structure of a recurrent situation, a metaphor overlays two such patterns and lets selected roles fuse into a new, blended pattern. In the same way, metaphors and KPs have predictable inference given by their components.

Wittgenstein and subsequently taken up to varying extent by Rorty and Hesse (1987) and Lakoff and Johnson (1999), among others, find that in studying meaning interrelations we are guided in thought by metaphors and “analogies in language”, without being aware of being guided by anything of the sort (Wittgenstein, 2005, pp. 409, 427)¹¹.

Metaphors, as we will see in the next chapters, reveal connections between seemingly disparate domains. By focusing on these interrelational patterns rather than on fixed definitions, we can develop semantic web resources that better capture the fluid, contextual nature of meaning that Wittgenstein’s later philosophy started paving the path for while still have the shared consensus component. In fact, Halpin’s conception of the Semantic Web as a form of life composed of distributed, asynchronous language-games may hint at the fact that in this context, metaphors, by extending meaning across domains, are central mechanisms for such sense-making, as they generate new language-games by inviting reinterpretation and reconfiguration of KPs.

Because metaphors are, in effect, cross-domain knowledge patterns, they give us the most demanding test-bed for seeing whether our pattern-centric view can really ground meaning in computational applications.

¹¹We do not strictly distinguish, throughout the dissertation, metaphor and analogy, but rather use the word metaphor in a loose sense, in a way compatible to many scholars such as Aristotle (Turbayne, 1962a) and Wittgenstein (Wittgenstein, 2002), who uses the term analogy and metaphor to refer to the same general phenomena. In fact, we adopt the prominent view of cognitive science, where it is argued that most metaphors are predominantly relational comparison, and thus are essentially analogies (Gentner, 1983). As E. Melandri stated, all analogical concepts are metaphors, even if not all metaphors are analogical concepts (Melandri, 2025, p. 176). Besides, the boundary between the two is widely acknowledged to be blurry (Nersessian, 2015): in the same way as Carrillo (2023), we note, there is an overlap between cognitive metaphor and analogy if the latter is understood as a kind of reasoning prompted by the comparison of two subjects. When the term “metaphor” is used to refer to a cognitive as opposed to linguistic phenomenon, it is possible to understand analogy as a special case of metaphor or as a development of metaphor (e.g., Lakoff and Johnson, 1980; Hoffman, 1995), as we do here.

2.2 The nature of metaphor

By bridging two typically incommensurable domains to form a coherent and emotionally vivid concept, metaphor has always been a mysterious phenomenon. The real marvel, however, is that metaphor is repeatedly called an enigma, yet we use it all the time.

In fact, metaphors appear in about a third of the sentences we use in everyday life (Steen et al., 2010), they emerge in all media from text to images and music (Indurkha, 2013), and through emotive impact, they make information understanding easier with respect to literal language (Gong et al., 2020; Fainsilber and Ortony, 1987; Fussell and Moss, 1998; Blanchette and Dunbar, 2001; Crawford, 2009; Mohammad, Shutova, and Turney, 2016). Overall, metaphorical language collapses category boundaries by projecting properties across domains, and relies on vast stores of tacit and ever-changing cultural knowledge: phenomena that even the most advanced computational systems tend to be ill-equipped to capture, realizing the famous U. Eco (1986) prophecy that no algorithm can exist for metaphor¹².

It is thus evident why metaphor is more “the resolution of an enigma than a simple association based on resemblance; it is constituted by the resolution of a semantic dissonance” (Ricoeur, 2004, p. 52), and in that, it exposes precisely the blind spots that current technology must overcome. Even now, as we will show in Section 3, automatic pipelines routinely misclassify figurative usages and reasoning engines overlook the inferential leaps that metaphors license (Wijesiriwardene et al., 2023; Nezhurina et al., 2024).

When we consider the utility and ubiquity of metaphor, these gaps matter: if we look at scientific discourse, for instance, it borrows heavily from everyday source frames (e.g., “genetic code”) and correct explanation of concepts through metaphors are also crucial at different stages of education apart from everyday social and political life (Sfard, 2012).

In this chapter, we will investigate this centuries-long challenge, examining what are the foundations of metaphors, what is the current state of theoretical research on metaphor, and why scientific metaphor is a privileged test case for both AI and theoretical metaphor research.

2.2.1 Conceptual and grounded foundations of metaphor

Historically, metaphor has been treated primarily as a rhetorical mechanism rather than a window into thought. Despite these efforts, its status as a cognitive device traces back to Aristotle¹³. As Aristotle famously described, a metaphor involves “the application of a strange (alien) term[...]from one class to another, or else by analogy” (Rapp, 2002, §8.2). The result of this application is at once a source of clarity and enigma. Indeed, he continues: “Now strange words simply puzzle us; ordinary words convey only what we know already; it is from metaphor that we can best get hold of something fresh...When the poet calls ‘old age a withered stalk’, he conveys a new idea, a new fact, to us by means of the general notion of bloom, which is common to both things” (Rapp, 2002, pp. 183-184).

On this matter, Aristotle was ahead of its time in many ways. Contemporary research continues to recognize that metaphor is fundamental to thought as our entire conceptual system is built on metaphorical mappings. As Lakoff and Johnson (1980)

¹²“No algorithm exists for the metaphor, nor can a metaphor be produced by means of a computer’s precise instructions, not matter what the volume of organized information to be fed in.”, *ibidem*, p.7.

¹³As also Eco argued: see Eco (2014).

put it, “the human conceptual system is metaphorically structured and defined” (p. 3). In practice, this means metaphor operates at the level of ideas to the point thought is primarily conceptual in nature, and abstract reasoning itself is thought to occur by way of metaphorical transfer. Thus metaphor is now viewed as one of the main cognitive mechanisms through which we comprehend abstract concepts and perform abstract reasoning. We will return to this aspect shortly.

But the second important aspect that Aristotle highlighted is that metaphors frame concepts in a new way than merely the sum of its components. It is precisely for this theoretical contribution that the consideration that only precise formal terminology captures truth while everyday metaphors are just loose talk has been highlighted as a trap (Varzi, 2008), since what we call reality might be described equally well (or even better, in some respects) in the imaginative terms of natural language, which in turn sparks a fresh way of thinking. This perspective reinforces a key point in metaphor theory: because the classificatory maps we project onto what Varzi calls “the metaphysical desert” are largely conventional, existing “only in our heads and in our practices”, the metaphors we use every day create and expose, rather than obscure, the conceptual schemes by which we navigate reality (Varzi, 2014, p. 2).

Figurative language, therefore, carries real ontological commitments. Saying that time is a resource, for instance, discloses a tacit model in which temporal intervals resemble scarce commodities, hence guiding our reasoning about waste, investment, and opportunity.

And in this process of language anomaly development, metaphors are monsters: whenever a cross-domain mapping spawns an entity that violates the expectations of its natural kind, we treat it as monstrous not because reality itself is deformed, but because our own conceptual boundaries have been stretched. In this light, metaphors are complex cognitive constructs whose parts can be re-cut, recombined, and occasionally misaligned in ways that reveal the limits of our categorizations. Embracing what A. Varzi (2011) calls a “modest realism” (p.15) reminds us that these figurative frameworks are indispensable yet always revisable; when one fails, we redraw it rather than blaming the world for non-compliance.

Embodied cognition further grounds the view, originally foregrounded by Lakoff and Johnson (1980), that metaphors are constitutive yet expandable by linking metaphors to our bodily experiences (Lakoff and Johnson, 1999). Numerous empirical studies have shown that people systematically use sensory and motor experiences to structure abstract ideas (Barsalou, 1999; Barsalou, 2003; Barsalou, 2008; Casasanto, 2009; Casasanto, 2011; Casasanto and Chrysikou, 2011; Casasanto and Henetz, 2012; Pulvermüller, 2005). Gibbs Jr and Franks (2002) review extensive evidence that metaphoric mappings trace back to the body. Indeed, it is claimed concrete sensorimotor patterns project directly into cognition, so we talk about concepts using the language of kinesthetics and spatial relations. For instance, listeners easily understand “rising prices” or “falling stocks” because everyday experience ties vertical motion to increase/decrease, and describe friendship in terms of proximity or warmth. In another example, our experiences of vertical space and quantity give rise to our conceptual metaphor that associates MORE with UP, and our tactile/thermal experiences give rise to our consideration that AFFECTION IS WARMTH¹⁴.

Therefore, metaphor is embodied: recurring bodily interactions serve as building blocks for more complex mappings. These are called image schemas: preverbal, dynamic patterns abstracted from perception and action (e.g. PATH, BALANCE).

¹⁴Throughout this work, as customary in the literature, concepts are typographically marked in small caps (e.g., SUN); the corresponding lexical items are written in lowercase (e.g., “sun”).

They are first introduced by Lakoff (1987) as “dynamic analog structures arising from perception, bodily movements, manipulation of objects, and experience of force” (p. xiv) and capture structural relations like moving along a path or being inside a container, which we constantly experience in infancy and later on in life. These schemas are analog representations spanning modalities, underlying both spatial and abstract concepts. For example, the CONTAINER schema (in/out) comes from repeatedly moving in and out of rooms. On this view, Narayanan argues that many entrenched image-schematic mappings are not merely correspondences in long-term memory but event structures that can be mentally simulated: understanding a metaphor involves importing the relevant action schema and running its causal-temporal organization in the target domain, so inferences arise from the unfolding of the borrowed process (Narayanan, 1997; Narayanan, 1999).

From this standpoint, multiple theoretical accounts of metaphor have been proposed. In the following section, we will examine the main theoretical frameworks about metaphor.

2.2.2 Theoretical frameworks

This section describes the main theoretical frameworks that have been proposed for metaphor over the years.

2.2.2.1 Conceptual Metaphor Theory

In 1942, philosopher S. Langer proposed the idea that lived experience could be enhanced or extended through the use of analogy in speech (Langer, 2009). She showed that all thinking has a metaphorical status, which makes metaphor a building block of human expression as well as the building block of human experience, since not only do we continuously construct metaphors: they as well influence and shape how we experience being in the world. Seeing that we are “symbol-making animals” is thus the starting point for human understanding. Langer sees articulation not just as speaking or writing but as a bodily and symbolic structuring of experience. The concern is with how novel meaning emerges, through perceptual or symbolic integration that cannot be fully explained by prior components and sees metaphor as a creative, emergent act that fuses incompatible ideas into a new whole¹⁵.

This view of language inspired Lakoff and Johnson in their *Metaphor We Live By* (1980), a book outlining the renowned Conceptual Metaphor Theory (CMT). According to this theory, metaphors create parallels between domains, highlighting experiential correlations (Lakoff and Johnson, 1980). In particular, CMT distinguishes between a *source* and a *target* domain, where the source domain consists of a concrete source of knowledge, which is mapped onto a more abstract domain, the target of the metaphor. For example, they link the conceptual domain of physical HEIGHT with that of QUANTITY in the sentence “The number of books printed each year keeps going up”.

Decades of psychological research have demonstrated the influence of conceptual metaphor on non-linguistic behavior as well, for instance in measures of mental imagery, episodic memory, processing speed, and response time (Gibbs Jr, 2017). Countless articulations of the view have extended it beyond the cognitive sphere and into the moral, social, and political realms (Johnson, Lakoff, et al., 2002; Kövecses, 2020; Yu, 2022), and it remains a vital force in contemporary cognitive linguistics.

¹⁵Just like Aristotle: see Rapp (2002).

Even more so, considering CMT is successful even in technical domains. In a recent study, it was shown that the most frequently employed metaphor theory was CMT in studies about metaphor use in medical settings (Liu, 2025).

To better understand how CMT works, it is possible to take this example: “Let’s build a digital immune system that patrols the network, remembers past infections, and evolves fresh antibodies against new malware”. CMT here aims to show why English is packed with immune-style cyber terms, such as virus or quarantine: the sentence evokes the conceptual metaphor CYBERSECURITY IS AN IMMUNE DEFENSE, with several cross-domain mappings: the computer network is associated to a biological body, where malware is a pathogen, intrusion is disease onset, the incident response is the immune response, and so on.

Another key principle in CMT evident in this example is directional dominance (hereafter called “dominance”). As M. Black foregrounded in numerous works (Black, 1962; Black, 1977; Black, 1977), metaphorical mapping involves an asymmetrical relationship where the source domain structures our understanding of the target domain. In fact, we could say that the immune system works like computer security, but switching the positions of the two, thus saying computer security works like the immune system, would entail a different view and consequence. This asymmetry, also proposed in A. Tversky (1977)’s work on similarity judgments, shows that, at the very least, metaphor tends to be directional, as features and inferences usually flow from source to target, not vice versa. Many studies (among those, see for instance Boroditsky, 2000; Casasanto and Boroditsky, 2008; Casasanto, Fotakopoulou, and Boroditsky, 2010) indeed propose that in English, it is nearly impossible to talk about time without using words whose primary meaning is spatial (e.g. vacations can be long or short). Although it is also possible to use temporal words to talk about space, time-to-space mappings in language are far less common than the opposite.

From the example, though, it is also possible to glimpse CMT’s limitations (see Section 3.1.1.1 for an overview): however CMT treats the sentence as a routine activation of a stored mapping, it does not specify how different ideas such as patrolling, derived from a law enforcement frame, or evolving code, derived from a software development code, get integrated on the fly. It therefore cannot fully predict the mental model the speaker and listeners actually construct. Furthermore, the theory does not model how additional frames or contextual clues can combine in real time.

Beyond dominance, CMT assumes that metaphorical mappings preserve the image-schematic and relational structure of the source insofar as it is compatible with the target. This is the *invariance* hypothesis: a mapping projects structure from source to target in a way that does not violate the target’s basic ontology and constraints (Lakoff, 1990). Thus, in LOVE IS A JOURNEY, we inherit PATH roles such as travelers, goals, and impediments (“we’re at a crossroads”, “we’ve hit a dead end”), while non-compatible source features (e.g., literal refuelling schedules) do not automatically transfer unless conventionalized by discourse. Invariance therefore explains why inference flow in metaphors is both directional and selective: the target’s structure filters what can be imported from the source.

At the same time, the invariance hypothesis has faced criticisms over the years; for example, the fact that mixed (in the sense of the integration of multiple domains) and multimodal metaphors do produce meanings that were not in either domain. In fact, other theories make this emergent structure central and show how apparent violations of source/target topology are resolved by instead considering more in

general a phenomenon of *interanimation* (Fauconnier and Turner, 2008)¹⁶. Indeed, strong version of invariance struggle to account for literary effects that deliberately bend domain structure, needing softer interpretations of the phenomenon.

Thus, despite CMT still being the most prominent theory of metaphor even today, due to its limits, alternative and complementary models of CMT have been carried out.

2.2.2.2 Structure-Mapping Theory

While CMT focuses on why certain domain pairings are connected in metaphors, Structure-Mapping Theory (SMT), developed by Gentner, Holyoak, and Kokinov (2001), underlines that metaphors are a form of analogical reasoning, that is, reasoning involving comparisons while sharing primarily relational information. When we compare two situations, we generate analogical inferences.

SMT aims to show that metaphors operate by mapping a system of relations from a familiar source domain onto a target domain. In particular, SMT postulates the systematicity principle, according to which if a relation holds between objects in the source, then the corresponding relation should hold between corresponding object in the target, and that each element in the source maps to exactly one element in the target so that structural consistency can be maintained. In other words, a good metaphor is not just about two things sharing a simple feature; it is about a deeper alignment where the relationships between elements in one scenario mirror those in another. For example, saying “the mind is a computer” highlights an entire relational structure (processing information, storing data, running programs) mapped from computers to minds. According to Gentner, Holyoak, and Kokinov (2001), proof of SMT-like reasoning is observed by the fact that humans (especially during developmental stages) become adept at focusing on these relational commonalities. Her studies on child development, for instance, found that young children initially make metaphors based on common object attributes (e.g. calling a grapefruit “moon” because both are round), but only later do they grasp metaphors based on abstract relations (e.g. understanding “the mind is a sponge” in terms of absorbing information). This shift, called “the relational shift” (Gentner, 1988), reinforces that mature metaphor comprehension relies on structured mapping. In SMT, a mapping can highlight or import relations already in the base but can never mint brand-new relations. It is only because there is no novel structure emerging beyond candidate inferences already present in the base that SMT aims to make quantitative predictions about which inferences people will entertain.

Let’s take again this example: “Let’s build a digital immune system that patrols the network, remembers past infections, and evolves fresh antibodies against new malware”. SMT maps the alignment between the base (biological immune system) and the target (cyber security platform), where pathogen detection is analogous to threat detection, antibodies is analogous to defensive methods, immunological memory maps to attack, and so on. SMT here finds a one-to-one relational mapping and generates candidate inferences that already exist in the bases, e.g., the fact that the system can remember past attacks. At the same time, SMT cannot explain the organism-like feel of the whole network or that code itself evolves, features that are imported from the biology model. The selective projection helps in explaining how elements are aligned in a metaphor; however, metaphors also involve creativity, multimodality, and frame-shifting basing on new needs or circumstances.

¹⁶The use of the word in this sense seems to date back to Coleridge (1817); for a more exhaustive account of the word for conceptual metaphor studies, see Stockwell (1999).

In the context of our “digital immune system” example, more recent frameworks would emphasize how the analogy operates through categorization mechanisms that identify which biological immune system features are most relevant for cybersecurity applications. Rather than simply mapping existing structural relations, the categorization process actively selects and organizes features based on their relevance to the target domain’s conceptual needs. This perspective is a step forward towards the investigation of creative, meaning-constructive aspects that Conceptual Blending Theory (CBT) addresses.

2.2.2.3 Conceptual Blending Theory

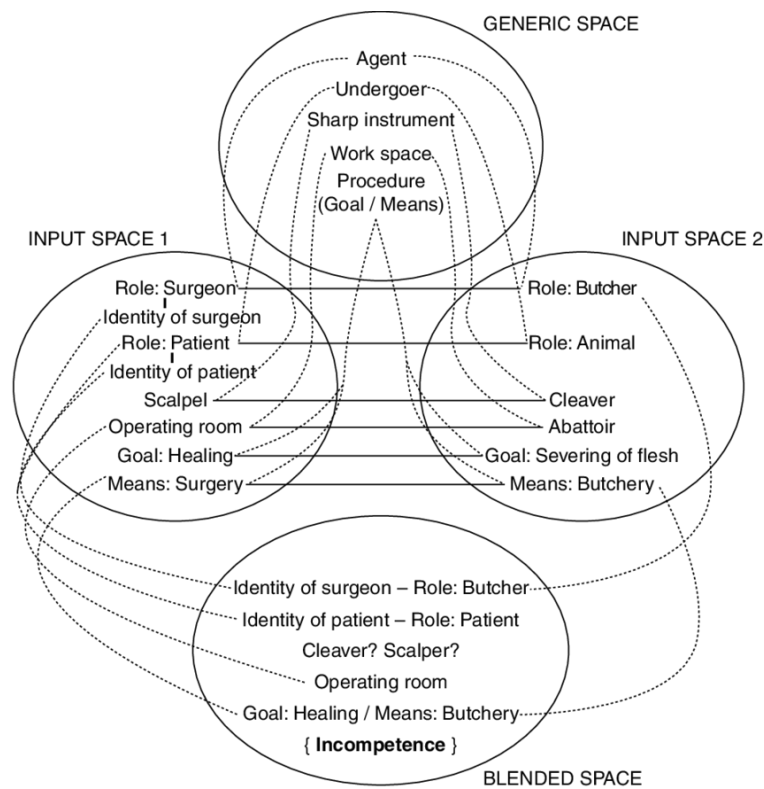
CBT, developed by Fauconnier and Turner (2003), introduces a dynamic model to understand how meaning emerges through the creative integration of conceptual elements. At its core, CBT operates through an integration network consisting of four interconnected mental spaces: two input spaces containing the conceptual domains being blended, a generic space that captures shared structural elements between the inputs, and a blended space where selective projection and emergent structure arise. This network architecture enables the systematic yet creative process of conceptual integration, where elements from the input spaces are selectively projected into the blend according to principles of cross-space mapping and optimization. Unlike SMT’s systematic one-to-one mapping, in working on generic concept combination, CBT also explains how metaphorical understanding often produces emergent meanings that cannot be traced directly back to either domain: the blend a metaphor creates is something genuinely new (as also seen in Section 2.2.1). Thus, in metaphor, the blended space combines and transforms the input elements into meanings that are not directly traceable to the original inputs, as shown in Figure 2.2.

CBT extends both CMT and SMT by addressing their limitations. While CMT identifies stable cross-domain mappings and SMT explains systematic relational transfer, neither theory fully accounts for the creative, emergent aspects of metaphorical thinking. CBT fills this gap by showing how novel conceptual structures arise through blending processes.

Additionally, CBT emphasizes the selective nature of metaphorical framing. When we employ metaphors, we inevitably make framing choices, whether conscious or not, that highlight certain aspects of a concept while obscuring others, depending on our communicative context. This selective emphasis reflects what Lakoff and Johnson (2020) term “metaphorical framing”. In CBT, this selective process operates through the strategic combination of elements from input spaces based on contextual salience and communicative goals. Framing research demonstrates how linguistic choices systematically influence audience interpretation and response (Flusberg et al., 2024): fundamentally, it is a matter of directing attention. This selective attention means that metaphors necessarily provide partial views of reality. Yet this limitation is also their strength: metaphors function as cognitive shortcuts, enabling inference and understanding in unfamiliar domains by leveraging familiar structural knowledge. These dual functions, framing and heuristic reasoning, are deeply intertwined: a metaphor’s ability to direct attention depends on the inferential pathways it opens, while its problem-solving value emerges from the specific aspects it renders salient. CBT helps explain this dynamic by showing how blended spaces selectively integrate elements to create focused understanding.

With respect to the relationship with CMT, CBT recognizes that conventional metaphors (as described in CMT) provide the foundational cross-domain mappings that enable blending, while online processing in working memory creates situated,

FIGURE 2.2: The conceptual blending depiction of the sentence “This surgeon is a butcher” as shown by Oakley and Coulson (1999). Here, the input spaces *butcher* and *surgery* have different goals. While the former aims to kill and disassemble, the other aims to heal and repair. In the blend, these goals are juxtaposed, creating a new emergent meaning. As a result, blending theory can capture complex metaphorical meanings that are not fully explained by traditional CMT, posing that a surgeon who operates as a butcher is incompetent.



context-specific meanings. This dual-level operation explains how metaphorical thinking can be both systematic (drawing on established conceptual mappings) and creative (generating novel insights through blending), which led the theory to obtain particular success in computational treatments of conceptual combinations (see Section 3.0.2). For example, in Wang et al. (2023d), the authors propose conceptual blending strategies in advertising that connect two domains, such as Star Wars and shampoo, using a linking concept based on semantic similarity or automatically generated responses.

More on this topic will be discussed in Section 3.0.2.

Let's apply the example analyzed in the previous sections to CBT as well: "Let's build a digital immune system that patrols the network, remembers past infections, and evolves fresh antibodies against new malware". In CBT, here, two input spaces are *immune system* and *cybersecurity*. The generic space evolves a property of defense, and the blend is the metaphor of a living network guardian, software-based but similar to an organism. CBT is able to predict that the network itself is conceptualised as a body with its own immune staff, which is not present in any single input. Furthermore, the mental spaces fuse into one coherent scenario. But also it is easy to imagine follow ups that CBT's completion-elaboration cycle naturally forecasts.

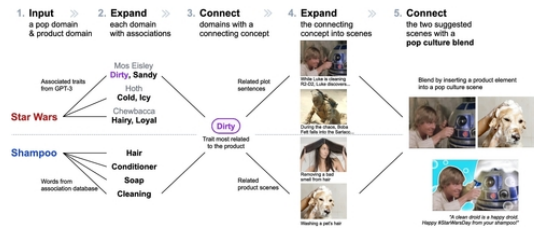


FIGURE 2.3: The conceptual blending strategy in Wang et al. (2023d).

It also might seem that in CBT there is not a strong dominance principle comparable to the one of CMT. Instead, in double (or multiple)-scope networks, all inputs contribute to a novel organizing frame; projection is selective and guided by competing optimality principles, which function as guidelines rather than hard constraints (thus also overcoming the criticisms to the invariance hypothesis: §2.2.2.1). As a result, dominance is typically an analytic outcome of a particular blend (shaped by discourse goals, background knowledge, and human-scale compression) rather than a theoretical axiom to impose in advance. CBT's topology principle is broadly compatible with CMT's invariance hypothesis: both constrain mappings to preserve relational structure, though CBT still allows emergent structure in the blend beyond either input. From an embodied simulation perspective, CMT's soft invariance can be recast as an enactability constraint: only relations that can be executed coherently in the target's ontology, processes, and goals, survive projection. This, too, softens invariance without abandoning it. For the 'digital immune system' example, 'fever' drops out because it is not enactable in a data center, while immune memory persists as a signature or policy store precisely because it is enactable within cybersecurity operations.

This is a deliberate choice: CBT frames blending as a general, routine process that largely escapes conscious awareness, not a rule that speakers explicitly decide to apply. A consequence is that CBT covers symmetric/bidirectional cases and emergent structure more flexibly, but at the cost of weaker *a-priori* predictiveness, a point raised in methodological critiques calling for more testable hypotheses that could clarify whether CBT is a predictive heuristic or an as-if gloss (Gibbs Jr, 2000; Glebkin, 2013). V. Turner himself acknowledged it as a challenge to have psychological studies about when and how blending happens (Huang, 2019). Gréa and Desagulier (2023) explicitly press the falsifiability issue, using arboriform vs forest-like analyses to argue that Conceptual Integration risks explanatory overreach unless it yields testable constraints. Today, this program has seen partial progress: ERP and eye-tracking work on metaphor/humor show online frame-shifting, integration costs, and late positivities, patterns consistent with CBT's integration claims, but still lacks a field-standard diagnostic (Coulson and Van Petten, 2007).

2.2.2.4 Extended Conceptual Metaphor Theory

Extended Conceptual Metaphor Theory (ECMT) is an evolution of Lakoff and Johnson's original CMT in that addresses several criticisms of CMT and expands its framework while simultaneously providing a way to integrate main theoretical frameworks.

As an explicitly integrative proposal, ECMT can be read as a partial armistice in the "metaphor wars", i.e., long-running debates about embodiment and universals vs. cultural variation, literal-first vs. direct pragmatic processing, which mechanism

dominates (CMT/SMT vs. CBT; comparison vs. categorization), and what counts as evidence (intuitions, corpora, experiments) (Kövecses, 2020; Gibbs Jr, 2017)¹⁷.

Kövecses (2020) thus proposed this extended view to answer questions CMT left open, such as the role of context in metaphor, the pervasiveness of metaphor in even literal language, and how multiple metaphors can mix in discourse. As seen in Section 2.2.2.1, traditional CMT defined a conceptual metaphor as a systematic mapping between a source and a target domain, largely rooted in universal embodied experience. However, decades of research and the emergence of new theories such as CMT raised issues like: How do we account for cultural and contextual variation in metaphors? What about mixed metaphors? Is there really a sharp line between literal and figurative language?

ECMT was developed to extend CMT and offer a more comprehensive model that integrates these considerations. While ECMT maintains the core idea that our conceptual system is largely metaphorical, it broadens the architecture of that system and situates it in context. In this view, conceptual metaphors are flexible constructs operating across multiple cognitive and abstraction levels rather than domain mappings in the mind.

A central innovation of ECMT is its four-level hierarchy of conceptual structures. Kövecses (2020) argues that each conceptual metaphor operates simultaneously on four levels of schematicity: image schemas (see Section 2.2.1), domains (see 2.2.2.1), frames (see Section 2.1.2) and mental spaces (see Section 2.2.2.3). These four levels form an interlocking vertical hierarchy: higher levels are more schematic, lower levels more specific. In this context, metaphors exist across all four levels at once. This multilevel approach brings coherence to phenomena that seemed distant from each other: what older theory treated as separate “deep” vs. “surface” metaphors are now just manifestations at different levels of schematicity. For instance, conventional metaphors stored in long-term memory (like ARGUMENT IS WAR) in general reside at the schema/domain/frame levels (“deeper” structure), while a novel metaphorical phrasing in discourse is a mental-space level instantiation (“surface” metaphor) that activates those deeper structures. Within ECMT’s four-level hierarchy, invariance operates primarily at the image-schema and frame levels as a background constraint on what can be recruited into mental-space blends, while the blend itself realizes context-driven elaborations online. ECMT also identifies several layers of context that influence metaphor usage and emergence: situational, discourse, conceptual-cognitive, bodily. Based on these contextual factors, it is argued speakers unconsciously select metaphors.

Situational context encompasses the immediate physical and social environment in which communication occurs. For instance, discussing economic challenges in a war-torn region might naturally evoke military metaphors (“battling inflation”), while the same topic in a medical conference might trigger health-related metaphors (“diagnosing economic symptoms”). Discourse context refers to the broader conversational or textual environment surrounding the metaphor. If a speaker begins with “navigating the market”, subsequent metaphors are likely to maintain the journey frame (“reaching our destination”) rather than switching to unrelated domains. Conceptual-cognitive context involves the cultural knowledge and ideological frameworks that speakers bring to the interaction. Different cultures may systematically prefer different source domains, for instance, Western business discourse often draws on sports metaphors (“winning strategy”), while other cultures might favor family

¹⁷We will return to these issues in Section 3.1.1.1.

or harmony-based metaphors. Bodily context acknowledges that our physical embodiment fundamentally shapes metaphorical thinking. Our upright posture and sensorimotor experiences create systematic patterns like MORE IS UP or EMOTIONAL INTENSITY IS HEAT, providing universal grounding for many metaphorical mappings.

ECMT treats conceptual blending as the highest level of metaphor production, operating in the mental space level of the hierarchy. Mental spaces are in fact where we actively combine elements from various inputs to make sense of a specific scenario. In that, ECMT explicitly acknowledges that CMT and CBT are complementary, not competing. The integration network of blending is seen as the on-the-fly execution of metaphorical thinking, while the more fixed cross-domain mappings of domains provide much of the background knowledge for that process. In other words, in ECMT, mental space blending is the online realization of offline metaphorical mappings.

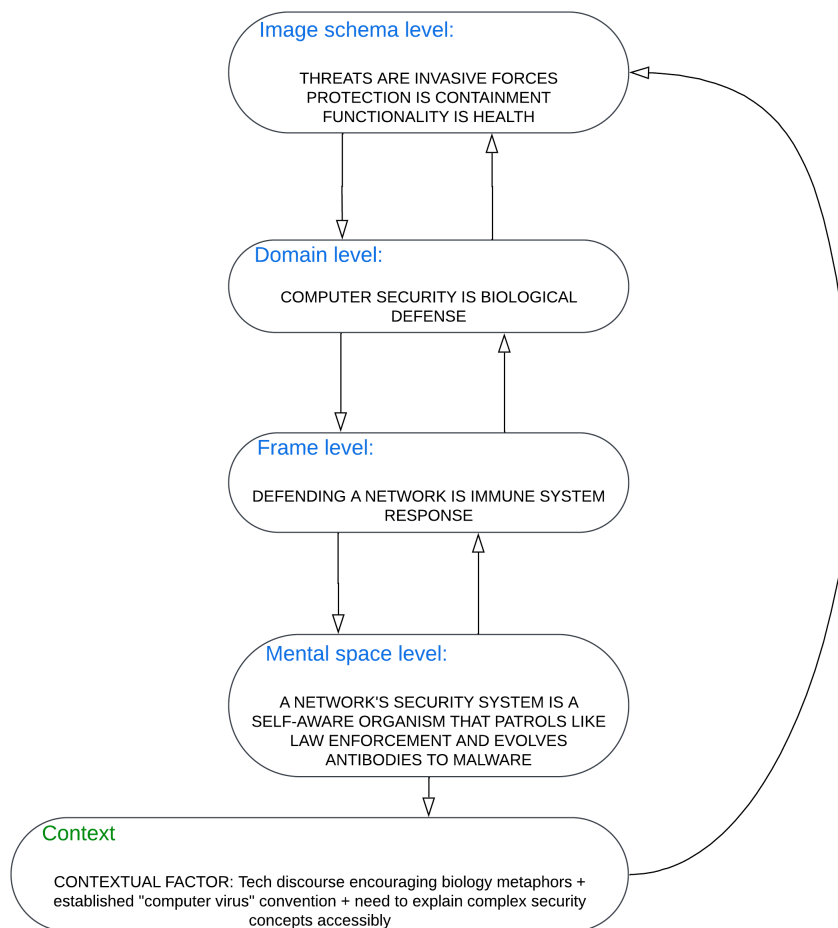


FIGURE 2.4: ECMT analysis of the metaphorical sentence "Let's build a digital immune system that patrols the network and evolves antibodies to malware", showing layers of abstraction across four levels. The progression from basic image schemas through contextually-constrained blending illustrates how online metaphorical reasoning produces novel conceptualizations beyond established mappings.

For example, as shown in Figure 2.4, when recalling the aforementioned example “Let’s build a digital immune system that patrols the network and evolves antibodies to malware”, we are blending at least two input frames in a novel way. Classic CMT can explain entrenched mappings like COMPUTER VIRUSES ARE VIRUSES (with words like “virus” or “quarantine”), but it doesn’t specify how we dynamically merge additional frames (“patrolling” comes from a policing frame, “evolving code” from a biology frame) into a single imaginative scenario. ECMT handles this via mental-space blending: multiple frames can be imported into the blend, yielding an emergent metaphorical scenario (a “network immune system” acting like a self-aware organism) that is not found in any one input domain. ECMT argues this online blending does not occur in a vacuum, but rather draws on established mappings (e.g. our longstanding conceptual analogy between computers and biological contagion) to facilitate the integration, and the choices of frames are constrained by context (e.g. a tech discourse encouraging both biology and security frames). ECMT thus blurs the line between literal and metaphorical. It proposes that what we call literal meaning is often just highly conventionalized metaphor that we no longer recognize as figurative.

2.2.2.5 Pragmatics

According to Mey (2009), pragmatics is “the study of language in a human context of use”, and in being that, a domain that goes beyond semantics, where the actual meaning in a context of use is determined, and is usually implicit. As such, pragmatics aims to study phenomena such as linguistic acts, conversational implicatures, and figures of speech.

As metaphor shows, not all statements state something: often, in fact, instead of wanting to communicate information, we use words to do (*perform*) something: we owe this distinction to J. Austin, though in this consideration he was close to the late Wittgenstein philosophy of language (§2.1.3.2) (Austin, 1962).

In fact, Austin (1962) showed that when we utter a sentence, we are actually performing three simultaneous actions (speech acts: from here, we talk about Speech Act Theory (SAT): (i) the fact of engaging in meaningful communication (*locutionary* act); (ii) in uttering a sentence the producer is doing something specific with their words, for example the action of performing a statement, a request, a warning, etc. (*illocutionary* act). By illocutionary force it is meant the conventional act type performed in saying something, such as asserting or requesting; (iii) the effects that the utterance has on the receiver (*perlocutionary* act).

Classic pragmatics and SAT thus highlight that what is *done* with an utterance can diverge from what is literally said. An everyday complaint like “You’re standing on my foot” typically functions as a request, not a description.

A metaphor usually inherits the illocutionary force of the sentence it appears in¹⁸. When one uses a metaphor in a declarative sentence, they are typically making an assertion, just not a literal one. However, there is disagreement about what exactly is being asserted. On one influential line, the sentence keeps its literal meaning and force while the speaker uses it to convey something extra by pragmatic means. D. Davidson is the clearest defender along this line, arguing that metaphors mean just what the words literally mean; their special contribution lies in use and effect (Davidson, 1978). On this view, a metaphor in a declarative still performs an assertion, but any metaphorically-adjacent insight fits into the perlocutionary one.

¹⁸Even though Austin himself did not extensively discuss metaphors in his work on speech acts, later pragmatists did extend SAT to metaphor (Searle, 1979).

Others argue that a metaphorical assertion is still an assertion but with nonliteral content. Work in this vein includes Black's interaction view (Black, 1962), and treatments that model metaphors as yielding an adjusted concept while keeping assertoric force, as in Relevance Theory (RT).

RT argues for direct pragmatic access to metaphorical meaning (Sperber and Wilson, 1995; Carston, 2002). Rather than treating metaphors as maxim violations requiring special interpretive procedures, RT reframes them as ad hoc concept construction, where hearers pragmatically narrow or broaden lexical meanings until an optimally relevant interpretation emerges. Here, hearers flexibly adjust concepts to meet communicative goals without necessarily computing literal meanings first. In this way, metaphor is treated as an inferential route to implicatures and evaluative stance (Sperber and Wilson, 1995; Wilson and Sperber, 1988), emphasizing that figurative language processing can be as automatic and efficient as literal language comprehension.

A third cluster gives metaphor a distinctive pragmatic profile, for example K. Walton's "make-believe model", which construes metaphor as inviting participation in a game of imagining, shifting what kind of act is being done in context (Walton, 1990). A further line proposes explicit metaphorical speech acts, though even here the typical force is said to be inherited from the host sentence rather than being *sui generis* (Cohen, 1975; Huttar, 1980). These views are not fully reconcilable, and the debate is still very much alive.

Another thing to note is that a single metaphorical utterance can simultaneously perform multiple illocutionary acts.

In fact, the illocution is not determined by clause type alone, and pragmatic force is often graded or blended. Metaphors rely on shared cultural knowledge to determine intended force, making interpretation more variable than with literal speech acts. Treating intentions as properties attributed to acts can make them more tractable for annotation and computational modeling (Gibbs, 1999; Michelli, Tong, and Shutova, 2024).

Aligning with ECMT's multi-level contextual constraints, pragmatics thus aims to explain how hearers recover speaker meaning in context and why some metaphors feel effortless while others are attention-grabbing. Differently from so-far described theories of metaphor, which focus on what structures make a metaphor possible, pragmatics aims to understand which interpretation of metaphor can become current for the actual audience, given tone, style, and social situation, and under what appropriate circumstances this happens or not.

P. Grice's influential theory of conversational pragmatics offers one account, compatible with the interactionist view, of how metaphors work in context. His Cooperative Principle assumes that speakers generally try to be truthful (maxim of Quality), informative (Quantity), relevant (Relation), and clear (Manner) in their contributions to conversation (Grice, 1975). When utterances seem to flout or exploit these maxims, as metaphors would if read literally, hearers infer additional meanings to preserve the assumption of cooperation. On a traditional two-stage reading, comprehension proceeds by first accessing a literal interpretation and then deriving a figurative one when the literal reading proves contextually inappropriate, but this sequencing is not required by Grice's own framework (Grice, 1975). Work influenced by J. Searle has often treated metaphor as a non-literal layer derived from prior literal interpretation (Searle, 1993). By contrast, RT denies any mandatory literal-first stage and allows direct pragmatic access to a metaphorical interpretation when it best satisfies relevance constraints (Sperber and Wilson, 1995; Carston, 2002).

On a speech-act picture, cues help fix or re-calibrate illocutionary force *in situ*. A rising contour, slowing tempo, or a smile voice can shift an apparently assertoric metaphor toward a joke or a tease; compressed tempo and hard onsets can tip it toward a warning. These non-lexical cues encode procedural meaning that guides the hearer's inferential search (Blakemore, 2002; Wilson and Wharton, 2006; Wharton, 2009). They prioritize hypotheses (e.g., irony, urgency, reassurance), mark focus, and calibrate stance, which helps explain why the same metaphor can perform different social actions in context. In interaction, the illocutionary force achieved (request, warning, joke) is not determined by clause type alone, but by these pragmatic resources operating with contextual assumptions (Levinson, 1983). Politeness and facework further recalibrate metaphor's force via hedges and mitigation (softening face-threats) or boosters (amplifying commitment) (Brown and Levinson, 1987). Finally, conventionality and salience modulate access: salient (often literal or well-entrenched metaphorical) meanings are recruited early and can facilitate or compete with novel readings (Giora, 2003). These mechanisms predict variation by genre, expertise, and culture, thus force is realized by how the utterance is instantiated across channels, not by clause type alone, as Popa-Wyatt (2022) argues¹⁹.

Applied to our running example ("build a digital immune system that patrols the network and evolves antibodies"), a pragmatic account models how hearers adjust *patrol*, *antibody* and infer a proactive, adaptive stance with minimal extra machinery. It also clarifies when a metaphor is processed as deliberate: if it is designed to shift perspective and manage attention, hearers adopt comparison-based processing rather than routine categorization (see Section 2.2.2.6), and its form-value pairing interfaces naturally with semiotic analyses of multimodality (see Section 2.2.2.7).

2.2.2.6 Deliberate Metaphor theory

Influenced by pragmatics, and aiming to explain metaphor in terms of reasons as opposed to causes (as happens in traditional metaphor theories such as CMT), Deliberate Metaphor Theory (DMT) has emerged as a new theoretical framework (Steen et al., 2015). It was set to resolve the "metaphor paradox": why most everyday metaphors do not seem to involve conscious cross-domain mapping, still there are some that are used deliberately (therefore not all metaphor in the structures of language and thought may be processed by means of a local analogy, cf. Holyoak and Stamenković, 2018).

Such theory is an attempt to deal with the intention behind metaphor use systematically (Steen, 2008; Steen, 2017), as it defines deliberate metaphor use as the intentional introduction of a topic shift or perspective change to the discourse. According to C. Steen, a radical way of responding to the growing attention to variation in metaphor and its processing is to adopt the position that not all metaphor in language and thought works in the same way as metaphor in communication. The paradox can then be resolved when the perspective is broadened beyond metaphor in language and thought to metaphor in use. The shift of attention to communication reveals another important function: perspectivization (and its relation to mental world-making)²⁰.

¹⁹According to Popa-Wyatt (2017), within a multi-channel view, SAT also helps to understand how figures like ironic metaphor are composed. The same metaphor, with a flat affect and deadpan timing, can convey a ridiculing act; packaged instead with warmth, it conveys consolation. A full treatment of these compound figures, however, lies beyond the scope of this section.

²⁰On this view, we follow Michelli (2025), according to who what matters is whether hearers can supply reasons for why that metaphor was used in context. Building on Anscombe's distinction between

For example, taking the sentence we used in previous sections as example, we could note the sentence could constitute a deliberate metaphor because it involves novel language choices such as patrols that are not conventionally used in cybersecurity discourse. A non-deliberate but similar metaphor would be: “The antivirus software detected and quarantined the infected files to restore system health”, which uses the same biological source domain but operates as conventional metaphor. Still, DMT keeps the focus on the on-line comprehension of concept blending, thus remaining agnostic about other aspects: whether invoking biological immune systems actually leads to better cybersecurity solutions, whether it creates problematic blind spots, or whether it will successfully influence practitioners and policymakers in the intended direction.

From this example, it is then evident that despite the emphasis put by DMT on a new aspect of metaphor theory, there are still some gaps to fill in²¹. The first one involves the classification of deliberate metaphors, i.e. the difference and relation between different types of deliberate metaphors are not shown. Furthermore, since the object of previous studies is English deliberate metaphors, the status of other languages remains unclear: DMT was criticized about its possibility to study deliberate metaphors in different languages. Apart from languages, other factors such as geographical environment and cultural tradition should also be taken into account to study metaphor understanding. Finally, one could ask what is the relationship between conscious and unconscious thoughts, thus the relationship between deliberate and nondeliberate metaphors.

Here, some scholars such as Z. Kövecses believe that nondeliberate metaphors base deliberate metaphors, which involves multi-level cognition. However, it is in general not clear how nondeliberate metaphors affect deliberate metaphors and how long-term memory and online processing work together (Kövecses, 2020). For this aspect, CBT can come in hand, since as seen in Section 2.2.2.3, conceptual blending operates across both conscious and unconscious cognitive levels, providing a framework for understanding how nondeliberate metaphorical mappings from long-term memory can be dynamically integrated with deliberate metaphorical constructions during online processing, potentially through the blending of multiple mental spaces that bridge automatic and controlled cognitive operations. Still, there does not seem to be established literature explicitly connecting DMT and CBT as complementary approaches.

For now, however, attention to metaphor is the key issue in the question about the relation between embodied cognition and social interaction. This aspect is elicited by deliberate metaphor use, and since deliberate metaphor is rather infrequent in most discourse contexts, according to Steen, this is the moment when embodied cognition and social interaction can meet.

causes and reasons for action, this account introduces intentional^A metaphors, namely uses where reasons beyond mere communication are commonly attributed. This reframing preserves DMT’s focus on perspectivization while avoiding reliance on conscious access and explains production–reception asymmetries as a matter of socially shared attribution rather than introspection. In fact, Intentional^A differs from both deliberate metaphor, which requires being used and processed as a cross-domain mapping, and purposeful metaphor in Critical Metaphor Analysis; the former is a processing claim, the latter an analyst’s reconstruction of goals, whereas intentional^A is a community-level judgment about reasons in use.

²¹For a more complete insight on the debate, see Michelli (2025).

2.2.2.7 Semiotic theory

Semiotic approaches to metaphor examine how metaphorical meaning emerges through systematic sign relationships that operate across multiple planes simultaneously. Rather than focusing primarily on conceptual mapping between domains, semiotics investigates how metaphors function as cultural signs embedded within broader systems of meaning-making. This perspective reveals metaphors as complex artifacts that generate significance through both their representational content and their formal properties, operating within culturally specific frameworks that shape interpretation. In other words, an important premise in this perspective is that content and form are not discernible: the sign also contributes to what content can express.

The semiotic analysis of metaphor builds on R. Barthes' three-level structure of signification (Barthes, 1964): denotation, connotation, and myth. At the denotative level, metaphors establish connections between conceptual domains through recognizable comparisons. At the connotative level, they activate networks of cultural associations that extend beyond literal meanings. At the mythological level, they participate in ideological systems that appear natural but are culturally constructed. This stratification explains how metaphorical signs evolve from fresh comparisons into habituated cultural forms that eventually function as naturalized myths.

Classic semioticians like C.S. Peirce classified metaphor as a type of iconic sign (a *hypoicon*), one that represents a concept by resemblance or parallel structure to another. Peirce saw metaphor as a form of iconic reasoning, capable of bringing novel insights in thought²².

Following Peirce, modern semioticians emphasize that the meaning of a metaphor is not fixed by an innate mapping alone; instead, it emerges through an interpretative process involving the interpreter's knowledge and context, as shown by Barthes' levels.

The cultural embedding of these sign operations occurs through what Eco describes as the cultural encyclopedia. Metaphors prompt interpretive leaps through shared cultural codes, making their comprehension dependent on culturally specific sign systems. Lotman's concept of the semiosphere situates this process socially, explaining how cultural positioning determines where metaphors will be successfully decoded and where they will encounter resistance or misinterpretation (Lotman and Clark, 2005).

Central to semiotic metaphor analysis is also the recognition of two distinct but interacting planes of meaning. The pictorial plane presents recognizable objects, scenes, and representational content, while the plastic plane operates through pure formal relationships independent of representational meaning. Floch (1985)'s distinction between figurative and plastic meaning demonstrates that metaphors generate significance not only through what conceptual domains they connect, but through how their formal structure shapes interpretation. When Floch, Sherry, and Bodkin (2001) analyzed corporate visual identity, they showed how IBM's horizontal stripes versus Apple's curved forms create opposing value systems (complex/simple, cold/warm, rigid/flexible) through structural contrast alone, demonstrating that plastic elements contribute independent meaning to metaphorical construction.

This cultural specificity reveals how metaphors function ideologically within semiotic systems. By operating through formal relationships that appear natural,

²²"For a great distinguishing property of the icon is that by the direct observation of it other truths concerning its object can be discovered than those which suffice to determine its construction" (Peirce, 1932, p. 2.279). See also Lattmann, according to who Peirce defined metaphors as a genuinely iconic sign that represents "a novel sign relation and, hence, novel thought" (Lattmann, 2012, p. 536).

metaphors make certain worldviews seem inevitable rather than constructed. The plastic level becomes particularly significant here, as formal elements carry ideological weight through systematic oppositions that organize entire frameworks of understanding. Recent research in visual narrative has begun to recognize these principles empirically. El Refaie (2003) first highlighted that manifestations of conceptual metaphors must account for the specific affordances of visual modes, while subsequent work (El Refaie, 2009; Teng, 2009; Bounegru and Forceville, 2011) has demonstrated that CBT-based approaches handle multiple input spaces operating across different semiotic planes more effectively than approaches focused solely on conceptual mapping (Bergen, 2004; Cohn, 2013).

However, as Schilperoord and Maes (2009) observe, such analysis remains “driven by genre-related, pragmatic considerations” (p. 219) and cannot escape pragmatic affordances. This recognition reinforces the semiotic insight that metaphorical meaning emerges through systematic relationships between signs operating within specific cultural contexts.

The semiotic approach thus examines metaphors as complex cultural artifacts whose formal properties carry independent significance within broader sign systems. By analyzing both conceptual mapping and semiotic operation, this perspective reveals how metaphors participate actively in constructing the frameworks through which reality becomes intelligible, making certain interpretations appear natural while obscuring their cultural constructedness.

Current semiotic studies of metaphor often seek to integrate with cognitive science research while maintaining the semiotic focus on signs and interpretation. Efforts to do this, for instance, show that when multimodal work appears in cognitive approaches such as ECMT (e.g., Forceville, 2002), it typically demonstrates that conceptual mappings can arise across media without examining how visual grammar itself contributes independent meaning. Semiotic analysis addresses this limitation by providing systematic tools for examining how formal properties interact with cultural positioning to generate ideological effects, revealing metaphors as sites where meaning, power, and cultural reproduction intersect through sign relationships.

Another example of integration attempts of semiotics is given by M. Danesi’s *metaform* theory as cross-code products of blending/metaphors, and explicitly constitutes “one attempt [...] to amalgamate the work in CMT with work in sign theory”, thus treating blends as metaforms whose results show up in language, art, math, and science (Danesi, 2013, pp. 397–399).

However, many semioticians explicitly critique and differentiate their stance from prevailing cognitive theories of metaphor. They argue that a purely cognitive-linguistic account can overlook how metaphors function as signs within communication, and how meaning is negotiated by interlocutors, contending that static mappings do not accommodate the polysemy of signs. Furthermore, semioticians following Peirce and Eco insist that nothing is a metaphor unless it is *interpreted as* a metaphor (Jiang, 2017).

Semiotic theorists also resist the career of metaphor hypothesis (§2.2.3), which implies that a metaphor’s life ends as soon as it conventionalizes. From a semiotic perspective, even dead metaphors remain part of the sign system and can be reactivated or reinterpreted in new contexts, showing an endless meaning potential.

2.2.3 Processing metaphorical meaning in natural cognition

The last 40 years, driven by these fertile methodological and theoretical accounts of metaphor, have witnessed sustained efforts to understand how we process metaphorical meaning and make inferences based on the elements that connect two domains in a metaphor. Central to this aim has been resolving whether literal meaning has primacy over metaphorical meaning, and understanding the cognitive mechanisms that enable metaphorical comprehension. This chapter examines current debates surrounding metaphor processing in natural cognition, focusing on the comparison versus categorization debate, and methodological challenges facing the field.

As already introduced in Section 2.2.1, one of the main critical turnarounds in metaphor research is that metaphor is fundamentally embodied. Recent empirical studies on metaphorical meaning processing confirm this: performing or even imagining body movements that depict metaphoric actions (e.g. “pushing an argument”) enhances metaphor comprehension, and watching action-related videos improves language comprehension (Khatin-Zadeh et al., 2023; Garello et al., 2024). However, the case of sensory impairment (Phillips, Grenoble, and Mason, 2023) challenged strong embodiment by showing that individuals without somatosensation can still comprehend and produce tactile metaphors, suggesting at the very least that linguistic experience can compensate for absent sensory experience. This likely occurs through multiple routes. Abstract conceptual knowledge can be built through distributional patterns in language, contextual associations, and explicit explanations. Moreover, metaphor processing involves multiple mechanisms that can operate on linguistic representations independently of sensorimotor simulation. These processing mechanisms have been long studied, and continue to reveal interesting results.

One study carried out by Nadler et al. (2025) evaluates embodied vs. statistical accounts of metaphor through color metaphors, comparing color-seeing adults, color-blind adults, painters, and text-only computational systems. Humans, including color-blind participants, were seen to draw on embodied reasoning for novel metaphors, while computational systems diverge from human patterns, especially under negation and novelty: novel metaphors more often require simulation²³ of imported structure to cohere (be *plausible*: §3.0.2). On a widely cited account, metaphors recruit embodied event structures: a familiar action schema (e.g., SOURCE–PATH–GOAL, force dynamics) is mapped into the target and run there as a mental simulation (Narayanan, 1997; Narayanan, 1999). The selective inferences we draw (obstacles, detours, slowing, suspension and resumption, completion) arise from the unfolding of the borrowed control structure under the target’s constraints; in this sense, projection is allowed by enactability and not mere co-occurrence.

On the Neural Theory of Language account (Lakoff and Narayanan, 2025), understanding is carried by simulation over embodied schemas and executing networks (X-Nets), which specify roles and a control flow for events (goals, enabling steps, obstacles, completions). These control structures are supramodal: the same runnable pattern can be anchored in vision, touch, proprioception, affect, or linguistic imagery. As a result, metaphorical mappings do not depend on any single sensory channel; they recruit whatever cross-modal resources suffice to run the

²³Following S. Narayanan’s model, event knowledge is encoded as interconnected x-schemas (executing schemas). Metaphor maps align these schemas, carry quantitative parameters (e.g., speed to rate of progress, distance to degree of completion), and align participants; executing the mapped schema writes consequences into a temporally extended belief model (Narayanan, 1997; Narayanan, 1999).

imported event logic in the target domain. In this sense, we argue knowledge patterns are likewise supramodal: they encode process structure that can be instantiated by different modalities without loss of inference. This helps explain why sensory-impaired participants can pattern like the sensory-typical: multiple modalities can stabilize the same control schema, yet text-only systems, which lack embodied control circuitry, diverge under negation and novelty, precisely when successful interpretation requires running the pattern rather than retrieving co-occurrences.

According to Bowdle and Gentner (2005)’s “career of metaphor” hypothesis, metaphor comprehension can arise from both comparison (Gibbs Jr, 2017; Gibbs Jr and Colston, 2012) and categorization mechanisms (Glucksberg and McGlone, 2001; Glucksberg and Keysar, 1990; Glucksberg, 2008). In this view, comparison-based theories suggest that when processing a nominal metaphor (e.g. JEALOUSY IS A TUMOR), we align the properties of the target (JEALOUSY) with those of the source (TUMOR) and look for an overlap (e.g. *both* grow uncontrollably). In contrast, categorization is based on semantic categorization, involving mappings between conceptual domains (such as the form “A is B” metaphors), so that the target of the metaphor is treated as a member of a more generic category exemplified by the source. Thus, in this perspective, the comprehension of a metaphor proceeds through a process of abstraction, suggesting we create an ad-hoc category, or “ground”, (e.g. *anything* that grows uncontrollably) that encompasses both the source and target term. According to this view, the metaphor JEALOUSY IS A TUMOR is interpreted as a statement that jealousy belongs in the category of anything that grows uncontrollably.

It is argued that analogical reasoning fundamentally operates through categorization processes that solve “the problem of feature relevance in analogies, especially in the case of conceptual innovation and knowledge increase in scientific domain” (Bianchini, 2016, p. 239).

According to this argument, analogies require models to establish salient properties and relations that enable knowledge transfer between source and target domains. It is indeed suggested that models function both within analogies (to connect situations) and as analogies themselves (providing elements for scientific discovery), as opposed to traditional approaches that treat models as either one and another (Bianchini, 2016). This categorization approach helps explain how analogical reasoning determines which structural features are relevant for mapping, a question that SMT’s systematicity principle alone cannot fully address.

The literature proposes an evolutionary trajectory from comparison-based to categorization-based processing as metaphors become conventionalized. DMT in fact proposes that deliberate metaphors engage comparison-based processing through conscious cross-domain mapping, while non-deliberate metaphors use categorization without metaphorical awareness (see Section 2.2.2.6).

According to this framework, metaphor comprehension shifts from comparison to categorization as metaphors are conventionalized, with the processing model determined solely by the conventionality of the metaphorical source.

Modern cognitive research in fact confirms both comparison and categorization processes can occur in a non-mutually exclusive way. The prevailing view is

that the route taken depends on factors like conventionality and context, as conventional metaphors behave like literal language and may be processed via categorization with minimal comparison, while novel metaphors require active comparison to establish a new mapping (Gokcesu, 2009). This phenomenon is consistent with evidence from reaction time and neuroimaging studies, where conventional metaphors are often comprehended as quickly and effortlessly as literal statements, whereas novel metaphors show slower comprehension (Koller, Müller, and Kauschke, 2022). However, some researchers argue that the strict comparison/categorization dichotomy is oversimplified. For instance, Barnden (2018) suggests that understanding a metaphor may lie along continuous dimensions.

In this context, Eco distinguishes between expressions produced by *ratio facilis*—where symbolic relationships are conventional and easily decoded—and those generated through *ratio difficilis*, which demand creative interpretive labor from the recipient. Powerful metaphors operate precisely through this *ratio difficilis*: they resist immediate comprehension, forcing the interpreter into an active process of meaning construction that often yields novel information (Eco, 1976; Eco, 1986).

This concept connects directly to CBT's integration networks: metaphors operating through *ratio difficilis* create blended spaces with greater emergent structure precisely because they require more extensive cross-domain mapping and creative projection. The cognitive effort involved in processing the difficult reasoning of powerful metaphors parallels the complex integration processes that CBT describes, where meaning emerges from the creative tension between input spaces rather than simple transfer²⁴.

Another foundational question about metaphor processing is whether we interpret a sentence's literal meaning before any figurative meaning (the indirect vs. direct access debate: see Genovesi, 2023, and §2.2.2.5). The first thing we get to know in education when we hear or speak about metaphor, in fact, is that there is a separation between literal and metaphorical meaning. The question that arises, then, centers around the primacy of literal meaning and whether it always is a necessary step for the lexical access of the metaphorical meaning of the same statement or words. In the indirect access view by Grice (1975), the idea is that the speaker can assign a metaphorical meaning to the utterance only after knowing its literal meaning, or in other words that there is a difference between the linguistic meaning and what the speaker implies. In this way, the literal meaning would help to understand the property²⁵. A similar position is held by Searle (1979) according to which metaphor processing involves recognizing that the speaker cannot literally mean what they say, leading the addressee to seek a meaningful interpretation through pragmatically driven implicature. Thus, according to his view, metaphorical meaning emerges when literal interpretation would violate principles of cooperative communication.

Another side suggests that metaphorical meaning can be accessed directly (the direct access view: Harris (1976), Gildea and Glucksberg (1983), and Gibbs Jr (1984), so literal meaning does not have unconditional priority. Empirical evidence from cognitive neuroscience has shown mixed results into this debate, showing for instance that even when a metaphorical meaning is accessed, literal semantic features may be activated in parallel (Weiland, Bambini, and Schumacher, 2014). On the other hand, many behavioural studies have shown that context and familiarity modulate

²⁴This, of course, does not mean that metaphors that are supposed to require less effort to comprehend because they are conventional are less relevant, as shown by semiotic theory: §2.2.2.7.

²⁵For an analysis claiming these stances have been at least in part misunderstood, see Genovesi (2023).

the need for literal processing (Weiland, Bambini, and Schumacher, 2014). In this context, Giora and Fein (1999) proposes that the salience of a meaning determines access in the Graded Salience Hypothesis. Thus, the consensus for now seems to be that literal meaning does not always have primacy, instead, factors like context and familiarity drive the speed and order of meaning activation.

Furthermore, in current cognitive research, metaphor aptness, conventionality, and deliberateness have emerged as key dimensions for metaphor comprehension. In fact, Gibbs Jr (2023) emphasizes that people use metaphorical language not just to convey meaning but to achieve various social and pragmatic goals.

Metaphor aptness, that is, the extent to which a metaphor captures essential features of the topic (Banaruee et al., 2024), is often measured via normative ratings of a metaphorical statement on a Likert scale. As it captures something about the internal coherence of the metaphor, empirical studies show that controlling for familiarity, higher aptness metaphors yield faster reading times (Thibodeau, Sikos, and Durgin, 2018; Stamenković et al., 2023). Also, it seems people prefer highly apt metaphors. However, there is some debate about whether aptness is really distinct from similarity, suggesting aptness essentially reflects the perceived quality of the commonalities between source and target. Aptness is independent from conventionality: a novel metaphor might be extremely apt, a conventional one inapt (though its survival in time demonstrates its aptness).

Conventionality refers to how common and established a metaphorical expression is within a community. Conventionality, too, is typically operationalized through rating scales or corpus frequency measures. Koller, Müller, and Kauschke (2022) found that among 116 neuro-studies of metaphors/idioms, familiarity was the most frequently controlled factor: 69 of 116 studies took it into account in stimulus selection. Often, novel metaphors are thus operationalized by contrast as those with low familiarity ratings. Furthermore, some studies treat conventionality as binary while others as a continuum; in recent work however the tendency is to quantify as a continuous predictor in analysis, as metaphors can indeed vary in degrees of conventionality. In general, conventional metaphors are easier and faster to understand: Yao et al. (2024) shows that conventionality and context allow faster performance on novel metaphors in supportive versus literal contexts.

Neuroimaging reviews (e.g., Bohrn, Altmann, and Jacobs, 2012) report that novel metaphors uniquely activate certain brain regions (such as right inferior frontal gyrus and right anterior cingulate cortex) that are not significantly active for conventional metaphors (Koller, Müller, and Kauschke, 2022). In behavioral terms, studies like the one by (Bowdle and Gentner, 2005) found that participants tend to prefer an explicit simile form for novel comparisons ("X is like Y"), but are comfortable with the direct metaphor form ("X is Y") for conventional metaphors, aligning with the idea that novel metaphors need a bit more comparative scaffolding while conventional ones are readily seen as categorical statements (Gokcesu, 2009).

Finally, deliberateness in metaphor usage is a concept introduced by recent theories (most prominently in Steen (2017), see Section 2.2.2.6) that distinguishes metaphors by the speaker's communicative intention and the resulting processing mode for the listener. Deliberateness is a more qualitative construct and is harder to operationalize than aptness or familiarity. Recent studies have even used eye-tracking to see if readers spend more time on metaphors when they are presented as creative literary instances (which might prompt deliberate reflection) as opposed to casual conventional phrasing. For example, Reijnierse et al. (2015) asked whether highlighting a metaphor's source domain imagery would change processing: this kind of design tests deliberateness by making a metaphor more or less salient to readers.

Alongside these aspects is that of visual properties and how they influence meaning processing. While semiotic theory has long shown that metaphors' plastic properties such as visual, sensory, and formal qualities contribute to meaning generation (see Section 2.2.2.7), empirical investigation of these claims remains limited. The few existing studies offer support: Dou, Liu, and Chen (2023) used image-based visual analysis to show that literal and metaphorical meanings of Chinese color terms are perceptually distinguishable, providing nonlinguistic evidence for embodied color metaphors. Similarly, Schloss (2024) demonstrated that people infer meanings through color-concept association networks. However, these isolated findings highlight a significant gap: the rich theoretical insights from semiotic approaches to metaphor have yet to receive systematic empirical validation, particularly regarding how formal and sensory properties influence online processing. This represents a promising direction for future research that could bridge semiotic theory with cognitive processing accounts.

2.2.3.1 Fast and slow thinking

To understand how metaphors achieve their performative effects, we need to examine the cognitive mechanisms through which they operate.

Bleakley (2017) uses a classic X-ray vignette to show how metaphors accelerate expert thinking. A senior radiologist inspects a barium-enema film, as shown in Figure 2.5, notices a concentric narrowing whose outline "gives an apple-core appearance", and—in a single sentence—pronounces "This is colonic carcinoma".



FIGURE 2.5: Apple-core lesion.

This process can be explained by dual-system theory in cognitive psychology, which holds that human thought is governed by two qualitatively different families of processes. In the medical literature they are usually called Type I and Type II reasoning; in Kahneman (2011)'s best-seller *Thinking, Fast and Slow*, they appear as System 1 (fast) and System 2 (slower)²⁶.

How does the cognitive shift work from deliberate, rule-based Type II reasoning (listing and excluding differentials) to fast Type I pattern recognition, as shown in the medical metaphors visual example? Years of experience consolidate countless images into a compact schema; the fruit metaphor functions as a retrieval cue that

²⁶This process is however not a new discovery. In particular, William James noted the contrast between train-like associative thought and stream-like reflective thought as early as 1890 (James et al., 1890).

activates this tacit archive almost instantaneously. Because the metaphor contains both surface shape and the underlying process (a once-intact cylinder now eaten away), it performs cognition by dictating how fast and on which channel the clinician thinks. The performative edge cuts both ways in metaphor: when the first frame that flashes up is the entrenched “war on cancer” metaphor, oncologists may default to aggressive “combat” interventions, whereas a gentler “surface-tension spread” image could cue watch-and-wait strategies instead.

Tversky and Kahneman (1974) showed that many everyday judgements rely on intuitive shortcuts. Stanovich and West (2008) later formalised the terminology Type 1/Type 2 and linked Type 2 performance to cognitive ability.

Empirical support for this metaphor comprehension dimension shows for example that deliberate metaphors can induce deeper thought, better retention of the underlying message, and even distinct neural responses related to theory-of-mind or domain switching. For instance, one hypothesis tested by Cuccio, Steen, et al. (2019) is that only deliberate metaphors fully activate embodied simulations of the source domain (e.g., metaphors about physical motion activate motor areas strongly if the metaphor is used deliberately, but not if it is processed mundanely). Another study with Chinese speakers found evidence that metaphors flagged as deliberate led participants to perform differently on interpretation tasks compared to non-deliberate metaphors, suggesting an added layer of processing when a metaphor is recognized as such (Wang et al., 2023b). Eye-tracking research has shown longer fixation times on novel metaphors in literary texts, interpreted as readers slowing down to interpret a deliberate metaphorical meaning. From an embodiment perspective, expertise compiles reusable, runnable schemas: once the apple-core pattern is mapped, the clinician can enact its causal-temporal organization (erosion, narrowing, failure of lumen continuity), which explains the fast, inference-rich leap from image to diagnosis. This is the same mechanism posited for novel metaphors: executing an imported event structure yields predictions and explanations in the target (Narayanan, 1999).

In conclusion, the past four decades of research on metaphorical meaning processing in natural cognition have revealed a phenomenon far more complex than early approaches suggested. While convergent evidence shows that context and conventionality reliably modulate processing, and that both bottom-up and top-down mechanisms are consistently implicated, the field struggles to integrate these findings into a unified framework. The persistence of multiple theories (CMT, SMT, CBT, etc.) reflects not just theoretical preference but fundamental challenges in mutual integration and subsequent empirical implementation. Because of that fragmentation, researchers disagree on what counts as “context”, how conventionality develops, and even where the literal/figurative boundary lies (Boyd, 1979; Kövecses, 2020; Zlatev, Jacobsson, and Paju, 2021). From the reviewed body of work, consensus seem to be on the fact that first of all, metaphor processing routes are not mutually exclusive: novel metaphors may require effortful comparison while conventional ones access category structures directly, yet both can involve parallel activation of literal and figurative meanings. About the distinction between literal and metaphoric, the evidence decisively refutes the classical sequential model where literal meaning must always be computed first. Salience, context, and individual differences modulate this process, making metaphor comprehension a dynamic interaction between bottom-up lexical activation and top-down contextual constraints. At the neural level, multiple cognitive systems are responsible for different aspects of metaphor processing. In this context, the theoretical frameworks proposed for

metaphor over the years increasingly appear as complementary perspectives on different aspects or timescales of metaphor processing. The evidence reviewed here suggests that methodological implications for metaphor research are thus that that single paradigms or measures cannot capture the full metaphorical phenomenon.

Additionally, this theoretical synthesis provides the foundations for computational approaches, suggesting that effective metaphor processing systems require both stable semantic knowledge structures and flexible online integration mechanisms. which will be presented in Section 3 of the next chapter.

As we will see, state-of-the-art neurosymbolic pipelines already require a hybrid substrate: (i) a stable, queryable knowledge layer of metaphor mappings and frames, and (ii) a dynamic integration engine that composes novel blends in context (Hicke and Kristensen-McLachlan, 2024; Rana, Chikkara, and Sharma, 2024). Until we understand how metaphor operates across levels and contexts, any system will be brittle: perhaps good at either conventional metaphors or creative ones, rarely both.

The goal is not to eliminate the discovered complexity but to work toward models that can predict and explain how all these factors interact in real-time comprehension, for example with suitable use cases. It is thus in this context that scientific metaphor can emerge as a privileged testing ground.

2.3 The prevalence and necessity of metaphor in science

Although different aspects come into play when we cognitively process metaphors, and although they are still not completely clear, one possible step to look at these phenomena more closely, especially those of conventionality and literal meaning access, is to look at the specific case where metaphors have been stabilized over years into conventional ones: those of the scientific type.

First of all, scientific and technical terms begin as live metaphors, then sediment into technical ones. That entire life-cycle is documented in articles and textbooks, which is ideal for modelling conventionalisation.

Secondly, in science, metaphors drive hypothesis-generation (seeing DNA as “text” prompted sequencing algorithms is a claim well documented in the literature; see, for example Searls, 1997). Throughout history, competing metaphors often coexist, so we can watch how one framing outperforms another (Turbayne, 1962b; Nersessian, 2008) or replaces the next one in diachrony.

Related to this aspect is the fact that, when we employ a metaphor, we also adopt a perspective on a problem (or frame a concept: §2.2.2.3. As whatever perspective we take dominates how we perceive the situation, and what possible actions we consider, sometimes a change of metaphor can be instrumental in solving a problem or proposing a new theoretical framework, especially in science: it has an heuristic power.

Turbayne (1962b) argued that viewing reality through different metaphors, or different models, increases our knowledge of reality, because taken together both metaphors provide more information about the reality than either one alone. This epistemic function of metaphor becomes particularly crucial in scientific discourse, where metaphorical frameworks can either constrain or enable theoretical breakthroughs. In this section, we examine how this heuristic power operates in scientific discourse, where the stakes of metaphorical choice are particularly visible and consequential.

Despite the standard view treats scientific metaphors as useful but ultimately dispensable pedagogical aids, the particular case of scientific metaphor and its crucial role for research and discovery have long been acknowledged (Black, 1962; Boyd, 1979; Hesse, 1965; Hoffman, 2018; Indurkha, 1992; Kuhn et al., 1979; Rothbart, 1984; Turbayne, 1962a). It is not surprising, then, that one corpus study reported an average of 15.6% metaphorical content words in academic prose (Do Dinh and Gurevych, 2016).

Metaphor is evidently very present in science. It suffices to think about all the metaphors we adopt at school when learning science concepts, or the metaphors used in political discourse. Scientific metaphors are also reflected upon in real time to comprehend their role in people's understanding of matters: in July 2020, during the COVID-19 pandemic, the World Health Organization was critical of the ubiquitous metaphor of multiple "waves" of the pandemic because it underplayed the role of human behaviour in controlling the spread of the virus. Even then, this did not stop politicians from using them (Semino, 2021). Thus, scientific metaphor remains inescapable, even when it employs technical or specialized domains.

Indeed, as Richards (1936) argued in analyses later developed by Black (1962), metaphor is far from being an exceptional deviation in language use. As Tietge (1998) shows, rhetorical and literary tropes are necessary components to a linguistic understanding of complex concepts, included scientific ones. We can thus say, to define scientific metaphors, that they are conceptual tools that systematically map knowledge to generate new theoretical insights, guide research programs, and structure scientific understanding in ways that cannot, in the same way, be expressed literally.

The specificity of scientific metaphor reveals three crucial distinctions. First of all, in scientific discourse, the difference between theory-constitutive (those essential to theoretical frameworks) and non-constitutive (those used merely for explanation) metaphors is particularly salient, yet this distinction often remains under-examined by users. Secondly, this aspect foregrounds metaphor's fundamentally relational nature. In science, where establishing relationships between phenomena is central to theory-building, this relational aspect becomes especially critical. Third, it is this very relational aspect that deals with the fine line between what was once metaphoric and currently literal that we really see the performative aspect of metaphor. Scientific metaphors can succeed or fail: their high stakes make them ideal for studying metaphor's cognitive and pragmatic functions.

In this chapter, we will examine all these levels, showing how scientific metaphor serves as a privileged site for understanding both the processing mechanisms and the epistemic functions of metaphorical language.

2.3.1 Constitutive and non-constitutive metaphors

In the philosophy of science and cognitive linguistics, a distinction is made between theory-constitutive metaphors and non-constitutive (or ornamental/explanatory) metaphors. A constitutive metaphor is one that fundamentally shapes a theory or conceptual framework: it introduces new terminology and ways of thinking that become integral to how we understand a domain (Boyd, 1979).

Boyd (1979) noted that theory-constitutive metaphors provide a way to refer to aspects of the world we suspect to exist but lack literal terms for; they allow scientists to introduce a terminology that serves as provisional reference for novel phenomena, even if initially in a vague or suggestive way. Over time, as knowledge advances,

these metaphorical terms can be refined and their referents better understood. Classic cases like the genetic “code” began as metaphor and, through theoretical development and measurement procedures, became constitutive of molecular biology’s conceptual toolkit (Kay, 2000).

By contrast, non-constitutive metaphors are those used mainly to explain or illustrate concepts without becoming part of the core theory. For instance, describing an electric circuit as “water flowing through pipes” can help a student understand electricity, but the field of electronics does not treat “electron water” as an actual component of its theory: rather, it is a pedagogical aid. The idea of stipulation ties in here: often a metaphorical expression can be stipulated (formally defined) as a scientific term, at which point it graduates from a loose analogy to an official concept. This happened with many scientific metaphors that started as vivid images and ended up as standard terminology once clearly defined. In this view there is a continuum from creative, exploratory metaphor to established, literal term: metaphors may start as exploratory comparisons and then crystallize into technical terms.

The process through which many scientific terms actually originate from everyday words, and often via metaphor, is a process that can be characterized as “semantic re-determination”. For example, in physics terms like “mass”, “force”, “wave” were originally common words or metaphors that were repurposed and formally defined in science. This process, also called “terminologization”, is also shown when disciplines borrow terms between each other (De Mauro, 1984; De Mauro, 2018; Serianni, 2011). In the famous linguist T. De Mauro’s view, scientific vocabularies emerge at the intersection of a “semantic dimension”, “pragmatic”, and “syntactic” of sign systems, and it is within this ecology of relations that a once-figurative label can acquire operational value in a field (De Mauro, 2018). This fits in with the sociocognitive and communicative turn in terminology studies, where terms are treated as dynamic, discourse-anchored categorizations (Cabr  Castellv , 1999; Temmerman, 2000; Huetter-Almerigi et al., 2025).

Within the pragmatic acts, as they organize use, we can treat definitions as a family of practices, each with a different function in inquiry and communication (De Mauro, 1988): (i) lexical (records current use of a word in a speech community); (ii) stipulative (introduces a technical sense for a term or coins a neologism); (iii) explicative (sharpens an already-used term by replacing a vague or everyday concept with a more exact surrogate suitable for calculation or inference)²⁷; (iv) ostensive (fixes meaning by pointing to procedures, measurements or paradigmatic cases); (v) regulative (sets rules for correct application inside a specialist practice). Terminologization narrows senses by stipulation and shared practice, much as a precisification makes a vague term exact for a given purpose; in scientific genres, that communal precisification is what secures the field’s rigor. In other words, performing scientific discourse starts by fixing technical terms via stipulative definitions and proceeds by providing explications, namely explicit criteria for use and truth, within a chosen domain of experience (*ibidem*, pp. 7-8)²⁸.

For what concerns metaphors in scientific language, they often appear at the entry points of conceptual construction but are domesticated through stipulation (to

²⁷This is very close to Carnap’s explication. On “explicative” definitions in the Carnapian sense: Carnap calls explication the replacement of an inexact pre-scientific concept (*explicandum*) by a new exact concept (*explicatum*), guided by similarity, exactness, fruitfulness, and simplicity (Carnap, 1950, pp. 3, 7). See also the overview in the Stanford Encyclopedia of Philosophy on explication, which reconstructs Carnap’s program and desiderata (Carus and Friedman, 2021).

²⁸See also De Mauro (2021, pp. 37-74), De Mauro (1990, pp. 89-90).

introduce a controlled sense) and explication (to sharpen an inherited term for calculation). In De Mauro's terms, the move from figurative usage to a *special language* ("lingua speciale") narrows the referential latitude of the sign and fixes its position within a specialist lexicon²⁹. It is a natural consequence that scientific language is one where maximum formality and explicitness are required. Metaphors can motivate and structure the explication, but the definitional act fixes how the term will function in the specialized language. Recent studies in philosophy of science recast this process as an abstraction path guided by metaphors that progressively scaffold model construction and category formation (Carrillo, 2023). Contemporary studies in science communication similarly show that while both constitutive and non-constitutive metaphors aid comprehension, their framing consequences must be monitored to avoid conceptual drift or misleading entailments (Reijnierse, Stee, and Mulken, 2025; Semino, 2021).

Too often, however, conventional metaphors slide into sort-trespassing: treated literally (*as if* they were literally true without re-determination), they smuggle in excess or inappropriate attributes from source to target, so camels become nothing but dogs, sounds nothing but vibrations, and the human body nothing but a machine (Turbayne, 1962a, p. 26). In this way, metaphor ossifies into a hidden metaphysics³⁰. When metaphors become constitutive, so does its framing effect and the relations it presupposes.

2.3.2 The relations of scientific metaphors

In Section 2.1.3, we argued that, in a way that has been long philosophically motivated, by putting emphasis on the relations between data and how we conceptualize the world, KPs are fundamental, cognitive-grounded mediator structures that help formalizing and extracting relations among in-use experiential patterns. In metaphor, similar innate structures are both very helpful and highly complex. The reviewed studies in the past sections also showed that a metaphor's meaning is not merely given from the juxtaposition between source and target, but from their *interanimation*, the relations between the two domains, and other extra-linguistic aspects.

Indeed, Black (1977) notes interaction is a metaphor itself and it is better to talk about not only the objects of the metaphor, e.g. the source and target domains, but rather as "system of objects" and it is this exact relationship between objects in the system, which emphasize or hide something we can define metaphor³¹.

The language of science is indeed so metaphoric that, according to Kuhn et al. (1979), even theory change must be expected to be accompanied by a change in metaphors and in the corresponding parts of the network of language similarities through which terms attach to nature. When geneticists found themselves in need of a new way to frame their understanding of genetics, the language adopted by

²⁹With this, De Mauro does not characterize scientific language as a separate one, but rather as a regulated use of the common language organized by goals, procedures, and explicit criteria (De Mauro, 1988, pp. 7-8).

³⁰This concept is in line with Nietzsche's constitutive reading of metaphor as opposed to the ornamental one; see for example his attack on philosophers who reduce reality to *Begriffs-Mummien*, conceptual mummies (Nietzsche, 1980a), herein truth develops through a process of abstraction via blurring differences and social convention, and finally through sedimentation in what is a mobile army of progressively worn-out metaphors (Nietzsche, 1980b).

³¹In the same way, and broadening the scope, E. Melandri talks about a "tropology": a system of metaphors, of theoretical constructs, and so on (Melandri, 1968, p. 546)

molecular biology, particularly the concept of the “genetic program”, came to assume many of the roles that had previously been attributed to the “action” of individual genes: a change in metaphor corresponds to a change of relations (and vice-versa).

Thanks to this interactive view, it is also possible to exemplify the higher heuristic power of metaphor. According to T. Boyd (1979), metaphors of the scientific field invite to explore similarities and analogies between sources and target that have not been fully discovered with respect to the relationships. This “inductive opening” ends when terminology becomes conventional.

Diagnostic practice, too, shows that what matters is not only resemblance but relational mapping, and clinicians train their senses around those mappings. Thus relations make the performative effects of scientific metaphor evident, which we will investigate in the next section.

2.3.3 The performative consequences of metaphorical choice

We begin with Melandri’s distinction between extensional and intensional (analogical) induction Melandri (1968): whereas the former generalizes from enumerated instances, the latter transfers structural relations across domains. In science, induction follows the tracks of exception and anomaly rather than normalcy.

In his *La Linea e il circolo*, resumed in introduction to H. Blumenberg’s *Paradigms for a metaphorology*, Melandri uncovers what he calls “a problem in a broadly moral sense” embedded within scientific metaphors (Melandri, 1968, p. 544). As we have seen, these metaphors do more than describe; they embody fundamental decisions about how to conceptualize reality. Following Aristotle’s concept of *dianoia* (intellectual virtue or practical wisdom), Melandri argues that every system of metaphors carries an implicit “moral”, not in the ethical sense, but as a fundamental orientation toward knowledge and action in a decisional sense (Melandri, 1969, pp. vii–viii). The question then becomes: what worldview does this metaphor advance?

Through language, we exercise a form of linguistic agency (Huetter-Almerigi et al., 2025): we recalibrate our grip on phenomena and on what actions appear available (see also according to pragmatics, which supposes we can *do* things with metaphors: §2.2.2.5). Metaphors work by altering salience (Ramberg, 2000; Hütter-Almerigi, 2020): they reorganize what stands out as relevant, which in turn guides modeling, experimentation, and communication: we call performative power of metaphor what causes this salience alteration³². In what follows, we trace these

³²Notice the relation with the Austinian tradition; (Austin, 1975).

effects³³ first in research practice, seeing how metaphors shape the content and conduct of inquiry; then in clinical discourse, discussing how metaphors shape interaction and uptake. This distinction matters: constitutive claims's efficacy tends to be evaluated by predictive payoffs, while the communicative one by comprehension; furthermore, a metaphor can be fruitful for modeling yet harmful in clinic, or clear for patients yet misleading for research design³⁴.

2.3.3.1 Constitutive effects in research practice

This claim finds historical resonance in the writings of physicist Niels Bohr, who, between 1926 and 1927, wrote: "The difficulties of quantum theory are connected with the concepts, or rather with the words that are used in the customary description of nature, and which all have their origin in the classical theories." (Bohr, 1985a, pp. 418–421). "These concepts leave us only with the choice between Charybdis and Scylla, according to whether we direct our attention towards the continuous or discontinuous aspect of the description" (Bohr, 1985b, pp. 462–463). With this text, Bohr was explaining an opening into the heuristic function of metaphors, asking for a redefinition of the analogy with the solar system and the introduction of scientific concepts whose peculiarity is in their being undefinable in the light of current knowledge. The concepts we inherit from earlier theories both enable unambiguous talk about experiments and force us toward particular descriptions.

Bohr's dilemma illustrates what we have described so far: the metaphors we use or inherit embody fundamental orientations toward knowledge that constrain future inquiry. The classical metaphors constituted the conceptual space within which physicists had to work, forcing them toward particular descriptions while rendering others unthinkable.

With this mechanism in view, consider how contrasting metaphorical framings in biology orient inquiry toward different ends.

Machine and computational metaphors exemplify one tendency towards a rationalizing approach. When we describe genes as "programs", cells as "factories", or organisms as "biological machines", we invoke a specific epistemology that privileges control, prediction, and engineering. These metaphors reduce phenomenological complexity to manipulable representations: organisms become assemblages of parts whose behavior can be reverse-engineered and redesigned. As Melandri warns, rhetorical figures ultimately express "not knowledge, but a certain will, project,

³³On an ecological view, what a situation "affords" is not a private mental property but a relation between an agent's abilities and features of the environment: an action possibility directly available in perception (Gibson, 1979; Chemero, 2003). In design, this becomes the idea of perceived affordances that cue use (Norman, 2008; Norman, 2013), while recent work in ecological psychology maps whole landscapes of affordances shaped by skills, practices, and sociomaterial arrangements (Rietveld and Kiverstein, 2014); see also Hutchby (2001) on how technologies both enable and constrain through their affordances. Read through this lens, metaphors in science and medicine are affordance-configuring devices: they reorganize which moves are seen as intelligible or apt in a field. Call this a metaphor's metaphoric affordance. Empirically, science-education research already speaks of the "affordances of metaphors" for making sense of processes (Pettersson, 2025), and framing studies show that alternative metaphors shift perceived options and priorities, even when literal content is held constant (Taylor and Dewsbury, 2018; Flusberg et al., 2024). On this picture, the performative effects we attribute to metaphor are the practical upshot of altered affordance structures: what lines of inquiry, interventions, or policies are rendered salient, licit, or urgent by a given figurative frame.

³⁴Here we continue to follow what Melandri states about science: it is not merely the ability to predict, otherwise our expectations would transform into norms; it is rather the ability to predict along with the constant expectation of the anomaly (Melandri, 1968, pp. 462–463).

or initiative of world domination: whose character is ultimately political and not scientific" (Melandri, 1968, p. 564). The synthetic biology movement, with its promise to *engineer* life from standardized biological parts, exemplifies this trajectory.

Abandoned the pretense of *tout-court* objectivity, in science we can proceed with an alternative approach, though more elusive to characterize directly, that in describing scientific objectivity maintains what Melandri calls "the synthesis of predictability and anomy" (Melandri, 1968, pp. 462–463), the capacity for prediction coupled with constant openness to the abnormal or anomalous. Rather than imposing mechanical schemas, metaphors in this mode (such as organisms as "autonomous systems"³⁵ or ecosystems as "webs of relationships") preserve the irreducible complexity and contextual dependency of living systems. This stance treats scientific objectivity not as mastery but as "the way through which we can make the alien intrinsic to ourselves" (*ibidem*), a form of understanding that acknowledges rather than eliminates otherness.

This theoretical contribution shows why metaphor choice shapes scientific practice. Machine metaphors in biology, for instance, actively orient research toward particular ends (control, standardization, commodification) while obscuring others (emergence, context-dependency, evolutionary contingency). The metaphors we (temporarily) choose prefigure both what questions we ask and what kinds of answers we accept. They are, in this sense, constitutive of scientific knowledge itself.

We can therefore say that the selection (but even the unconscious adoption) of scientific metaphors represents a fundamental epistemological and political choice about our relationship to the natural world.

The scientific literature has analyzed the performative effects of metaphor in different fields. As Perret and Longo (2016) have noted, applying information theory to biology belongs to a form of reductionism and the theoretical consequences enhance our position according to which information is not an appropriate notion for biology.

For instance, as Taylor and Dewsbury (2018) argue, while the metaphor of genes as "blueprints" has long guided molecular biology research, this view is deterministic, oversimplifies complex gene-gene and gene-environment interactions, and conflicts with recent advancements in developmental biology and epigenetics.

For another example, W. L. Johannsen introduced the term "gene" in 1909 (Johannsen, 1911) by combining two separate and seemingly incompatible metaphors. On one hand, a gene was seen as the fundamental building block of life, akin to the particle-like nature of atoms in physics and chemistry. On the other hand, it was also viewed as an active agent able to guide the development of specific traits. Indeed, the dominant metaphors within a given field of study can inevitably shift focus away from certain phenomena or dynamics that may suggest alternative explanations or characterizations.

Recently, Carver, Waldahl, and Breivik (2008) and Keller (2020) noted that genetic codes notion being the overevaluation of the genetic contribution on one's overall health as opposed to environment, and that non-coding DNA is useless, while in fact it plays crucial regulatory and structural roles.

2.3.3.2 Communicative effects in clinical practice

We now turn from research practice to clinical discourse, where metaphors shape reasoning, pedagogy, identity, and ethics. To make the performative landscape analyzable, we adopt a functional lens. The metaphorical nature of any scientific theory

³⁵ Varela (1981).

means that the theory creates its own ontological representation of reality, and a cognitive agent believing in a theory experiences the reality in terms of the concepts and structure of that theory. This means scientific metaphors have an impact on material and social worlds.

If we ask what metaphors actually *do* in scientific work, the effects line up into a handful of families. Some consequences are epistemic and heuristic, in that a metaphor scaffolds where to look, what hypotheses to entertain, and which models seem apt, especially in the case of constitutive metaphors. Some, tied to the contingent nature of non-constitutive metaphors, are operational and instrumental, while others are pedagogical and communicative. Still, others are social and identity-forming: a term can coordinate a community and mark boundaries. Finally, as we have seen, there are normative and ethical effects: metaphors carry values and frames that color risk perception and responsibility. All these functions, nonetheless, are a consequence of the performative character of metaphor itself, as seen in Section 2.2.2.5.

Bleakley (2017) builds on Peña and Andrade-Filho (2010)’s identification of four core functions of resemblance in medicine (naming, remembering, reasoning, and learning) by expanding them into a nine-function grid, as shown in Table 2.2. According to this framework, medical metaphors like the classic apple-core lesion for annular constriction of the colon (Figure 2.5) operate across multiple overlapping domains: they are cultural emblems, diagnostic heuristics, pedagogical tools, ethical provocations, and more. With the grid, Bleakley emphasizes metaphors shape how medicine is practiced, taught, and ethically reflected upon.

TABLE 2.2: Bleakley’s Nine Functions of Medical Metaphors.

Function #	Role of Metaphor
1	Emblematic of medical culture
2	Socializing ritual / professional identity
3	Imaginative transformation of medical culture
4	Diagnostic heuristic (pattern recognition)
5	Pedagogical mnemonic (learning aid)
6	Medical-cultural / historical marker
7	Encourages analogical biology
8	Ethical provocation / reflection
9	Explaining to patients (lay translation)

In the case of the apple-core lesion, the metaphor: (i) is emblematic of medical culture: knowing the term signals insider fluency within radiology; (ii) functions as a diagnostic heuristic: the core/hollow relation primes malignancy scripts in Type I reasoning (see Section 2.2.3.1); (iii) serves as a pedagogical mnemonic aid through the vivid food imagery.

Because this single metaphor maps onto multiple rows of the grid, it exemplifies how metaphor’s power in scientific discourse lies, once again, but in the dense relational networks it activates.

At the same time, the grid shows that we can only employ it if we analyze metaphors through their use, or action. Scientific language thus brings over a performative aspect that needs to take into account the distinction between how and where terms are borrowed and used over time, since as we said, metaphor in science at the very least influences the way we structure and build our social and material worlds (Cabassa, 2022). Through the frame-shifting effect (see Section 2.2.2.3),

metaphors can constrain our understanding of natural processes. They may uphold outdated scientific paradigms, leading to misunderstandings among the general public.

For instance, as Tate (2020) notes, war and machine metaphors script the clinical encounter as combat or maintenance; they hard-wire interruption of the patient and a preference for aggressive intervention. Recasting tumour spread as “surface tension” rather than “invasion” can, for instance, reduce stigma and psychological burden for patients living with cancer. At the same time, the study reports military metaphors are also often viewed as harmful, causing suffering to the patient, but empirical evidence does not conclusively support this claim. Food metaphors too, such as the example of the apple-lesion (with other examples as “strawberry tongue”, “anchovy-sauce sputum”) may fail or alienate outside their original culinary culture, exposing the colonial reach of supposedly benign metaphors. Inappropriate metaphors such as referring to a patient as a “vegetable” can dehumanize patients and cause distress to families (Langlume et al., 2020). Vagueness in the providers’ metaphors can lead to misunderstanding or poor understanding of communicated concepts (Rossi, Macagno, and Bigi, 2022).

Hauser and Schwarz (2015) have studied the effect of exposition to war metaphor frames to see if they could prevent people’s intention to limit their behavior to prevent cancer, for instance limiting sun exposition and consuming less meat. When patients read messages that mentioned cancer calls “hostile and aggressive”, people thought about prevention less with respect to the message that omitted these words (the neutral frame). In a longitudinal study on breast cancer, it was found that when patients chose war metaphoric frames as opposed to other framings, there was also association to higher depression and lower quality of life (Degner et al., 2003).

In the context of metaphors of illness, the notion of agency is central. As Semino et al. (2017) notes, if empowerment and disempowerment are defined as increase or decrease in the degree of the patient’s perceived agency, this manifests in the metaphors: e.g. in the perceived ability to control or react to events for one’s own benefit, where this ability is desired by the patient and not externally imposed.

When the patient feels disempowered, illness and treatment are perceived to fight the patient, who cannot “beat it”; biological and physical aspects of illness are emphasized at the expense of psychological and social dimensions; personal experiences of patient are marginalized, and in general they do not stimulate prevention intentions because focused on attack and aggression.

Instead, when the patient feels empowered, metaphors can stimulate a context of mutual support, cheer patients up, work better in critical stages of illnesses, can be used to describe the patient’s perceived difficulties inside the healthcare system, and can work in general to incentivize social and economic funds to medical research.

These dynamics of empowerment exemplify what we have already identified as the moral dimension of metaphor, which aims to fundamentally orienting our collective stance toward suffering, agency, and care.

As metaphors in contexts of illness are not only used by patients, but also by healthcare practitioners, friends, and family, having choices of how to attribute meaning to experience is fundamental. Especially in medical education, and despite their centrality in clinical reasoning and communication, metaphors remain underexamined. Students receive little formal training in their use (Kanthan and Mills, 2006), and a study by Rees, Knight, and Wilkinson (2007) analyze how metaphors shape perceptions of doctor–patient relationships, yet they rarely receive formal curricular framing.

For this reason, *the Metaphor Menu* by Semino has been launched as a project inspired by a range of different people who have experienced cancer³⁶. In the wake of a renewed interest about medical communication, another wave of recent studies have investigated how metaphors bridge the communication gap between providers and patients, aid in doctors' persuasion, but also better understand the patient's needs, help patients adapt to pain and improve quality of life, and open avenues of psychological exploration (Liu, 2025)³⁷.

In facing the discourse on whether wrong metaphors can exist or not, beyond academic writing, mainstream news articles have been written and continue to be written, especially during and after the COVID-19 pandemic, showing it is still an important point of discussion in medical communication³⁸.

2.4 The ultimate test case for computational meaning

In the last section of this chapter, we saw that the peculiar nature of scientific metaphors highlights some intrinsic features and issues of metaphors. On the one hand, scientific metaphor enables its being a cognitive tool for novel phenomena; on the other hand, it has a constraining function on paradigms. Thus we need metaphorical awareness in scientific practice, but we also need scientific practice to know what are the consequences of these metaphors on science and society. We have also seen that scientific metaphors can spark many interesting research topics that are still underexplored in contemporary accounts of metaphorical phenomena.

Scientific metaphors present unique epistemological properties that make them the ultimate test case for computational processing of meaning. First, they shift between literal, figurative, and technical status as scientific knowledge evolves. This nature is counterbalanced by their consequences: wrong metaphors in medicine tend to affect patient outcomes, while misleading metaphors in research can misdirect public policies and funding. Furthermore, to process scientific metaphors, it is necessary that a machine has information about the domain that is investigated. Even more so, when considering metaphors evolve in time and frame-shifting mechanisms can be tracked.

Being used across different domains and under different guises, scientific metaphors are particularly challenging to systematize. The same biological metaphor might function differently in medical, ecological, and synthetic biology contexts. Yet this heterogeneity provides ideal computational properties: scientific literature offers

³⁶<https://wp.lancs.ac.uk/melc/the-metaphor-menu/>

³⁷The interest in medical communication is not new. In a very interesting article, Donne (2020) examines doctor-patient communication in Hippocratic medical texts, identifying two contrasting models: one that involves minimal dialogue, bearing few questions from the physicians' point of view (thus an unbridgeable knowledge gap), and another that is idealized, in which physicians genuinely seek to learn from patients' experiences while teaching them. In the nature of ancient medical discourse, patients often relied on metaphors from everyday life, and at the same time the physicians were aware that effective treatment required genuine communication, prompting them to repurpose everyday language with specialized meanings. But Plato's *Laws*, in which the second model is presented, showed that while the aim was to envision medicine as fundamentally educational and collaborative, the reality was that time pressure, social hierarchies, and knowledge gaps often made such dialogue impractical. These dynamics persist to the present day.

³⁸See, for instance, (Fatehi et al., 2022; Brown, 2014). Following Wittgenstein, similarities in surface grammar can conceal differences in depth grammar, the roles expressions play in our language-games. When a term is extended across domains *as if* its depth-grammar carried over, we risk unhappy analogies: category mistakes that license bad inferences. In scientific discourse, this gives a principled sense in which metaphors can be "wrong": action-misleading mappings that systematically misweight salience or misfit the target practice (Yetman, 2025).

massive, structured corpora with explicit documentation of metaphor evolution, and multiple modalities enable cross-validation of genuine understanding. Especially if we adopt a semiotics perspective, according to which even conventionalized metaphors contain never fully exhausted meaning potential. The interest in scientific metaphors is confirmed by multiple accounts in the scientific literature that propose many open problems and gaps in research to align with contemporary scientific goals (Liu, 2025).

From a theoretical standpoint, studies show that to account for the different mechanisms of metaphorical processing, CMT (§2.2.2.1) and CBT (§2.2.2.3) are largely complementary: while the former addresses recurring patterns, the latter focuses on dynamic online processes, with conventional metaphors feeding the blending process. Therefore, while the comparison versus categorization debate has dominated processing discussions, CBT offers a framework that addresses both mechanisms in the dynamic integration of conceptual structures. Neural evidence confirms CBT's theory that integrating distant domains requires more conceptual work and that there is one general mechanism for meaning construction, not separate systems for literality and metaphoricity (Yang et al., 2013). Furthermore, it solves the debates around the limitations of the invariance hypothesis and the dominance claim (see §2.2.2.1): while traditional views assumed information flows mainly from source to target, Barnden et al. (2004) highlights that influence can, though less often, also run target to source, involving proposition transfer, query transfer, focus/highlighting and certainty adjustments. Despite these possibilities, Lakoff (2014) notes that psycholinguistic experiments often show unidirectional source-to-target effects. This apparent contradiction is resolved by CBT, according to which there are many different mechanisms for both domains to interact and contribute to an emergent blended space. Furthermore, CBT is compatible with DMT (§2.2.2.6) and semiotic theory (§2.2.2.7) as they all presuppose that complex meaning-making relies on the dynamic interaction between multiple representational spaces that interanimate in a blend. Therefore, we argue CBT is a particularly valuable framework for understanding the full spectrum of metaphorical processing.

Computational research can aid in these settings, particularly with the assistance of semantic web technologies, which provide stable but flexible structures for representing evolving metaphorical meanings and their mutual relations. From Melandri's contributions, as Cabassa (2022) notes, we can posit a more complex notion of structure, complementary to content: not merely a formal representational description, but an interrelation among facts, in a synthesis of the complementarity between form and content. In this view, and as defined at the beginning of the chapter (§2.1.3), KPs treat form as the relational abstract schema with its roles, constraints, and dependencies, while content consists of the concrete instances that populate that schema. Structure is therefore the shared and shareable network of relations established across multiple instantiations, which enables analogical mappings toward new domains. As a consequence, the guiding choices in the design of KPs are made explicit and, in doing so, renders visible the broadly moral orientation encoded in scientific metaphors.

Furthermore, this pattern-based approach would, yes, encode the cross-domain mappings identified by cognitive metaphor theory in a formal way, but could also be augmented with context slots and interpretive flexibility emphasized by semiotics. Despite maintaining the shared element, KPs are not strictly hard-coded and might be activated differently in different pragmatic circumstances. The framework is inherently open-ended, just as Eco described metaphor interpretation as tapping knowledge of all kinds.

To answer the RQ posed at the beginning of this chapter, namely *What is the role of KPs in developing systems that can move beyond fixed definitions to process the flexible and dynamic nature of meaning as found in human thought and language, and metaphors in particular?*, we conclude that KPs are supramodal, perspicuous layouts of roles and constraints that let systems move beyond fixed definitions by encoding frame-based structure while leaving meaning to be fixed by use.

In the view we propose, the KP is an interface (Hookway, 2014), functioning paradoxically as both a restrictive passage—a necessary conduit that constrains how we engage with technology—and a productive site where meaningful interaction emerges through use. This capability supports metaphorical mapping, blending, and multiple coexisting readings with provenance. Therefore, KPs act as reusable templates that enable flexible yet testable interpretations of metaphor.

In short, KPs reconcile the formalist view (that we need explicit structures and rules) with the cognitive view (that meaning is grounded in bodily and imaginative structures) and the semiotic view (that meaning emerges in context through interpretation).

To conclude, this chapter has set the conceptual stance of the thesis: scientific metaphors are action-guiding, so any account of meaning in this sense should (i) model analogical mapping; (ii) use frames; (iii) consider pragmatic intentions/deliberateness and (iv) study how metaphors alter salience and direct attention. In scientific discourse, canonical genres and review norms intentionally suppress speaker-specific intentions in favor of communal warrant, which makes (iii) comparatively underdetermined at the textual level we analyze. Likewise, frames and measurement conventions pre-stabilize relevance, so any claims about (iv) would require audience-level or experimental evidence (uptake, task effects) beyond the scope of this thesis.

Chapter 3

Computational solutions for metaphor processing

Metaphor has long been considered a distinctly human cognitive ability, making its computational processing a compelling challenge for AI research.

In this chapter, we show that metaphor’s resistance to algorithmic processing, even in state of the art systems, stems not from its complexity alone nor from the realization of U. Eco’s prophecy that no algorithm exists for metaphor (Eco, 1986), but from its fundamental reliance on the very interrelational patterns that knowledge engineering seeks to formalize.

The aim of this chapter is thus to answer to the following RQ: *What are the computational requirements to study the phenomenon of metaphors, either generic or scientific?*

The difficulty to handle metaphor computationally but also to reach a consensus theoretically, and the emergence of multiple scientific views that put forward as many alternative metaphors to established ones in different fields, constitutes what we call a “metaphor crisis”. It is only through a comprehensive view of the outstanding issues on the matter of metaphor that it is, in fact, possible to find ways to analyze this phenomenon. To conclude the chapter, we discuss the fundamentally bidirectional relationship between AI and metaphor (Section 3.2): how computational approaches have attempted to model metaphorical reasoning, and conversely, how metaphor serves as a crucial test case that exposes the limitations of current AI architectures. This reciprocal relationship suggests that advancing computational metaphor understanding is not merely about building better NLP systems, but rather about developing AI capable of the innate analogical reasoning and frame-shifting capabilities that characterizes human thought.

3.0.1 Metaphor datasets

The first major attempt at developing conceptual metaphor datasets was the Master Metaphor List (Lakoff, Espenson, and Schwartz, 1991), which compiled over 791 conceptual metaphor mappings, creating the first comprehensive evaluation benchmark. Mason (2004)’s CorMet system represented the first large-scale corpus-based approach to metaphor extraction, dynamically mining Internet corpora using selectional preference patterns. The development of reliable annotation schemes proved crucial for creating high-quality metaphor datasets. The Pragglejaz Group’s MIP (Metaphor Identification Procedure) (Group, 2007) provided the first explicit, systematic method for identifying metaphorical word usage. MIPVU (Metaphor Identification Procedure VU University), developed by Steen et al. (2010), refined and extended MIP with more detailed guidelines for borderline cases.

Early benchmark datasets contributed significantly to computational metaphor research, particularly for supervised detection and cross-lingual evaluation. The TroFi Example Base (Birke and Sarkar, 2006b), MOH-X (Saif Mohammad, 2016), and the Tsvetkov SVO dataset (Tsvetkov, Boytsov, and Gershman, 2014) have provided annotated instances of metaphorical and literal sentences drawn from news and parallel corpora. MetaBank (Dunn, 2013) and manual annotations from the British National Corpus (Shutova, 2010b) offered foundational resources for mapping metaphorical source–target domains and quantifying metaphor frequency. Additionally, efforts to leverage FrameNet and PropBank for metaphor annotation, though not their primary function, have intersected with semantic predicate research and semantic role labeling (Baker, Fillmore, and Lowe, 1998; Palmer, Gildea, and Kingsbury, 2005). The VU Amsterdam Metaphor Corpus (Krennmayr and Steen, 2017) became the field’s primary benchmark, containing approximately 190,000 lexical units from the BNC-Baby subset. The LCC Metaphor Datasets (Mohler et al., 2016) represented a leap in scale and linguistic diversity.

Recent developments have emphasized multimodal and multilingual expansion. The MultiCMET dataset (Zhang et al., 2023a) provides 13,820 text-image pairs from Chinese advertisements, representing the first large-scale multimodal metaphor dataset in Chinese. The MUNCH (Metaphor Understanding CHallenge Dataset) (Tong et al., 2024) provides over 10,000 paraphrases plus 1,500 inapt paraphrases, representing the first comprehensive benchmark for evaluating large language model metaphor understanding.

Furthermore, multimodal metaphor processing has emerged as a crucial frontier. The MET-Meme dataset by Xu et al. (2022), with more than 10,000 text-image pairs, enables cross-modal metaphor analysis. The Pittsburgh Ads by Hussain et al. (2017) was a large-scale dataset of advertisements with rich annotations for topic, sentiment, persuasive questions/answers, and symbolic references, though it is not explicitly about metaphors. Zhang et al. (2021) released MultiMET, a multimodal metaphor dataset that is annotated for metaphor occurrence, the source and target domain, the sentiment expressed, and the author’s intent. The dataset is not publicly released yet. MET-Meme was introduced by Xu et al. (2022) to propose a multimodal large-scale meme dataset in two languages. Each meme is labeled for metaphor presence, the source and target concepts of the metaphors, the sentiment, intended communicative intent, and offensiveness level. MET-Meme has been publicly released, though this binary approach to metaphor labeling has been criticized given theoretical arguments that all memes are inherently metaphorical (Scott, 2021; Dancygier and Vandelandotte, 2017). This raises questions about what the negative cases actually represent—memes without explicit metaphorical framing, or simply less salient metaphors—and whether the task is well-defined for computational modeling.

To foster richer research in visual metaphors, the MetaCLUE task set was introduced Akula et al., 2023, though its code and dataset are not publicly available.

IRFL by Yosef, Bitton, and Shahaf (2023) focuses on connecting textual metaphors to images. IRFL includes collections of idioms, similes and metaphoric phrases paired with candidate images. The majority is indeed idioms, while metaphors are 35. Furthermore, idioms also tend to be obsolete, as they come from crowd-source MAGPIE lists. Images are retrieved automatically so it might not be canonical figurative depictions.

Metaphoric compositions linked by an annotated attribute are available in the ELCo Dataset proposed by Yang, Zhang, and Miao (2024), where metaphorical composition in emojis is evaluated via a textual entailment task.

The Video Metaphor Corpus was released by Alnajjar, Hämäläinen, and Zhang (2022), and contains short video clips expert-annotated for metaphors. More recently, Kalarani, Bhattacharyya, and Shekhar (2024) introduced a Video Metaphor Captioning dataset, comprising advertisement videos with human-written metaphorical captions. This dataset is, at the time of writing this thesis, yet to be released.

Beyond modality, datasets are also built on the basis of task type. Recent datasets have operationalized entailment through various evaluation paradigms: binary classification tasks that assess whether expressions like “he absorbed the costs” entail “he paid for the costs” (Stowe, Utama, and Muresan, 2022), multiple-choice assessments where models select correct explanations from candidate options (Chakrabarty et al., 2022), and contradiction detection tasks that require generating explanations for why figurative and literal statements conflict (Mohler, Tomlinson, and Bracewell, 2013). The FLUTE dataset (Chakrabarty et al., 2022) provides 9,000 figurative natural Language Inference (NLI) instances, which evaluate whether a hypothesis logically follows from, contradicts, or is unrelated to a given premise, with explanations spanning metaphors, similes, and sarcasm, while IMPLI (Stowe, Utama, and

Muresan, 2022) contributes approximately 900 metaphorical entailment pairs focused on idioms and metaphors. This evaluation paradigm has recently extended to multimodal contexts through V-FLUTE (Saakyan et al., 2025), which frames visual figurative language understanding as explainable visual entailment, demonstrating the broader applicability of entailment-based evaluation for computational metaphor processing. Finally, Meta4XNLI (Sanchez-Bayona and Agerri, 2025a) extends the NLI framing to a parallel multilingual setting on English and Spanish by annotating NLI pairs for metaphor at token level and for interpretation at the premise–hypothesis level, enabling cross-lingual tests of whether inference depends on metaphor and how translation affects metaphor preservation.

3.0.2 Computational approaches to Conceptual Blending Theory

Early computational approaches to CBT (§2.2.2.3) established theoretical foundations that inform current work. Goguen’s algebraic semiotics provided one of the first rigorous mathematical models of conceptual blending, representing metaphors as combinations of structured sign systems and formally specifying how conceptual domains can be merged (Goguen, 1998). In Goguen’s work, concepts were modeled as sign systems and defined blends as mappings (“morphisms”) between these systems. In this way, metaphors could be represented as combinations of sign systems, providing a precise category-theoretic foundation for how two conceptual domains can be merged into a blend. This approach brought mathematical rigor to CBT as the original theory was formalized to measure how well a blend preserves source structures, showing that algebraic semiotics could be systematically analyzed and evaluated for quality of mappings.

Building on these formal foundations, CBT was extended into ontological frameworks, developing methods for algorithmically merging distinct domain ontologies to create novel hybrid concepts through shared structural constraints (Confalonieri and Kutz, 2020; Bhatt, Hois, and Kutz, 2012). In this way, such constraints could be merged into a new ontology that exhibits emergent structure. Unlike standard ontology matching or integration, ontological blending in this view aimed at creative combination of distant domains and allowed divergent elements to combine, producing hybrid concepts. Other work by Confalonieri et al. (2016) and Confalonieri et al. (2018) extended this framework with description logic formalisms and refinement operators to automate the search for useful blends. For instance, they developed a general computational framework for conceptual blending that can take formal ontologies as inputs and generate candidate blends through logical constraints to ensure coherence. These works thus show CBT can be used computationally for concept invention, keeping the focus on creating new concepts.

Veale’s computational models pioneered practical implementations of CBT, particularly in the field of metaphor and analogy. Veale and O’Donoghue (2001) and Veale (2019) were in fact among the first to propose a working computational blending model. They extended Veale’s earlier metaphor interpretation system that found analogical connections between domains to better align to CBT. This extended system reframed metaphor comprehension as a blending process: two input domains could be systematically mapped and merged, with Sapper’s network of conceptual associations providing the relational structure for the blend. However, as early adopters of CBT, Veale and O’Donoghue (2001)’s implementation only scratched the surface of CBT-inspired possibilities: later analyses noted that their system did not fully embrace the emergence of an independent blend space as envisioned by Fauconnier and Turner, since the blend remained embedded in one input domain’s

terms rather than forming a novel structure of its own. Nonetheless, this work was fundamental as it showed that a conventional metaphor interpretation engine could be adapted to perform conceptual integration.

Veale continued to champion CBT in creative computing, such as identifying practical forms of blending that computers could tackle. One approach was through what were “conceptual mash-ups”, blends generated by creatively reusing common-sense knowledge. This line of work maintained a strong generative creativity focus.

All the above approaches share certain limitations and leave open important challenges. First, CBT is complex to implement computationally, and translating its full richness into algorithms has proven difficult also because of the context-dependent nature of creative blends. Researchers note that the CBT framework is extremely abstract and generic (Pereira and Cardoso, 2002), so implementations of it must simplify or constrain¹ the theory, which means they tend to capture only aspects of what human blending can do. Additionally, there is still no general, domain-independent blending engine that can take undetermined concepts and produce a meaningful blend like humans can. This ties to the challenge of search and selection in blending: even moderately-sized knowledge bases can produce a combinatorial explosion of possible blends (as shown in the work on metaphors of Gangemi, Alam, and Presutti, 2018, see Section 3.0.3). Also, identifying the appropriate generic space is non-trivial and can require common sense that is hard to formalize (Confalonieri and Kutz, 2020)². Current systems often rely on heuristic optimality principles to prune the blend search spaces, but these are imperfect as automating the discovery of the right mapping between domains remains an open research problem. Finally, these problems are all the more compelling when we want to study a specific type of conceptual blending, that of metaphorical blending, which is more complex.

A clear limitation of the outlined approaches lies in evaluating and filtering blends for quality and coherence. Blending can yield many bizarre or useless combinations if not guided properly. However, this task requires what is for a human innate pragmatic and contextual knowledge. Confalonieri and Kutz (2020) highlight concept evaluation as a core unsolved aspect: determining the value or meaningfulness of newly invented blends algorithmically is still largely unsolved. In the same line of thought, Barnden (2008) observed that proposed metaphorical mappings in research are rarely evaluated for coverage, coherence, or side-effects. Because novel metaphors are relatively rare in natural discourse, researchers reuse familiar mappings, which can mask plausibility failures. In response to this plausibility challenge, recent work has explored using social choice theory or analogical coherence measures to rank blends (Confalonieri and Kutz, 2020). The issue, though has not been directly applied to metaphors, is particularly crucial for metaphor understanding: if a system produces multiple interpretations of a metaphorical phrase, how to decide which one is plausible³. CBT itself does not provide a straightforward quantitative metric for a good blend beyond informal optimality principles. Indeed, in theory, anything can be related to anything else according to the semantic infinity hypothesis (Hoffman and Honeck, 1976): the act and very purpose of making an analogy is to create new connections.

¹By which we mean to operationalize. See Chapter 5 for the meaning and its application in metaphor studies.

²We will specifically focus on this challenge in Section 5.3.

³And vice versa, how to decide when a monstrous metaphor is implausible?: it “has the following characteristics: (1) creating a monster analogy can drive the creator insane, (2) no existing computer system can accept a monster analogy as input without relying on redescription based on some ad hoc criteria, or (3) no existing algorithm can process a monster analogy. However, monster analogies are not necessarily difficult for humans to understand.” (Hoffman, 1995, pp. 20-21).

Specific to metaphor detection, a current research gap is the integration of CBT with modern computational linguistics techniques. Until recently (See Section 3.0.6), blending-based models did not leverage large-scale corpora or machine learning effectively, making them less robust for real-world language data. They often required curated ontologies or knowledge bases, limiting coverage. Meanwhile, data-driven metaphor detectors lacked the explanatory power of CBT, which also requires the relational structure that is so fundamental for metaphors.

The need for evaluation is also tied to the one for comprehensive benchmarks to assess blending systems on tasks like metaphor interpretation, to systematically compare CBT-inspired models with other approaches. In summary, while significant progress has been made in formalizing and using conceptual blending in AI, current approaches are still limited in scope, scalability, and metaphor-specific capabilities.

Thus, while these approaches established crucial theoretical groundwork, they primarily focus on creative concept invention and generation rather than metaphor detection and understanding. In the original theories, CBT treats metaphors as one instance of the broader integration process, but early AI implementations of CBT gravitated toward deliberate creative tasks rather than identifying metaphors in data. Indeed, computational approaches to CBT largely overlooked metaphorical applications, likely because blending theory’s complexity made it challenging to apply for the existing metaphor datasets and tasks. Overcoming these limitations will require dedicated datasets, techniques, and evaluation methods, which we will propose in Chapter 5.

3.0.3 Knowledge representation of metaphor

Early knowledge representation efforts handled metaphors in isolation, using SWRL rules to derive new facts from fixed triples (Hamilton, 2021; Mitrović et al., 2017).

The most famous metaphor knowledge base is the MetaNet project, which created a multilingual repository of conceptual metaphors anchored in FrameNet frames⁴. Alongside MetaNet, FrameNet has independently evolved to include metaphorical frame relations, supporting construction grammar approaches to metaphor understanding across languages and genres.

Newer lines of work align metaphors with frame resources: Panayiotou (2024), for example, links cross-domain figures through bibliographic ontologies. Panayiotou’s model incorporates the FaBio bibliographic ontology (Peroni and Shotton, 2012) to capture source information (e.g. literary references for a metaphor) and the OWL-Time ontology (Pan and Hobbs, 2005) to encode temporal aspects of metaphor usage. By introducing concepts for the source domain and target of a metaphor and their shared qualities, this ontology allows metaphors from different domains to be interconnected and queried in a uniform way.

At the same time, by modeling metaphors as fixed source–quality–target triples, this work captures only a static snapshot of such inherently fluid and context-sensitive phenomena. This rigidity limits the ontology’s ability to reflect variation in metaphor interpretation across texts, cultures, or communicative intent. Moreover, the representation of shared qualities has limited semantic depth and not grounded in formal property ontologies, which restricts inference and semantic querying. The lack of integration with major lexical and frame-based resources (e.g., FrameNet, WordNet, OntoLex) further reduces interoperability and hampers its use in natural language processing tasks. Since in the framework proposed in the paper metaphor entries

⁴<https://metaphor.icsi.berkeley.edu>

must be manually curated, the ontology also suffers from limited scalability and coverage, making it impractical for real-time or large-scale applications. Finally, the model does not support pragmatic or cognitive dimensions such as conventionality, novelty, or conceptual blending, which are crucial for modeling how metaphors emerge, evolve, and are understood in human cognition.

Martín-Chozas, Diab-Lozano, and Montiel-Ponsoda (2024) bring metaphor into Semantic Web terminology modeling. In the work, they present Termlex, a semantic modeling framework that bridges cognitive terminology research and Linked Open Data. They begin by extracting candidate metaphorical terms from specialized corpora such as cybersecurity or biotechnology texts using dedicated tools. Each term is analyzed to identify its source domain, target domain, and the key qualities transferred from one to the other. These components are then formalized within a custom OWL vocabulary (built upon OntoLex-Lemon and SKOS), allowing metaphorical mappings to be captured as semantic entities rather than mere lexical annotations. The authors further enrich these entries with sociocognitive metadata such as usage examples, provenance, and domain context, thus enabling complex queries about how metaphors are structured and evolve within specific discourses.

As a semi-automatic system and extended ontology, Termlex provides a flexible platform for representing metaphorical transfer in specialized domains. However, like many such frameworks, it remains limited by its reliance on manual annotation, lack of in-use modeling, and absence of mechanisms for dynamic inference or integration with broader lexical-semantic networks.

Amnestic Forgery (Gangemi, Alam, and Presutti, 2018) is an ontology that integrates MetaNet's metaphor data into the Semantic Web ecosystem. It extends the Framester knowledge graph to include MetaNet's metaphors. In fact, interoperability with large lexical semantic networks, such as WordNet or BabelNet, through integration within knowledge graphs like Framester, is an ongoing research focus to enhance semantic richness and inference capabilities for metaphor representation. In Amnestic Forgery, MetaNet's conceptual metaphors are converted into OWL/RDF format and aligned with the existing Framester frame ontology, making them queryable with standard semantic tools. The resulting resource contains 642 recorded metaphor mappings, connecting about 270 source frames to 221 target frames (i.e. 642 distinct source-target frame pairs). However, despite cataloging hundreds of metaphors, the authors note that such a static taxonomy is inherently incomplete, since new metaphors constantly emerge in usage; a property of Metanet also noted by Lieto, Pozzato, and Zoia (2025).

In fact, Lieto, Pozzato, and Zoia (2025) recently presented a system called MET^{CL} able to perform metaphor generation and classification by applying a formal operationalization of the CBT. The core of MET^{CL} is a reasoning framework specialized in human-like commonsense concept combination: Typicality-based Compositional Logic (T^{CL}), which is able to account for the composition of prototypical representations.

The reviewed works on metaphor representation highlight that formalizing metaphors is complex: metaphors are creative and context-dependent, so any fixed repository will inevitably miss novel or unconventional metaphors. However, despite having laid important groundwork, static ontologies alone cannot capture the endless creativity of metaphor. They can provide structured semantic representations of metaphoric mappings which can be leveraged to improve metaphor understanding, but they are increasingly seen as components to be complemented by adaptive, context-aware processing systems rather than standalone solutions. Therefore, researchers argue for hybrid approaches that combine these structured repositories

with dynamic systems (e.g. neural language models or reasoning algorithms) to handle metaphors on the fly.

3.0.4 AI for computational metaphor processing

With the rise of AI, increasing attention has been devoted to computational approaches to metaphor processing. In the following section, we review key developments in this area and highlight the main directions of current research.

3.0.4.1 Early approaches

Early work in computational metaphor relied on symbolic AI and knowledge-based techniques, often based on linguistic and cognitive theories⁵. A classic strategy is to detect metaphors via selectional preference violation (Dodge, Hong, and Stickles, 2015). For example, the system of Fass (1991) identified metaphors by finding semantic anomalies, such as animacy or thematic-role violations (e.g. “the car drank gasoline” violates the expectation that only animate agents drink). Such rule-based systems encode world knowledge (often via lexical ontologies such as WordNet) to flag words used outside their literal selectional constraint. Another rule-based technique uses hand-crafted lexical patterns. For instance, Dunn et al. (2014) designed syntactic patterns to extract candidate metaphoric phrases, then applied classifiers with features like concreteness norms and semantic overlap to decide if an expression is metaphorical. These studies leverage the intuition that a metaphor involves describing an abstract target concept using language more typically associated with a different, concrete domain.

Symbolic approaches have also tackled domain-specific metaphors by incorporating domain knowledge. For example, Perlerin, Beust, and Ferrari (2003) developed an NLP system to detect metaphors in stock market news by first cataloging conventional metaphorical mappings in that domains and then searching texts for those patterns. Another approach, CorMet (Mason, 2004), automatically derived domain mappings by comparing domain-specific corpora to improve precision, still under the assumption that specialized domains have characteristic vocabulary and metaphorical themes.

An interesting approach is found in Barnden et al. (2004), and criticizes the Dominance hypothesis (see Section 2.2.2.1) supposing that computational methods do not handle multiple ways of information processing. The system proposed instead is ATT-Meta, which uses rule-based, uncertainty-aware goal-directed reasoning, and can handle forward and reverse transfers for propositions, questions, and certainty levels. This method is different from previous AI systems such as MetaBank (Martin, 1994) or KARMA (Narayanan, 1997) that implement metaphor understanding by transferring propositions or features from a source domain to a target and do not compute reverse mappings. Systems also such as that of Hummel and Holyoak (2001) also treat transfer modelled in a forward direction.

While initial systems focused on selectional preference violation, later early work incorporated broader discourse context and syntactic features to improve metaphor detection performance. For instance, Jang et al. (2015) demonstrated the importance of combining local sentence-level and global discourse context via lexical chains and frame-semantic features. Many symbolic approaches also leveraged lexical-semantic resources such as WordNet, VerbNet, and FrameNet to encode world knowledge

⁵Here we mostly focus on early approaches from the early 2010s onwards; for a detailed account of the reviewed literature before then, see Barnden (2008).

and fine-grained predicate-argument expectations, which remain central pillars in computational metaphor analysis (Ma et al., 2020).

3.0.4.2 Embedding-based methods

In NLP and multimodal learning, an embedding is a numeric vector that represents a linguistic or visual unit—such as a word, phrase, frame name, or image—in a continuous space where geometric relations correlate with semantic relatedness. Classical static embeddings assign a single vector to each lexical type and thus conflate senses (e.g., Word2Vec, GloVe, fastText; see Mikolov et al., 2013; Pennington, Socher, and Manning, 2014; Bojanowski et al., 2017), whereas contextual embeddings produced by Transformer models assign a different vector to each token occurrence, conditioning the representation on its sentential context (e.g., ELMo, BERT; Peters et al., 2018; Devlin et al., 2019). Multimodal variants align text and images in a joint space (e.g., CLIP; Radford et al., 2021). These spaces enable operations such as nearest-neighbour search, clustering, and linear projections that can act as weak semantic signals when explicit knowledge is missing. In this chapter, embedding-based methods denote techniques that score or discover metaphor primarily from vector-space evidence, optionally combined with frame- or lexicon-level knowledge. One widely used strategy relies on deviations from learned selectional preferences: if a predicate-argument pair’s joint representation lies far from the predicate’s typical literal neighborhood, the usage is atypical and hence a candidate metaphor (Resnik, 1997; Shutova, Teufel, and Korhonen, 2013). A complementary approach exploits contextual representations, comparing a token’s contextualised vector in situ with a type-level or corpus-derived prototype; large contextual shifts signal incongruity consistent with metaphoric usage (see, for example, contextualised-embedding approaches to metaphor detection surveyed in Shutova, Teufel, and Korhonen, 2013 and extended with modern encoders).

Embedding-based methods are attractive because they are broad-coverage, inexpensive, and sensitive to graded regularities such as plausibility windows and directional dominance between domains. At the same time, similarity is not plausibility: a short geometric distance does not guarantee role compatibility, and raw vector spaces lack the typed, role-level structure needed for explanatory accounts of source-target projection. They are also sensitive to corpus artefacts, anisotropy, and bias, and can be unstable across domains and checkpoints (Ethayarajh, 2019).

3.0.4.3 The LLM era

The limitations above concern raw vector spaces used directly as task models. Modern large language models inherit the same representational substrate—token embeddings in high-dimensional spaces—but update these vectors layer by layer using attention, producing context-dependent representations at every position. In this sense, LLMs are not a departure from embedding-based methods so much as deep, context-sensitive embedding generators. This shift mitigates some problems of fixed spaces.

Large language models are neural language models trained to predict the next token in a sequence (Brown et al., 2020; Radford et al., 2019). They extend classical language models that estimate the probability of word sequences by using deep learning with very large parameter counts and large scale training corpora. Earlier approaches such as n-gram models rely on fixed context windows and struggle

with rare events and long range dependencies (Jurafsky and Martin, 2023). In contrast, modern large language models learn contextual representations that adapt to surrounding text and can generalize across domains without task specific feature engineering.

Large language models are most often built with transformer architectures. Self attention provides a way to compute context dependent representations by allowing each token to attend to other tokens in the sequence (Vaswani et al., 2017). The effectiveness of these models depends on the learned parameters that govern attention, representation, and prediction. A full treatment of transformer variants is beyond the scope of this chapter; the following overview summarizes the two model families most relevant for common language processing workflows.

Encoder and decoder based transformers include a stack of encoder layers and a stack of decoder layers. The encoder builds bidirectional contextual representations of the input. The decoder generates the target sequence step by step while attending to both the previously generated tokens and the encoder outputs through cross attention. This sequence to sequence setup is well suited to tasks that map an input string to a target string such as summarization, translation, and question answering (Vaswani et al., 2017; Raffel et al., 2020; Lewis et al., 2020a). Inputs and outputs are handled by distinct components, which encourages stable behavior when inputs are long but outputs are short and structured.

Decoder only transformers are autoregressive generators. They read a single concatenated stream that may contain instructions, optional demonstrations, and the place where the answer should appear. The model builds a representation layer by layer and predicts the next token conditioned on the prefix. This design is natural for open ended generation and instruction following, and it also supports classification by asking the model to print a label in a prescribed format (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023c; Touvron et al., 2023a). In this sense LLMs are thus said to be generative.

Either family can be used with prompting or with fine tuning. Decoder only models are convenient for instruction following in a single pass. Encoder and decoder based models are convenient when the objective is to compress a long input into a compact target string (Chung et al., 2022; Ouyang et al., 2022).

In context learning refers to the ability of large language models to adapt their behavior based on examples and instructions placed directly in the prompt. The prompt can include natural language instructions, a schema for the desired output, and optional demonstrations that show how inputs relate to outputs. This allows users to specify new tasks without additional gradient based training (Brown et al., 2020; Liu et al., 2023).

Zero shot prompting provides only an instruction and any necessary constraints on format; it is used when inputs are long or when consistent formatting is more important than domain adaptation (Brown et al., 2020; Liu et al., 2023).

Single shot prompting adds one input and output demonstration before the real query, which helps the model align to the desired style and output schema while keeping the prompt short (Brown et al., 2020).

Few shot prompting uses several demonstrations to shape model behavior on ambiguous or domain specific cases. It often improves robustness but increases prompt length and may require care in selecting representative examples (Brown et al., 2020; Min et al., 2022).

Reasoning traces can be requested as short rationales when helpful, but free form explanations should be treated as intermediate guidance rather than ground truth

(Wei et al., 2022; Kojima et al., 2022). Later chapters apply these ideas more concretely.

Pretraining exposes the model to diverse text so that it acquires broad linguistic competence. Adaptation after pretraining can proceed in several ways. Instruction tuning uses curated input and response pairs to make the model follow instructions more reliably (Chung et al., 2022; Ouyang et al., 2022). Supervised fine tuning on task specific data specializes the model to a narrow objective with relatively few updates. Parameter efficient methods such as adapters or low rank updates reduce memory and storage requirements by training a small set of additional weights while keeping the base model fixed (Houlsby et al., 2019; Hu et al., 2022). Retrieval augmented prompting can supply short definitions or factual snippets at inference time without changing model parameters, which is useful when prompts need external context (Lewis et al., 2020b). In fact, this pattern is widely referred to as *retrieval augmented generation* (RAG): a retriever queries an external index or knowledge graph, the retrieved passages are inserted into the prompt, and the generator conditions on them, grounding outputs in verifiable sources rather than relying only on parametric memory (Lewis et al., 2020b). We use “RAG” for the architecture and “retrieval augmented prompting” for the practice of supplying the retrieved snippets in the prompt.

Performance of LLMs depends on model size, training data coverage, and the match between prompt design and the target task. Longer prompts consume more computation and may limit context available for inputs. Output formatting should be specified clearly to reduce ambiguity. Even when models produce fluent text, factuality and calibration can vary by domain, so evaluations should reflect the intended use and the uncertainty inherent in human judgments. Sampling choices such as temperature influence fluency and diversity at inference time (Holtzman et al., 2020).

3.0.4.4 Theory-integrated metaphor detection and understanding

Recent advancements in metaphor detection have increasingly incorporated theoretical frameworks into computational approaches. Studies leveraging CMT have shown improved baseline results in metaphor identification (Mao et al., 2023; Tian, Xu, and Mao, 2024). For metaphor interpretation, various methods have been developed to discern the source domain implied by the metaphor, given a target domain. These include unsupervised techniques (Shutova, 2017), deep learning-based approaches (Rosen, 2018), and methods using LLMs (Wachowiak and Gromann, 2023). However, these methods generally depend on annotations of a single domain rather than exploiting unlabeled data. In Hicke and Kristensen-McLachlan (2024), when unlabeled data was exploited, the MIP was preferred as a theoretical framework, thus the metaphors were annotated by category but not by source and target domain. Indeed, the authors recognize this limitation, stating that with the MIP (and following models such as the Deliberate Metaphor Identification procedure), forcing the models to annotate word-by-word makes it challenging for them to identify metaphors comprising multi-word units. Other research focuses on extracting properties linking the source and target domains to interpret metaphors (Su, Huang, and Chen, 2017; Rai et al., 2019; Su et al., 2020). While these approaches yield promising results, they often do not adequately capture the underlying theoretical reasoning processes that are based on the definition of their structural similarity in terms of abstraction and relational dependency on the two metaphorical domains. Additionally, for these systems to work, they tend to require a specified target word.

For what concerns the pragmatic aspects of metaphor, extra-linguistic factors (roles, norms, power), many-to-many form–function mappings, and compositional intentions all complicate any deterministic labeling. Derived from this complexity is the fact that computational pragmatics still lacks robust models of *why* metaphors are chosen in particular contexts and with what effects, despite advances in contextual neural models (Hesse, 2023; Cappa, Lippolis, and Zoia, 2025).

In the context of visual metaphor understanding, while the work by He et al. (2024a) focuses on detection, it does not adequately address metaphor understanding according to existing theoretical frameworks, which claim that a comprehensive approach should at least involve identifying both source and target domains. A study by Mao et al. (2025) that uses MetaPro (Mao et al., 2023) on human and ChatGPT-produced metaphor shows that both humans and ChatGPT do not extensively rely on metaphors when providing responses to medical inquiries, however, that there is the most notable variation, thus not overlap, in frequent concept medical mapping with respect to other domains: where the differential usage of cognition and series of actions-related source concepts is the main dissimilarity between ChatGPT and humans. In this work, the term “metaphorical cognition” is employed to show cognition through the lens of metaphor, as suggested in CMT (§2.2.2.1). Another study on conceptual mappings by Mao et al. (2024) applied to various weather disasters reveals the ways to frame the confrontation with those natural phenomena in order to gain better policy recommendations. Despite applications across domains, MetaPro is trained on general, non-domain-specific resources and its current version cannot handle extended metaphors or document-level inference; consequently, robustness on domain-specific mappings and unseen metaphors remains unestablished in the published evaluations.

Computational metaphor entailment concerns the evaluation of logical inference relationships between figurative expressions and their literal counterparts through structured assessment tasks. This approach, pioneered by Agerri (2008), frames metaphor understanding as a recognizing textual entailment problem where systems must determine whether metaphorical statements logically entail, contradict, or are neutral with respect to literal interpretations. Performance evaluation on computational metaphor entailment tasks reveals significant challenges for current models, with results varying considerably across different evaluation paradigms and datasets.

On the other hand, it has to be noted that entailment tasks are inherently problematic when using AI, as highlighted by the three caveats identified by Leivada, Dentella, and Günther (2024) in their analysis of language model evaluation. The *problem of double standards* involves using evaluation methodologies that either cannot be applied to humans or are assessed differently across human and artificial systems (Leivada, Dentella, and Günther, 2024). Multiple choice entailment tasks create artificial evaluation scenarios that humans rarely encounter in metaphor processing. Additional problems include response instability, where models provide inconsistent predictions for identical inputs across multiple trials (Dentella, Günther, and Leivada, 2023), and systematic biases in evaluation frameworks that can misinterpret statistical regularities as genuine understanding (Dev et al., 2020). These limitations suggest that while computational metaphor entailment shows promising surface-level performance, current multiple-choice evaluation paradigms may overestimate genuine figurative language comprehension and mask reliance on spurious correlations rather than deep semantic understanding.

3.0.4.5 Computational approaches for scientific metaphors

A significant question is whether processing metaphors in scientific text requires special treatment compared to other genres. Many general-purpose metaphor systems have been evaluated on news, fiction, or conversational data (e.g. the VU Amsterdam Corpus, see Section 3.0.1). As seen in Section 2.3, however, scientific discourse has distinctive vocabulary and metaphor usage, but also different levels of conventionality.

A general metaphor detector might catch these if it has seen similar language, but certain terms are domain-specific, and their metaphorical usage might confuse a system not trained on biomedical text. A recent experiment on PubMed articles attempted to identify figurative usage of terms like “juvenile”, “baby”, “fish”, etc., by training models on biomedical corpora (Demner-Fushman, Rogers, and Mork, 2021). It found that no single model among those examined performed best for all terms, as some metaphors were easier to catch with simple models, others needed the contextual prowess of BERT. Thus metaphor detection can be term-sensitive and domain-sensitive.

Domain adaptation, however, is not a trivial task. If a metaphor model is trained on one type of text, its performance may drop on another. For example, models from the VUA shared task (news-heavy) might flag every instance of “X is a catalyst for Y” as metaphor because in general English “catalyst” is metaphorical (meaning “trigger”), but in a chemistry paper “catalyst” is literal. Conversely, a model might miss metaphors that rely on scientific knowledge (e.g. “the software virus spread”: one needs to know a virus is a biological concept being mapped to software). Some approaches to mitigate this include fine-tuning on target-domain data (if available) or using multi-genre training corpora. Adapting to scientific text may require handling varied metaphor registers: from formal research articles (where metaphors are often signaled by phrases like “as if” or are very conventional) to popular science articles (which may use more vivid, novel metaphors for pedagogy).

Some metaphors are essentially domain-bound. For example, the “lock-and-key” metaphor in biochemistry is so entrenched that it is a part of scientific jargon. A domain-general system might correctly label “enzyme” and “key” as different domains (biology vs. everyday objects) and flag it as metaphorical (Dodge, Hong, and Stickles, 2015). Domain-sensitive approaches might need to calibrate what counts as metaphor depending on user needs. Some recent research in science education examines how metaphors and analogies are used instructionally (Kampourakis, 2020), arguing that many core biology concepts are in fact metaphors, indispensable yet at times risky if taken literally. Used reflexively, metaphors help students think; used uncritically, they can reify, anthropomorphize, or smuggle teleology and thus mislead learners. These risks are associated with the inherently performative aspect of metaphor (§2.3.3) and discussed in more detail in §3.1.2.

Very recently, large transformer models have been used for metaphor generation to aid science communication. Kim et al. (2023) introduced Metaphorian, an interactive system that through GPT-3 and other LLMs, helps science writers create extended metaphors for concepts in astronomy, biology, physics and other domains. In the work, the authors constructed a dataset of 600 extended metaphors, that is, cohesive mutual analogies, spanning six scientific domains and three reading levels. Using prompting and few-shot examples, the system could generate candidate metaphors which were then refined with the users. The metaphors were evaluated by science writers on criteria like originality, metaphoricality, and scientific precision in both the source and the mapping explanation for quality. This evaluation shows

that in scientific metaphor generation, it is not enough for a metaphor to be novel or grammatically fluent (to be *plausible*: see §3.0.2). The use of LLMs here show promise as metaphors scored well on fluency and creativity, but also reveals limitations, as the model has sometimes produced metaphors that were scientifically imprecise, or mixed incompatible source domains.

The results from the reviewed works show that there is not much investigation on scientific metaphors, and also that while LLMs have potential in generating ones for different goals or levels of expertise, the lack of a symbolic structure that can help directing these systems towards less hallucinations, being more precise and adaptable to different contexts without causing harm, is fundamental to avoid found pitfalls.

Furthermore, evaluating scientific metaphor processing systems poses particular challenges since metaphor conventionality ranges widely in scientific discourse, sometimes blending into literal jargon. This continuum complicates establishing clear annotation standards and gold datasets, a topic addressed by Wan et al. (2024), who emphasize the need for domain-aware annotation protocols. Explaining metaphorical meaning in scientific contexts also demands interpretable models that align with scientists' epistemic practices.

3.0.5 Visual metaphor processing

Building on the Pitt Ads images dataset by Hussain et al. (2017), Ye and Kovashka (2018) propose ADVISE, focusing on symbolism and external knowledge in ads to recover the implied persuasive message of an ad. They used the Pitt Ads symbolic annotations and external knowledge bases to train models that link detected objects to abstract concepts. Their system learned a joint image-text embedding to match ads with their implicit message statements, though it does not work on metaphor structure per se thus only indirectly touches metaphor when symbol mappings resemble metaphorical transfer, and no dataset was released.

To foster richer research in visual metaphors, the MetaCLUE task set was introduced (Akula et al., 2023), though its code and dataset are not publicly available.

Yosef, Bitton, and Shahaf (2023), after releasing IRFL dataset, propose two benchmark tasks: a figurative understanding task, where a model must choose the image that best visualizes a given idiom/metaphor from a set of candidates, and a figurative preference task, which tests if models rank figurative interpretations above literal ones for a phrase. Results show that the language model struggled on the image selection task far below human performance. This performance could also come from the fact that today's LLMs struggle with explicit multiple choice reasoning (Leivada, Dentella, and Günther, 2024).

The Video Metaphor Corpus was released by Alnajjar, Härmäläinen, and Zhang (2022) and enabled experiments on video metaphor detection. The creators found that using only the subtitle text outperformed multimodal models, suggesting that many metaphors in their videos were primarily linguistic. More recently, Kalarani, Bhattacharyya, and Shekhar (2024) introduced a Video Metaphor Captioning dataset, comprising advertisement videos with human-written metaphorical captions. They propose the task of generating a one-line caption that describes the metaphor in a video, requiring the model to identify the primary concept, secondary concept, and shared property from the video. This dataset is yet to be released. However, the evaluation metrics proposed (Average Concept Similarity and Average Concept Distance) have limitations, as they rely on exact word matching and may not capture

the full range of valid human interpretations that could use synonyms, paraphrases, or culturally-specific expressions of the same metaphorical concepts.

A recent line of work is visual metaphor generation. Chakrabarty et al. (2023) and Su et al. (2024) formulate the task as, given a linguistic metaphor, generate a figurative image for it. Chakrabarty et al. (2023) tackled it by orchestrating a LLM with a diffusion-image based model. In this way, they generated a dataset of metaphorical images for more than 1000 metaphoric phrases. Professional illustrators judged many of these generated images to successfully convey the metaphors. In a related approach, Su et al. (2024) proposed a more structured pipeline to first parse the input metaphorical sentence to identify the target source and intended metaphorical meaning to retrieve or generate visual elements for the source domain to guide the image synthesis. Both works note the difficulty of standard diffusion models in handling metaphors: if prompted naively, the model might either produce literal elements or no elements at all. The LLM-guided approach helps by rephrasing the prompt to include the shared property. These early results are promising, but they also highlight challenges like compositionality and evaluation, which for this task is mostly manual.

Compositionality in particular is a known problems for LLMs. For metaphor understanding and generation, it takes a role-binding form. Systems must align a source domain to a target domain, bind properties to the right arguments, and preserve constraints across clauses or modalities. Recent LLM evaluations targeted at metaphor show that models can do reasonably well in-distribution but not so much when mappings or required inferences shift, and that pragmatic context often decides the intended reading. This fragility is consistent with broader pragmatic limits documented for LMs on implicature-like phenomena (Tong et al., 2024). In visual metaphor generation, diffusion pipelines inherit compositional weaknesses from text-to-image models: object-attribute binding errors, missing or duplicated entities, and incorrect spatial relations, all of which are fatal to figurative scenes that depend on precise multi-object composition. Chakrabarty et al. (2023) explicitly note under-specification and composition issues, and follow-up work in compositional text-to-image generation quantifies these failures and proposes fixes, but the problem is not fully solved.

Integrating heuristics based on existing ontologies on metaphor and knowledge graphs into the metaphor detection and understanding pipeline remains largely unexplored, despite its potential to render theories standardized and machine-readable across multimodal contexts, aid with compositionality problems, reduce hallucinations, adapt them for other automated reasoning tasks, and enhance system explainability.

3.0.6 Hybrid approaches for computational metaphor processing

In recent years, researchers have grafted structured semantic knowledge onto Transformer backbones to improve metaphor detection and interpretation. FrameBERT (Li et al., 2023) fine-tunes RoBERTa while jointly learning embeddings for FrameNet lexical units, yielding interpretable token-frame alignments. Lieto, Pozzato, and Zoia (2025) introduce MET^{CL}, a T^{CL}-based system that builds a single, emergent prototype for each metaphor by composing HEAD and MODIFIER concepts. The approach is anchored in the FrameNet-style frames embedded in MetaNet and therefore inherits the blind spots of that inventory: if no source/target frame exists, MET^{CL} must first extend MetaNet with heuristic candidates before it can even run its logic. MET^{CL}'s prototypes are assembled from ConceptNet triples, and risk importing

properties that are irrelevant to the discourse at hand; when concept information is missing, LLMs are proposed as substitutes.. Carenini et al. (2024), who adopts a very similar approach, finds the same issues: Bayesian typicality machinery works well for some types of metaphor, e.g. X is Y, but it remains to be enhanced for metaphor processes at large and creative ones.

A related neurosymbolic line brings symbolic conceptual structure to bear on visual persuasion, which is methodologically relevant to figurative understanding even when not explicitly modeling metaphor. The ADVISE model (Ye and Kovashka, 2018) links detected objects in ads to curated abstract symbol categories and learns a joint image–text embedding so that ads and their implied message statements co-locate. This is symbolism-driven pragmatic inference rather than explicit metaphor modeling, since it does not represent source–target mappings or cross-domain entailments. Still, by reasoning over high-level concepts (for example danger, freedom, love), ADVISE improved multiple-choice message retrieval over image-only baselines. In contrast, recent meme models introduce metaphor-aware objectives that directly supervise figurative cues; for instance, Wang (2022) adds an auxiliary metaphor classification task to strengthen the text encoder’s sensitivity to figurative language. Together these works illustrate a broader neurosymbolic trend: structured conceptual knowledge guiding neural representations toward more interpretable figurative and persuasive understanding.

In conclusion, computational work on metaphor has a long history, but only recently, and largely thanks to transformer-based architectures, have both the breadth of problems and plausible solutions come into sharper focus. Computational approaches grounded in CBT remain in their infancy. Purely symbolic systems are constrained by the need for continually curated knowledge bases, while purely neural/generative methods have their own limitations, especially concerning the lack of dedicated datasets and consensus regarding evaluation methods. A further source of friction, related to inconsistent evaluation, is inconsistent task nomenclature. For instance, the task of identifying a metaphor’s source and target domains is variously labeled “analysis”, “processing”, “extraction”, and it is sometimes folded into what papers call metaphor “understanding” or “interpretation”. Although these two terms are often treated as synonyms, and “understanding” is sometimes cast as broader because it may include paraphrase generation or affect detection, “interpretation” is frequently paraphrase-based as well. Notably, the source/target distinction is specific to CMT, where properties are not the primary focus; along this line, new labels appear, such as *cognitive metaphor*, to contrast with the linguistic metaphor commonly studied. The tasks range become more complex when, beyond source and target, a connecting property or a reason for the connection of source and target is also extracted (often under the broad heading of “processing”).

Metaphor-related computational tasks have thus shown a terminological variability. To make some order, Table 3.1 reviews currently reviewed approaches in the literature across four layers of increasing complexity and cognitive depth. Detection (Layer 1) represents the most basic level, testing whether a system can perform binary classification to identify metaphorical versus literal language⁶, a task that has dominated computational metaphor research for decades. Analysis (Layer 2) requires deeper semantic understanding, as systems must not only detect metaphoricity but also identify the specific source and target domains involved in the metaphorical mapping, demonstrating comprehension of the conceptual structure underlying the figurative expression. This layer also involves the generation of paraphrases apt

⁶In this sense, theory-based one, as we show in the related section.

to explain the metaphor and related tasks. In particular, since it conflates distinct subtasks, we differentiate between domain identification (finding source/target, 2a) and interpretation/paraphrase (explaining meaning, 2b). Orthogonal to this layer is also entailment, which can be considered as another type of interpretation. Reasoning (Layer 3) demands a higher level of cognitive sophistication, requiring systems to recover and articulate the blending property or shared, overarching characteristic structure that justifies why the source and target domains can be meaningfully connected, essentially explaining the analogical basis of the metaphor. Finally, Generalization (Layer 4) evaluates the robustness and consistency of approaches by testing how well the same architectural framework performs when these tasks are posed across different linguistic domains (e.g., scientific vs. everyday language) and across modalities (e.g., textual versus visual metaphors). The robustness axis here therefore concerns cross-domain robustness and cross-modal robustness. As evident from the table, the vast majority of existing work concentrates on Layer 1, with progressively fewer approaches addressing the higher-order reasoning and generalization capabilities that are crucial for genuine metaphor understanding.

Despite these advances, the proliferation of approaches and persistent challenges reveal a field in transition. The following section examines how these technical limitations, combined with theoretical fragmentation and harmful metaphor persistence, constitute what we term “the metaphor crisis”.

3.1 The metaphor crisis

The proliferation of metaphor studies, both theoretical and empirical, and their consequent fragmentation has inevitably made different views emerge. While some conflicts are undergoing resolution, some still remain open. In this chapter, we define what today is a phase of transition for metaphor studies, sparked by what we can call *the metaphor crisis*. We analyze the factors that have brought over this phase with respect to aspects in theoretical studies, in neuroscience and computer science, along with the solutions that have been proposed.

3.1.0.1 What is a crisis?

The concept of theoretical ‘crisis’ as driven by fragmentation and lack of consensus is well-established in academic literature, for instance, in the fields of strategy and organizational research (Volberda, 2004; Bundy et al., 2017). Following Y. Dafermos (2024), who advocates that crisis is a concept part of broader social theory, we intend with the term a historically immanent transitional phase.

Crisis as systemic rupture has been discussed in the field of philosophy of science. T. Kuhn identified crisis as a period characterized by “the proliferation of competing articulations, the willingness to try anything, the expression of explicit discontent, [and] the recourse to philosophy and to debate over fundamentals” (Kuhn, 1962, p.61). Kuhn argued that crisis emerges when accumulated anomalies undermine a field’s dominant paradigm, forcing practitioners into ‘extraordinary research’ without established frameworks to guide them. At the same time, the philosopher noted that this fragmented state, marked by “claims, counterclaims, and debates over fundamentals” (*ibid.*, p.61), characterizes philosophy and much of social science as their normal condition.

Beyond the evident similarities with Kuhnian’s definition of a crisis (from which descends the concept of the “scientific revolution”), to define the metaphor crisis,

we also resort to critical accounts that assess fields by the social organization of evidential standards (Longino, 1990). In this view, the present situation in metaphor studies is a crisis in the sense of a coordination failure: multiple traditions (cognitive, pragmatic, corpus/critical, experimental, neuroscientific, computational) solve different classes of problems with limited cross-talk. At the same time, a fragmentation perspective recognizes that contradictions are inherently indispensable parts of regular coordination practices, and as such, when we have different actors engaged in coordination processes, a multiplicity of interpretations arises and multiply, forming coalitions around different interpretations of the situation (Wolbers, Boersma, and Groenewegen, 2018). In interdisciplinary domains, durable progress typically proceeds via integrative pluralism (Mitchell, 2002) and collaboration in trading zones sustained by boundary objects: shared datasets, protocols, and task definitions that permit work even if without full theoretical consensus (Galison, 1997; Star and Griesemer, 1989).

We have previously mentioned the two manifestations of the metaphor crisis, methodological fragmentation and harmful metaphor persistence. They are not separate problems but mutually reinforcing aspects of a single phenomenon. Metaphor-unaware research methods make it difficult to identify which metaphors cause harm, while lack of consensus on how to study metaphors means no shared framework exists for systematic metaphor critique. Fragmented theoretical approaches cannot provide coherent alternatives to problematic metaphors, leaving harmful framings to persist by default. Meanwhile, entrenched metaphors resist investigation precisely because they appear commonsensical. This creates a vicious cycle: harmful metaphors persist because we lack adequate tools to study them systematically, while our capacity to develop better analytical tools remains constrained by the very metaphors that shape how we conceptualize metaphor itself. The crisis thus operates at both methodological and substantive levels simultaneously, with each dimension reinforcing the other's persistence.

As a first step, understanding this cycle requires examining how methodological challenges manifest across different domains of metaphor research.

3.1.1 Methodological issues

The metaphor crisis firstly stems from issues of method, whether it is applied on theory, on neuroscience or computation. Below, we address these three sides, each contributing equally.

3.1.1.1 Theoretical metaphor studies

Over the years, as many metaphor theories have been developed, equally as many theoretical accounts that advocate for the integration of such theories have been developed. In fact, contemporary metaphor studies are marked by their theoretical debates, methodological disagreements, and calls for greater integration.

The theoretical diversity in metaphor studies is particularly evident by standing methodological disagreements and terminological inconsistencies: still now, scholars lament a lack of standardization in how metaphors are identified, and analyzed across studies (Nacey et al., 2019). For example, different researchers can use different and often not fully explained criteria to decide what counts as a metaphorical expression. In response, while some groups developed explicit metaphor identification procedures like the MIPVU (Steen et al., 2019), which improved transparency and replicability, not all researchers have adopted it, and variability persists in metaphor

identification practices and how strict or inclusive the definition of metaphor should be. In this way, one study's metaphor might be another's dead figure of speech.

At the heart of the progressive fragmentation of metaphor-related theories and studies is the distancing from CMT, which, while still valid for some aspects of metaphorical treatment, requires more clarity in others. In particular, one point of contention concerns the empirical validity of CMT: critics argue CMT originally relied heavily on analysis' intuition and examples, lacking empirical proof that conventional metaphoric mappings truly shape how people think, thus feeling CMT's cognitive claims needed better experimental evidence (Gibbs Jr, 2017). As scholars scrutinized its claims, Gibbs Jr (2017) even described these disagreements as ongoing "Metaphor wars" (see Section 2.4).

Furthermore, the experiential (in the broader sense of the term) dimension of metaphor has traditionally been downplayed, with CMT-based research focusing primarily on metaphors as a mental and individual achievement; bodily, if anything. Studies have so far paid little attention to context and the collaborative production of metaphoric language in discourse, which is now arising as a fundamental aspect of metaphor-based work (Jensen, 2018).

This important aspect, relating to the communicative or situated aspect of metaphor, has emerged in recent years and highlighted a gap even in current theoretical accounts of metaphor, since it prevents from viewing metaphor as a multidimensional phenomenon (Semino, 2008). As we saw in previous chapters, metaphor systems, especially those appearing in science, are not neutral but reflect underlying belief systems that justify social actions and representations. In this view, language, and metaphor in particular, plays a key role in realizing these social and political values: texts are always "oriented social action" (Titchkosky, 2007, p. 21).

Linell's notion of an "interworld" provides a valuable theoretical framework for understanding these social dimensions of metaphor (Linell, 2009). Unlike traditional cognitive approaches that locate metaphor primarily in individual minds, the concept of *interworld* emphasizes how metaphorical meaning emerges (*interanimates*) through interaction in a shared communicative space. However, early critics like Wierzbicka (1986) already noted CMT's emphasis on bodily experiential *gestalts* risked blurring important cultural differences. Indeed, Gibbs Jr (2017), contrary to standard assumption within CMT which claim source domains of conceptual metaphors are primarily based on direct sensorimotor experiences, argues that metaphorical meanings do not necessarily arise from the mappings of purely embodied knowledge onto abstract concepts. Gibbs argues that the experiential basis of metaphor is not exhausted by simple sensorimotor mappings; ecological and interactional work further suggests that so-called source domains come already shot through with social, cultural, and affective meaning (*ibid.*).

This would also explain why sensorimotorily impaired people can still think of embodied metaphors (Phillips, Grenoble, and Mason, 2023). For example, in the conceptual metaphor LIFE IS A JOURNEY, the experience of physical journeys cannot be reduced to a mere physical experience but will, most of the time, also entail "existential, social, and cultural meanings" (Jensen, 2022). The take still reconciles with CMT, since one could argue that this is because linguistic understanding recruits simulation mechanisms that run embodied action schemas, binding words to structured motor-goal representations that already encode goals, causal structure, affect, and social roles. Metaphor maps these rich simulations from concrete domains to abstract ones, so the relevant inferences can be drawn even without concurrent or even prior somatosensory experience (Narayanan, 1997; Narayanan, 1999).

While theories like CMT advocate for widely shared associations, another contention point concerns specific cultural diversity examples⁷. For example, in the Mlabri language of Thailand and Laos, spatial-emotional metaphors directly reverse western patterns: Wnuk and Ito (2021) documented that the concept of “heart going down” represents happiness, contentment, and positive emotions, while “heart going up” signifies distress and negative emotions, contradicting the widely shared “happy is up, sad is down” metaphor. As another example, cross-linguistic research by Koptjevskaja-Tamm and Nikolaev (2021) examined the AFFECTION IS WARMTH metaphor across 94 languages worldwide and found it has remarkably skewed distribution across language families. The metaphor appears to be absent or extremely marginal in the majority of language families and linguistic macro-areas, challenging assumptions about widely shared embodied metaphors. Furthermore, in the same study, researchers observed that in a number of African languages it is the concept of “cold” that has pleasant associations, such as peace, serenity, comfort, contentedness, absence of calamities, while in many other cultures is the opposite. These pieces of research thus show that in spite of evidence supporting some common tendencies especially with primary metaphors (Grady, 1997; Kövecses, 2005; Kövecses, 2015), our experiences as described by metaphors are filtered through cultural values and norms along with social practices, and that for this reason they cannot be in any way considered universal.

Another example can be drawn from disability studies (Schalk, 2013). Scholars have pointed out that many common metaphors are ableist, equating disability with negative qualities (e.g. phrases like “blind to the truth”, “crippled by fear”, “lame”) such as using disability as a metaphor for ignorance, obstinance, helplessness or weakness. In this view, despite authors being well-meaning, they unwittingly reinforce an ideology of disability-as-deficit and in particular resume CMT’s principles even if it lacks the full set of tools to account for the sociocultural baggage of such metaphors. This represents a limitation of scope rather than empirical falsification: CMT addresses the reasons and mechanisms of conceptual metaphor but not so much their consequences that these critical approaches raise. For this reason, recent studies claim for a view on metaphor that is not just embodied, but inter-bodily, and to incorporate a broader sociocultural and ethical lens in metaphor studies. In such a context, constitutive metaphors (see §2.3), which have been widely studied, are only now seeming to be disputed as feminist practice, teamwork, and patient-centredness push towards more ecological or collaborative framings (Bleakley, 2017).

Connected to inter-bodily multidimensional accounts of metaphor and the critiques thereof is the metaphor resistance phenomenon, only recently studied, and the various reasons why it happens. For instance, people resist metaphors if they lack explanatory power or for a preference for alternative metaphorical concepts with respect to normative ones. In the cases where guidelines expose the reason why people deliberately reject metaphor, such as not using terms like “blind review” in academia as it casts blindness as a lack of insight and use “anonymous review” instead, it is easier to notice such shift. However, without a comprehensive metaphor study, it is not possible to know why some metaphors are not picked up (Gibbs Jr and Siman, 2021). Indeed, a research by Hart (2021) on extreme metaphors in UK

⁷It is important to note is that some scholars such as Gibbs and Kövecses argue that many critiques of CMT on the fact that CMT ignores culture or language entirely is false, however, the dialogue has pushed metaphor theory to broaden its scope and methods (Kövecses, 2008). Indeed, later work in the CMT tradition explicitly foregrounds cross-cultural, subcultural, and individual variation (Kövecses, 2005).

anti-immigration discourse found an interesting backlash effect, in that when readers see very extreme, overt metaphors, they can trigger skepticism and rejection of the message and in that case actually decreased public support for anti-immigrant policies.

The communicative aspect of metaphor is closely connected to context. If a metaphor can be extended, resisted, or interpreted differently across situations, then it is clear why a “pragmatic turn” in cognitive linguistics happened (Gibbs Jr, 2017).

Beyond CMT and the many subsequent, competing, and overlapping theories that have been developed over the years, many contemporary scholars argue that no single process can account for all metaphor usage, thus metaphor comprehension likely depends on multiple factors (Steen, 2023). People do not always use context in the same way for every metaphor, and that interpretation process vary with the situation and goals (Gibbs Jr and Colston, 2012), thus different theories might each capture part of the truth and operating in different conditions or cognitive stages. In this context, despite existing integration efforts that signal paradigm transition, insights from semiotic theory and pragmatics still result underplayed accounts of metaphor interpretation, with semiotic theory especially being not integrated in any existing multilayer theoretical account such as ECMT (see Section 2.2). The main issue in theoretical metaphor studies is thus that studies still do not follow homogeneous practices and multiple simultaneous theoretical underpinnings that can together explain different aspects of the metaphorical practice, due to the richness of metaphor as a phenomenon.

This fragmentation is also shown in linguistics for what concerns the multitude of corpus-based methods with respect to data sources and what counts as evidence. For instance, while corpus-based analyses add information beyond earlier intuition-only research, research also cautions that corpus findings still require interpretation and should be complemented by other evidence (Gibbs Jr, 2017). As a result, the field is gradually moving towards mixed-method approaches, although individual researchers often still prioritize one type of data.

This liminal phase between long established metaphorical frameworks and an emerging, competing (but also integration-oriented) paradigm represents a true crisis of metaphor. However, a complete analysis of the critical landscape requires examining recent neuroscientific studies that have yielded significant findings. In the following section, we will analyze the crisis markers arisen from these studies.

3.1.1.2 Neuroscientific metaphor studies

The work by Koller, Müller, and Kauschke (2022) is a comprehensive review of 116 neuroscience studies on metaphor and idiom processing. Rather than summarizing results of those studies, the authors critically examine how the studies were designed, focusing on definitions, theoretical assumptions, and stimulus controls. They identify several “elephants in the room”, i.e. often-neglected issues that could explain why results across studies are sometimes inconsistent or contradictory.

What they found is, first of all, inconsistent terminology or definition for figurative language and its subtypes across the analyzed studies. Different researchers operationalize metaphor in varying ways: some include only poetic metaphors, some include idioms, some treat similes as metaphors or exclude them. Similarly, the boundary between metaphor and other tropes (like metonymy or irony) is not consistently respected. The concept of figurative language itself is defined broadly by some and narrowly by others. This terminological inconsistency makes it hard to compare studies: one paper’s metaphor stimuli might overlap with what another

paper calls idioms or analogies. For example, some studies count any non-literal expression as a metaphor, while others only consider nominal metaphors and exclude verb metaphors or idioms. Koller, Müller, and Kauschke (2022) urge that researchers clearly define what subtype of figurative language they are examining, and they call for consensus on core definitions.

The same conclusion was drawn by Lakoff and Narayanan (2025), who explicitly caution that many studies on the neuroscience of metaphor have treated idioms as if they were transparent metaphors and then averaged across participants who may not even be construing the same metaphor. They show why this produces noisy or contradictory results: idioms vary in arbitrariness, mapping may apply to the construed image rather than the surface words, and body-part idioms can activate motor maps for reasons other than metaphor. They urge researchers to verify that participants share the same underlying metaphor and to avoid conflating idiomatic meaning with metaphorical structure.

Another point raised concerns by Koller, Müller, and Kauschke (2022) was the status of literal and figurative meaning, and in particular whether it is a binary distinction or a continuum. Some studies implicitly treat the two as distinct categories, but others (inspired by cognitive linguistics) see figurativeness as a matter of degree. This philosophical difference influences stimulus selection, thus lack of clarity in this field can affect results.

The analysis also concerns the influencing factors in metaphor processing, or better, how the stimulus is controlled to report the various factors known to affect metaphor processing. Koller, Müller, and Kauschke (2022) catalog a range of psycholinguistic factors (e.g. familiarity, conventionality, frequency, aptness, imageability, length) and structural factors (e.g. syntax, context provided) that can influence results. They found that many studies did not adequately control for these, or at least did not report doing so. For example, as mentioned, familiarity was controlled in roughly 60% of studies, meaning 40% of studies potentially mixed familiar and unfamiliar metaphors without accounting for it. Aptness was rarely explicitly controlled in the set of studies, even though it can modulate difficulty. Other factors such as imageability and semantic domains were also inconsistently handled.

Among these factors, the confusion of familiarity versus conventionality is more problematic. 18 out of 33 papers that said they examined familiar as opposed to conventional metaphors actually treated the terms as equivalent, usually defaulting to a definition akin to familiarity. Consequently, some conventional metaphors in experiments were merely those rated as highly familiar by individuals, rather than an objective measure of conventional usage. The authors argue it is crucial to clearly define these: one is subjective frequency, the other concerns collective knowledge.

Finally, some of these factors, such as familiarity, were determined as continuums also in the literature, but then are operationalized binarily.

The work serves as a caution that in order to solidify consensus on these theoretical issues, researchers must also address methodological consistency. On this topic, Lakoff and Narayanan (2025) argue that studies should be designed from an explicit account of the conceptual units involved, whether they are schemas, frames, or source-to-target mappings, so stimuli and contrasts correspond to real theoretical distinctions, and so metaphor is not operationalized as any nonliteral idiom.

Secondly, they propose using simulation semantics as the linking hypothesis, where comprehension recruits embodied simulation cascades, so that experiments should target the circuitry implicated by a given simulation. To generate testable predictions that connect cognitive-linguistic structure to neurally plausible circuits, they introduce a structured neural computation model to design tighter experiments.

In connection to this claim, we share the philosophical target: to avoid operationalizing metaphor as any nonliteral idiom, and to design studies from explicit conceptual units such as frames and patterns. Where Lakoff and Narayanan propose structured neural computation and simulation semantics as the linking hypothesis between conceptual units and brain circuits, our method operationalizes those same conceptual units as explicit knowledge patterns and ontologies that are explicitly defined and guide learning and inference at the algorithmic level, without committing to a particular neural mechanism.

The problems so far emerged in theoretical and neuroscientific accounts of metaphors reflect most of all in computational accounts of metaphor.

3.1.1.3 Computational metaphor studies

As partly glanced at in Section 3.0.4, computational metaphor studies exhibit three hallmarks of the metaphor crisis: fragmentation, stagnation, and technical limitations.

Fragmentation. A first problem that directly mirrors the fragmented landscape of metaphor theories is the lack of evaluation and standardization: computational tasks in metaphor studies range from detection to interpretation, yet, as we have seen in Section 3, tasks, definitions, metrics, and supported linguistic forms vary wildly across studies (see also Table 3.1).

If we look more into metaphor interpretation and understanding, especially conceptual ones, this usually requires either generating a paraphrase or identifying at least one of the two domains from a given piece of data (usually a sentence).

For the latter, while systems are often developed to identify that two domains are involved, most systems rely only on at least one specified target word, thus they do not study how to detect source and target at once. Furthermore, pinpointing the precise aspect(s) of the source that maps to the target is an unsolved challenge. For instance, high-level mapping in a metaphor can be clear, but the precise implication is subtle. Human metaphor interpretation operates on multiple levels, and current algorithms typically stop at identifying source and target domains and, in rare cases, a general shared property. Developing methods to extract or generate the specific analogical correspondences between the two domains would greatly enhance interpretability, though that would require the presence of gold-standard datasets that (i) are currently missing and (ii) would still hinder the possibility to evaluate metaphors in discourse, which is when they tend to happen in everyday life. Over the four layers we saw in Table 3.1, the vast majority of computational approaches remain concentrated on Layer 1 (detection), with progressively fewer systems attempting the other tasks.

Stagnation. If we delve deeper into the metaphor mapping system, computational studies, especially on metaphor detection, mainly leverage the *incongruity* of source and target domains, neglecting similar attributes shared between source and target concepts in metaphorical expressions. Here, we are talking about systems that use selectional preference violation as a baseline and feature, along with metaphor identification by comparing concreteness/abstractness or implicitly capturing incongruity through lexical semantic features of a metaphorical construction (Shutova, 2015; Tsvetkov, Boytsov, and Gershman, 2014; Turney et al., 2011). However, basing on theoretical accounts, it is the *similarity* between domains, explained through a

connecting property, and alignment of relational structure that drives the metaphorical composition through mapped features (Gentner, 1983; Bowdle and Gentner, 2005). Still, even in theory it is not clear how to establish this ground, given the unpredictability of metaphor. Only a few computational works have researched on this topic, and they mostly concern interpretation/explanation of metaphorical mappings. Here fit computational accounts of similarity-driven composition, which treat metaphor as feature attribution via shared properties or relational similarity between domains (Kintsch, 2001; Veale and Li, 2012; Shutova, 2010a; Bollegala et al., 2013; Tian et al., 2023), though no gold standard dataset continues to exist beyond a handful of examples and remains difficult to evaluate, adding up to a stagnant situation. A study additionally claims that there is also not much research on the cognitive patterns of AI through the lens of CMT despite them being crucial for human cognition and psychology analysis (Mao et al., 2025). Indeed, prompts crafted to test LLMs might not reflect how these models are actually used in everyday situations.

Furthermore, there is no mention on how to establish shared cultural and social context for effective communication to happen (as advised, for instance, by pragmatic theory: §2.2.2.5). Concerning cultural specificity, metaphors are inevitably culturally specific, and cross-lingual metaphor processing is in its early stages due to the complexity of creating such resources. Beyond commonsense or prototypical knowledge, recent theories of metaphors have commented on the lack of inclusion of personal and sometimes contextual aspects that influence knowledge acquisition and interchange, which reflects in computational studies. Concurrently, computational studies miss contextual cues, as datasets often focus on isolated sentences or images at most: what is the context of this data, what is the tone of voice of this data, what are the social consequences and the cultural background from which data originates, are all aspects still not accounted for⁸.

Indeed, as semiotic and pragmatic theories remain largely neglected in theories of metaphor (see Section 2.2), this reflects in computational studies: while CMT reigns supreme and CBT holds secondary prominence, computational studies fail to engage with the (generally fragmented) theoretical proposals that have emerged since these foundational frameworks. We have already explained the reasons for this lack of accountability. It is still worth to notice, however, that the plastic categories of semiotics and the insights of pragmatics could for instance facilitate better integration of multimodal approaches, yet they are conspicuously absent from current computational metaphor research. To the best of our knowledge, only one study on the annotation of pragmatic intentions has been made, with only one annotator and another for cross-checking, and showed LLMs are bad for intention mining (Michelli, Tong, and Shutova, 2024). Another study on ChatGPT shows pragmatic inferencing is not fully accounted for in LLMs (San Pietro et al., 2023).

Although these studies highlight pragmatic limitations, they also risk over-interpreting results, since the very testing paradigms may not transfer cleanly to LLM evaluation (see Leivada, Dentella, and Günther, 2024). Furthermore, the point in using computational methods for theoretical problems is the possibility to integrate and operationalize them in pattern-based ways that cannot neatly identically transfer to the

⁸Already in 2008, Barnden observed that although AI metaphor processing was yet to fully incorporate context. Contextual reasoning goals could serve as powerful drivers of metaphorical interpretation, helping to resolve meaning indeterminacy while filtering out irrelevant aspects of the source domain (Barnden, 2008).

real world, yet somehow still work even as predictive engines. Despite the possibility to explore this option, computational methods stagnate by repeating old errors, thus still failing to keep up with the theoretical landscape of metaphor.

Apart from theoretical metaphor studies, insights from cognitive science at large have also been largely downplayed. Among these, the pervasiveness and importance of noise and its role in shaping efficiency in communication systems, such as what happens in metaphor in communication, largely acknowledged by information theory (Dingemanse and Enfield, 2024), thus lacking an important foundation of human language.

Technical limitations. This theoretical myopia is not justified by high capabilities in the tasks computational systems aim to carry out. In fact, even computational tasks that, by now, should now be trivial, are still challenging for machines. For instance, detecting novel metaphors remains harder than catching common ones. Many current models, especially ones based on neural networks, tend to memorize metaphors that appeared in training data. A novel metaphor might only be caught if the model picks up on semantic conflict signals. Some knowledge-based methods can flag novelty by comparing to known metaphors (Dodge, Hong, and Stickles, 2015). However, grading metaphor novelty is outside the scope of most systems.

More fundamentally, in metaphor processing, even LLMs, the state of the art in metaphor processing, struggle to distinguish deep relational mappings from mere associations, particularly in complex or multimodal metaphors (Wijesiriwardene et al., 2023; Nezhurina et al., 2024). Recent empirical evidence confirms these limitations: the study by Sanchez-Bayona and Agerri (2025b) on metaphor interpretation shows that when investigating the performance of LLMs on four metaphor interpretation tasks, the output seems more influenced by features like lexical overlap and sentence length rather than by metaphorical content, concluding that any alleged emergent abilities of LLMs to understand metaphorical language are the result of a combination of surface-level features, in-context learning, and linguistic knowledge (basically, these systems tend towards shortcuts).

Another difficulty in detecting metaphors is shown by domain specificity: some metaphors are domain specific and domain transfer cannot always be used. In the case of scientific metaphors, which are still underexplored, novelty can be double-edged: a highly novel metaphor might be illuminating but also risk misunderstanding, thus ethical reflections on this use should be concurrent with metaphor processing systems.

Part of this problem, as already suggested, stems from datasets and specifically the lack of heterogeneous, annotated datasets that could show that in the first place. Creating high-quality metaphor annotations is laborious, as it requires trained annotators, which can also be time consuming and subjective. Existing datasets provide a good foundation, but more specialized datasets that could really elicit analogical reasoning beyond surface annotations are needed. Comprehensive datasets for metaphoric usage remain scarce, and building new ones is hindered by frameworks that go no further than simple domain mappings.

Connected to the problem of datasets and lack of resources is the one of interpretability, or better, final usability⁹: there is a gap between high-performing models,

⁹Final usability emphasizes practical utility from the end-user perspective, rather than technical interpretability from the developer perspective, focusing on whether explanations actually help users trust, verify, or improve system outputs for their specific tasks.

which are the final goal of most computational systems, and human interpretability, which is desired by users to understand what happens in computational systems to give a specific output. While neural models increasingly augment performance, symbolic approaches offer explanation. A promising direction in this case is thus neuro-symbolic methods that can provide human-readable justifications. And the error analysis shown in Sanchez-Bayona and Agerri (2025b), and the claim that LLMs themselves rate LLM-generated metaphors high even if they are low-quality according to human experts (Heilig, 2024), provide evidence for arguments that it is still not possible to evaluate LLMs' behavior in metaphor processing without extensive human evaluation, thus evaluation by those that need to use these systems in the first place.

Improved metaphor processing could significantly enhance other NLP applications, especially in scientific contexts. For instance, better metaphor understanding might improve scientific text simplification or machine translation of scientific articles. Moreover, in scientific knowledge discovery, metaphors could facilitate analogical connections between disciplines and potentially advance interdisciplinary research. In the metaphor crisis, however, we are not only facing methodological issues in metaphor analysis, whether theoretical or computational. The studies reviewed here show it is also a crisis of metaphors that currently dominate fields of studies and our lives and that need to be conceptually re-engineered.

3.1.2 The metaphors of crisis

While theoretical and empirical studies struggle to provide a comprehensive account of metaphorical processes, recent academic scholarship reveals a fundamental shift in how scientists themselves view dominant metaphors in their fields. A surge of critical research across multiple disciplines shows that many traditional scientific metaphors are not just inadequate but, at times, actively harmful.

Recently, research has moved beyond simple metaphor replacement toward consequentialist evaluation frameworks. Rather than judging metaphors by their accuracy, scholars now assess them by their effects on understanding, research, directions, and social outcomes (Stuart and Wilkenfeld, 2022). This approach is in line with the field of conceptual engineering which focuses on questions of the normative kind with respect to the descriptive one: instead of asking what is something, it is asked what is that something *for*. A natural consequence of such process is reflecting on one's own linguistic agency, wherein language is a tool for agency and at the same time it is paramount to claim agency on language (Huetter-Almerigi et al., 2025).

3.1.2.1 Against computational metaphors

Metaphors derived from engineering and computer science are the ones that for the longest time have been put at the center of discussion.

Even before computers, one source of metaphorical thinking is that of machinery. For example, in evolutionary biology, the metaphor of the organism as a machine has been criticized for implying teleological design and supporting intelligent design arguments: the metaphor suggests discrete mechanical parts when biological systems are more fluid and context-dependent (Nicholson, 2019; Taylor and Dewsbury, 2018; Brito and Marques, 2014; Boldt, 2019). The adjacent factory metaphor for cells is also problematic, reinforcing mechanistic thinking that misses the self-organizing nature of cellular processes (Pearce and Miller, 2025).

Brette (2019)'s critique launched a reassessment that continues to exist through neuroscience. The author's analysis of the neural coding metaphor revealed fundamental problems: the alleged codes depend on experimental apparatus rather than neural properties, the causal structure mismatches brain dynamics, and the metaphor conflates information with meaning. The response has been a major interdisciplinary project of experts that reconsider the brain-computer analysis. Matassi, Mishra, and Martinez (2023)'s historical analysis revealed how brain metaphors evolved from hydraulic to electrical to computational, suggesting that like in other cases in history, current computational metaphors may be another phase.

Embodied cognition research provides further evidence against computational assumption: for instance, Finley (2025) shows that computational metaphors constrain theory evaluation, creating bias toward disembodied approaches. Furthermore, motor actions are incompatible with computational models (Bongard, 2021).

The alternative frameworks that are gaining traction emphasize dynamic, embodied and environmentally-coupled approaches. For instance, network and internet metaphors shift focus from computation to communication, opening new avenues of inquiry.

3.1.2.2 Against war metaphors

Metaphors of war are used everywhere, but perhaps one of the fields in which they are most used is medicine. Concerns about war metaphors in medicine have deep roots. S. Sontag was among the first to argue that metaphors surrounding illness were active agents that amplified suffering and created stigma. She claimed that war metaphors particularly burdened patients with responsibility for both causing and curing their diseases, while punishment metaphors transformed illness into moral failing (Sontag, 1990). Her prescient analysis of cancer and AIDS metaphors anticipated many problems that recent research has now empirically confirmed.

More specifically, one recently emerged case study confirms the long-standing assumptions that war metaphors are more harmful than beneficial.

The COVID-19 pandemic provided an unprecedented real-time experiment in metaphor effects. The extensive public use of war metaphors generated immediate scholarly response documenting harmful consequences across multiple domains (Taylor and Dewsbury, 2018; Olza et al., 2021). Political research revealed concerning authoritarian implications: war metaphors increased acceptance of civil liberty restrictions and promoted obedience over awareness (Panzeri, Di Paola, and Domaneschi, 2021a; Benzi and Novarese, 2022). The transformation of citizens into soldiers shifted focus from individual agency to hierarchical compliance. International research documented how war metaphors contributed to nationalism and xenophobia, particularly during the "China virus" framing that increased anti-Asian racism.

Furthermore, while healthcare worker heroization seemed positive, it created problematic expectations and shifted responsibility from institutional failures to individual sacrifices, and rendered patients as passive victims or casualties (Marron et al., 2020). Empirical studies found war metaphors ineffective at increasing compliance with health measures compared to non-militaristic language (Taylor and Dewsbury, 2018). E. Semino led the *#ReframeCovid* initiative, which collected over 550 alternative metaphors to the dominant ones of COVID-19 across 30+ languages, showing how academic criticism can turn into practical, community-driven action (Olza et al., 2021).

These findings demonstrate that metaphors generate emergent conceptual realities rather than simply describing existing ones: war metaphors did not just describe the pandemic, they created new social and political possibilities.

In general, as already shown in Section 2.3.3, research reveals systematic problems with traditional medical metaphors beyond pandemic contexts. War metaphors in cancer care lead to feelings of disempowerment, guilt and fatalism among patients (Taylor and Dewsbury, 2018). Treatment described with violent metaphors is perceived as more difficult and undermines passive prevention strategies such as smoking cessation (Marron et al., 2020). In musculoskeletal medicine, machine and “wear and tear” metaphor reinforce kinesiophobia and catastrophic thinking. Paradoxically, studies found that metaphors validating patient misconceptions received higher trust ratings, explaining their persistent use despite harmful effects (Fatehi et al., 2022; Chandler et al., 2024). This body of research reveals an ethical dilemma in clinical communication, as metaphors that reinforce misconceptions are associated with higher ratings of communication effectiveness and trust, but at the same time creates tension between immediate rapport-building and long-term patient understanding.

To be fair, there are contexts in which militaristic framing can be functional. Qualitative and mixed-methods studies report that some patients, especially in earlier or potentially curative phases, use battle imagery to sustain agency, organize action, and cope during highly demanding treatment windows; in acute institutional crises, war framing can also help leaders impose order and coordinate collective effort. Yet these effects are contingent on perceived controllability and stage, and they are not obtained by default across audiences (Semino et al., 2017; Liu, 2025; Bodd et al., 2023; Naamati-Schneider and Gabay, 2022; Panzeri, Di Paola, and Domaneschi, 2021b).

Alternative frameworks emphasize care over cure. For instance, hospitality metaphors position physicians as hosts and patients as guests (Cox, 2016). Quilting metaphors for medical research emphasize craft, collaboration and community rather than mechanistic problem-solving (Taylor and Dewsbury, 2018). In general, however, providing options alongside established ones helps in empowering choice rather than reinforcing thought schemas.

Outside medicine, which is our main case study, war metaphors have also been challenged in climate change studies, which confirm these promote adversarial thinking over collaborative solutions and can actually backfire by creating climate doom-ing. Alternative metaphors proposals concern focusing on those borrowing from medical language (Johnstone and Stickles, 2024; Forgács, 2024; Mangat and Dalby, 2017).

3.1.2.3 Evolutionary biology, climate science, chemistry, metaheuristics

The aforementioned problems can also be seen from examples that span the life sciences and fields that borrow heavily from them. Evolutionary biology and chemistry form the foundation of life sciences, climate science studies planetary systems that sustain life, while metaheuristics explicitly borrows biological metaphors to solve computational problems.

Contemporary researchers in evolutionary biology advocate moving beyond the survival of the fittest phrase, which misrepresents Darwin’s actual views and create false impressions of competition. The metaphor ignores ecological niches allowing multiple species to be fit for different reasons and has been used to justify capitalism and social darwinism (Abebe et al., 2017). Furthermore, the “selfish gene” metaphor coined by Dawkins (1981) faces sustained criticism for oversimplifying gene-gene

and gene-environment interactions while reinforcing individualistic, competitive worldviews. Research shows how this metaphor has been reified as a fact rather than remaining metaphorical (Lent, 2017).

Alternative framings emphasize cooperation and networking: Margulis and Sagan (1997) highlight that life “did not take over the world by combat but by networking” (p. 29).

Another problematic metaphor is the “genetic blueprint” metaphor, having been heavily criticized for being deterministic and oversimplifying gene-gene and gene-environment interactions, along with being incompatible with recent advances in evolutionary biology (Taylor and Dewsbury, 2018).

In climate science, metaphors receive mixed evaluation. The metaphor of the greenhouse effect faces criticism for mechanical inaccuracy and public confusion, while war on climate change metaphors promote adversarial thinking over collaborative solutions (Deignan and Semino, 2019). Alternative metaphors like “heat-trapping blanket” and “earth’s fever” show promise for more effective communication (Deignan, Semino, and Paul, 2019; Johnstone and Stickles, 2024).

In chemistry, molecular machinery metaphors are criticized for oversimplification and inappropriate scale assumptions. The metaphor suggests discrete, mechanical parts when molecular interactions are more fluid and context-dependent, while inadvertently supporting teleological reasoning (Borsley, Leigh, and Roberts, 2022).

Another problem concerns borrowing biological or life sciences terms to create metaphors in metaheuristics (Aranha et al., 2022). In this paper signed by more than 100 researchers, the authors argue against metaphor-based metaheuristics and commit to no longer publishing them. The researchers contend that metaphors in this context are often superficial, loosely mimicking real-world processes without genuine correspondence. Metaphors like *Grey Wolf Optimizer*, *Whale Optimization*, *Bat Algorithm*, all are shown to be reformulations of existing particle swarm optimization or evolutionary algorithms. When poorly implemented, these can obscure the actual mathematical operations being performed. Instead, they advocate for algorithms described through their actual computational mechanisms rather than biological or physical analogies. The core issue at hand is not metaphor use *per se*, but the selection of misleading metaphors to justify poorly implemented algorithms that are then marketed as novel contributions to the field.

3.1.2.4 Metaphors of AI

Metaphors of AI present problems that extend far beyond semantic precision to shape technological development, public understanding, and policy decisions. Anthropomorphic language for AI systems, especially LLMs, describing them as “thinking”, “learning” or “understanding”, systematically exaggerates capabilities by attributing human-like traits to statistical pattern-matching systems (Salles, Evers, and Farisco, 2020). This linguistic anthropomorphism has been noted and warned about from decades. In his Turing Award Lecture, ‘On the Cruelty of Really Teaching Computer Science’, E. Dijkstra criticized the field’s reliance on anthropomorphic metaphors, particularly in programming. He likened this habit of speech to the ‘intellectual stagnation’ of the Middle Ages (Dijkstra et al., 1989, p.1399)¹⁰.

The reason why this phenomenon happens is metaphors become jargon when used performatively and taken too literally. Specialists may treat their metaphors as if they were facts, letting the technical caveats drop until the figure is confused with

¹⁰Clearly, he was wrong about the intellectual stagnation of the Middle Ages. Among the many sources on the topic, a comprehensive one is presented in Hannam, 2011.

reality. Novices then mimic the vocabulary without grasping its substance (Johnson, 1994).

Baudrillard, too, had glimpsed the problems that could stem from metaphor surrounding AI, while for him, AI is clearly not artificial but rather “pseudo-artificial”: it lacks artifice, metaphor, the unspoken, and the secret, therefore it lacks intelligence. In other words, AI had lost the original purpose of a gateway to deeper meaning and transcendence (Baudrillard, 1990).

And as we have seen with metaphors, they have a performative effect: empirical studies demonstrate that users who anthropomorphize AI systems exhibit misplaced trust, overconfidence, and responsibility displacement (Bi and Huang, 2023; Waytz, Heafner, and Epley, 2014). The “mind in the machine” effect leads users to over-rely on systems they perceive as intelligent (Verberne, Ham, and Midden, 2015), while simultaneously obscuring the human agency behind AI development and deployment (Coeckelbergh, 2009). Watson (2019), in fact, argues that anthropomorphic rhetoric in AI is “at best misleading and at worst downright dangerous” (p. 417), creating obstacles to properly conceptualizing ethical challenges posed by emerging technologies.

These consequences are all the more compelling considering that any critical comparison with the computer presupposes that we first understand the machine, yet our grasp of the computer is itself built by analogy with the human. Over time, AI-related concepts originated as metaphors sediment into normalized literal meanings: “intelligence” is no longer reserved for organic or conscious beings. Thus, rather than only reading computers through human capacities, we increasingly come to read the human through computational ones (Boon, 2025)¹¹.

The “black box” metaphor, too, which describes AI systems as inherently mysterious and unknowable, exemplifies how metaphorical framing can actively harm understanding and governance as it discourages regulatory oversight and public engagement, even when many AI systems are interpretable. Current academic and industry guidelines have shifted toward describing AI systems as “opaque” rather than black boxes, recognizing the metaphor’s problematic implications for transparency and accountability¹².

The quick popularity and great capabilities of LLMs, visible to billions of users in the world, have brought over the need for terms to make these systems understandable, and with them, generated particularly contentious metaphorical debates. These metaphorical disputes matter because they shape both technical research directions and social expectations. The “stochastic parrot” metaphor, coined by Bender et al. (2021), characterizes LLMs as systems that “haphazardly stitch together linguistic forms without reference to meaning” (p. 7), sophisticated mimicry without understanding. This metaphor sparked intense debate, with critics arguing it smuggles in reductionist assumptions about LLM capabilities (Venkatasubramanian, 2025), while defenders maintain it accurately captures the absence of genuine comprehension behind human-like text generation. Competing metaphors, from “autocomplete on steroids” (Marcus, 2024) to “geometric understanding”, reflect ongoing uncertainty about how to accurately characterize systems that produce coherent text through statistical prediction rather than semantic understanding.

¹¹Though this is not inherently problematic: see Section 3.2.

¹²See, for instance, the work by Mittelstadt, Russell, and Wachter (2019), Burrell (2016), and Ananny and Crawford (2018), ACM guidelines https://www.acm.org/diversity-inclusion/words-matter?utm_source=chatgpt.com, Google Developer Documentation https://developers.google.com/style/word-list?utm_source=chatgpt.com and Apple style guide https://help.apple.com/pdf/applestyleguide/en_US/apple-style-guide.pdf?utm_source=chatgpt.com.

As AI systems become more sophisticated and widespread, the metaphors we use to discuss them increasingly function as implicit theories that guide development priorities, regulatory approaches, and public acceptance, making metaphor choice a form of technological politics with far-reaching consequences.

The metaphor crisis is the cognitive-linguistic face of a broader crisis in AI¹³: both reveal a gap between symbol manipulation and grounded conceptual understanding. Scientific metaphors highlight where this matters the most: when outdated analogies misshape theories and when AI tools cannot tell which mappings exist, which perspectives they highlight and which mislead. The reason metaphors have such powerful performative effects is precisely because they create emergent conceptual realities that cannot be reduced to their source and target domains.

In such a context, what are the solutions? Apart from the obvious response—more studies, more datasets, more empirical rigor and link to contemporary theoretical accounts of metaphor—two remedies are actionable: awareness and pragmatist evaluation. Awareness means surfacing when metaphors ossify into quasi-literal jargon, i.e. become constitutive, and either inventing replacements or, at minimum, explicitly naming the live metaphors we rely on (cf. Johnson, 1994, pp. 5, 22–23).

Secondly, the so-called pragmatist turn in conceptual engineering, which designates the shift from evaluating concepts in terms of their truth-conditional accuracy to assessing them in terms of their practical consequences for inquiry, enhances our ability to evaluate metaphors in different contexts. Rather than following the traditional assumption that metaphors should be (only) judged by their truth value or descriptive fidelity, the pragmatist approach includes a consequentialist evaluation: the adequacy of conceptual devices is not exhausted by their descriptive accuracy but by their consequences for inquiry, communication, and social coordination (Huetter-Almerigi et al., 2025). Practically speaking, we can extract and formalize patterns of metaphors into KPs, which by design imply participatory development and review, and therefore go in the direction of the aforementioned two solutions.

In this view, and recalling the *shared* aspects in ontology development (§2), knowledge engineers act less as policemen of metaphor, seeking universal replacement strategies, and more as stewards, attending to the situated effects of metaphor use across contexts. Such an orientation acknowledges both the productive power and the potential risks of metaphor, and recommends context-sensitive practices such as explicit metaphor awareness training, periodic metaphor audits, and the participatory evaluation of metaphors by diverse communities of practice.

3.2 AI for metaphor, metaphor for AI

This chapter has revealed a dual metaphor crisis: fragmentation in theoretical frameworks coinciding with uncritical persistence of harmful metaphors in scientific practice (§3.1). The AI-metaphor relationship offers a path forward for both dimensions: AI, and LLMs in particular, offer unprecedented capabilities for understanding figurative language at scale. Computational models can now analyze metaphorical

¹³By AI crisis, we mean the persistent gap between systems that excel at producing well-formed symbols and systems that exhibit grounded conceptual understanding of the world, as evidenced by brittleness under distribution shift, hallucinations, and shallow causal modeling. See, for example, Harnad (1990), Searle (1980), and Bender et al. (2021). Mollo and Millièrè (2023) recast symbol grounding for neural vectors and argue that referential grounding can be met by certain LLMs through causal-historical links, not necessarily through embodiment.

patterns across vast corpora, whether it being textual, image, or video, revealing usage trends and cultural variations (or misconceptions) that manual analysis cannot quickly capture. Beyond analysis, AI can detect novel metaphorical mappings in scientific literature as they emerge, potentially identifying conceptual innovations in real-time. However, this is not a primary track of inquiry: according to the framework of the four levels of interpretation shown in Table 3.1, we have seen current research is still mainly focused on the generic metaphor detection part, and not without challenges.

Inevitably, AI struggles with metaphors due to a mismatch in theory and implementation, an incompleteness of methodologies, and a general lack of heterogeneous datasets to test from, but also due to a general lack of systematic protocols to see if AI, in metaphor and concept making, bridges distant parts of its knowledge base beyond being a pattern matcher (Pezzulo et al., 2024) (§3.1.1.3).

In that fundamental question, metaphor and AI are bound together: generative AI systems do not simply mirror our understanding; they repackage these understandings. With mixed results (see Section 3.1), they suggest bridges between distant parts of the world-model we have uploaded into our data streams. This distinctively human form of epistemic self-engineering allows AI to elevate to a new level (Pezzulo et al., 2024).

Conversely, as AI systems possess specific, sophisticated established architectures, they can in turn be used to structure our own understanding, provided that a more nuanced notion of ‘understanding’ is required, one that transcends classical dichotomies and reflects the diverse capabilities of both living and artificial systems.

Metaphor for AI reveals itself as the ultimate cognitive benchmark, exposing fundamental limitations in current AI-based systems¹⁴. Understanding metaphors demands pragmatic, culturally-situated reasoning that exposes the brittleness of systems trained only on statistical correlations and that should instead obtain relational reasoning. More profoundly, metaphorical thinking exemplifies the analogical reasoning and creative conceptual combination that defines human cognition, operating at different levels from simple similarity detection to complex creative blending. As a skill test, metaphor processing asks at its very core: what metaphor is plausible to create from an human point of view? What cognitively embodied structures does a metaphor underline? The answers to these questions reveal whether AI systems can engage in the kind of meaning construction that emerges from interaction rather than fixed reference, the real path for developing truly intelligent systems.

As Barnden (2008) noted, AI has three aims: an engineering one, a psychological one, and the philosophical one. Metaphor is important for the engineering aim precisely because such a complex language phenomenon is pervasive but difficult to handle computationally. In that, the evaluation must go beyond quantitative metrics towards instead the indirect, long-term, and subjective sense of the extent to which the development contributes eventually to other fields such as Philosophy or Psychology, leading to satisfying the other two aims. And in the broader sense, “much of the point of developing a detailed computational account, implemented or not, is aid in the development of principles, methods and theories in more detail and with greater security than would otherwise be likely” (*ibid.*, p.5).

¹⁴As we have seen, Baudrillard noted it ahead of time—AI is dull because it lacks the power of metaphorical thinking (Baudrillard, 1990).

The reciprocal shaping of metaphor and computation creates a recursive enhancement loop: metaphor helps AI behave¹⁵ more like humans through analogical reasoning and creative conceptual combination, while AI helps us understand metaphorical processes more systematically and at scale than ever before. AI attempts to process metaphor reveal gaps in both computational architectures and theoretical understanding, forcing refinements in both domains. This recursive relationship suggests that advancing computational metaphor understanding thus is not merely about building better computational linguistics systems, it is about developing AI that engages with meaning through the same creative, analogical processes that characterize human thought.

Zlatev, Jacobsson, and Paju (2021) propose that a set of desiderata for any contemporary account of metaphor are: (i) to account for both communication and cognition, (ii) to explain both universal and culture-specific aspects, (iii) to achieve a balance between stable structures and contextual processes, (iv) to apply not only to different languages, but to other semiotic systems such as gesture and depiction (and their combination), (v) to provide clear theoretical and operational definitions.

We argue that the AI-metaphor studies cooperation is now able to address these requirements. Without AI, metaphor studies cannot be studied at scale in the contexts of (i) and (ii), but without a consensus on theoretical definitions, we would not have (v) nor would be able to know how to implement (iv) in practice.

In this landscape, (iii) in particular serves as the connecting ring between the other requirements, presupposing, in our view, both metaphor theory and AI working in tandem through the lens of KPs. For AI systems, they offer structured semantic scaffolding that moves beyond pattern matching toward genuine metaphorical understanding. For metaphor research, they provide standardized, machine-interpretable frameworks that allow testing whether proposed mappings reflect stable cognitive patterns or ad-hoc associations.

To answer our guiding RQ—what are the computational requirements for studying metaphors, generic or scientific—we mainly identify three needs. Data requirements demand domain-specific and cross-domain corpora with graded metaphoricity, explicit source–target mappings, blend properties, and pragmatic context. When metaphors are visual, these corpora must extend to visual modalities. Representational requirements call for a blending-oriented ontology aligned with frames that accommodates contextual and pragmatic cues while supporting executable constraints. Methodological requirements necessitate a goal of awareness and consequentialist accounts of metaphors. To satisfy this aim, we argue we need neurosymbolic pipelines with consensus-aware metrics for detection, and spanning interpretation and reasoning, together with strong reproducibility practices including versioned graphs, prompts, and tests.

The next chapter examines how large language models can assist knowledge engineering processes before Chapter 5 instantiates the program.

¹⁵Though this is obviously an anthropomorphic metaphor for AI, which as we have seen bears more limitations than benefits, and has thus to be taken as a merely rhetorically fit word for this context.

TABLE 3.1: Categorization of computational metaphor approaches by cognitive complexity and robustness axes.

Layer 1: Detection	Layer 2a: Domain identification	Layer 2b1: Interpretation	Layer 2b2: Entailment	Layer 3: Reasoning
<i>Binary classification</i>	<i>Source/target recognition</i>	<i>Paraphrase, meaning explanation</i>	<i>Multiple choice evaluation</i>	<i>Property extraction</i>
Fass (1991)	Mao et al. (2023)	Su et al. (2020), Rai et al. (2019)	Agerri (2008)	Lieto, Pozzato, and Zoia (2025)
Dunn et al. (2014)	Wachowiak and Gromann (2023)	Kim et al. (2023) [†]	Chakrabarty et al. (2022)	Confalonieri and Kutz (2020)
Mason (2004) [†]	Shutova (2017)	Yosef, Bitton, and Shahaf (2023) [⊗]	Stowe, Utama, and Muresan (2022)	Confalonieri et al. (2016), Bhatt, Hois, and Kutz (2012)
Tian, Xu, and Mao (2024)	Rosen (2018)	Kalarani, Bhattacharyya, and Shekhar (2024) [⊗]	Mohler, Tomlinson, and Bracewell (2013)	Carenini et al. (2024)
Li et al. (2023)	Barnden et al. (2004)	Chakrabarty et al. (2023) [⊗]	Saakyan et al. (2025) [⊗]	Veale (2019)
He et al. (2024a) [⊗]		Su et al. (2024) [⊗]		
Perlerin, Beust, and Ferrari (2003) [†]				
Hussain et al. (2017) [⊗]				
Demner-Fushman, Rogers, and Mork (2021) [†]				
Hicke and Kristensen-McLachlan (2024) [†]				
Wang (2022) [⊗]				

Robustness Axes:

Cross-Domain Adaptation (*same modality, different domains*): Scientific domains: Demner-Fushman, Rogers, and Mork (2021) (biomedical), Hicke and Kristensen-McLachlan (2024) (science texts), Kim et al. (2023) (science communication). Specialized domains: Perlerin, Beust, and Ferrari (2003) (stock market), Mason (2004) (domain-specific corpora).

Cross-Modal Transfer (*different modalities*): Visual metaphors: Ye and Kovashka (2018) (advertisements), He et al. (2024a) (visual metaphor detection), Hussain et al. (2017) (Pitt Ads dataset), Chakrabarty et al. (2023), Su et al. (2024) (visual metaphor generation). Multimodal: Wang (2022) (memes), Yosef, Bitton, and Shahaf (2023) (IRFL dataset), Kalarani, Bhattacharyya, and Shekhar (2024) (video metaphor captioning), Saakyan et al. (2025) (V-FLUTE visual entailment).

Legend: [†] = Cross-domain robustness; [⊗] = Cross-modal robustness.

Chapter 4

Methodologies for metaphor pattern extraction and automated knowledge engineering

Natural language artefacts such as scholarly articles, industrial standards, or everyday discourse, all encode regularities that can be formalised as patterns. As we showed in Chapter 2, when those patterns are expressed in logical form, they can become a bridge between unstructured text and AI, with the potential of improving complex tasks that require relational reasoning.

Chapter 3 has also shown that computational metaphor processing should depend on background commitments and plausibility constraints: mappings between a source and a target only carry over when the right vehicles and preconditions are present, otherwise the metaphor is not recognized.

In this chapter, we turn to LLM-assisted methodologies for knowledge pattern engineering, because they provide the scaffolding to make those commitments explicit, operationalize usage criteria from Semantic Web practice (Gangemi, 2005; Gangemi et al., 2016), and reduce the entropy of possible readings by fixing testable constraints on meaning. The chapter complements the operationalization of metaphor in the subsequent one by providing the foundation that makes them computable and extensible.

The aim of this chapter is thus to survey the methodological continuum that has emerged for pattern-based ontology engineering, from traditional methodologies to the contemporary LLM era. In particular, it aims to answer the following Research Question: *How has automation in knowledge engineering evolved with LLMs, and how can LLMs be systematically integrated into a reproducible, human-in-the-loop pipeline for KE automation of metaphor processing?*

The discussion around the answer is organised into three main parts. Section 4.1 begins with an overview of traditional methodologies, contrasting top-down ontology-driven frameworks with bottom-up, data-centric pipelines and highlighting the collaboration models and tool support that shaped early practice. The chapter then turns to AI-assisted extraction methods, and reviews the progression from neural discriminative models to contemporary LLM-based systems, showing how these advances have expanded the scope of ontology engineering tasks, from entity linking to full ontology generation and evaluation (Section 4.2). In particular, we focus on the Ontogenia method for automated ontology generation, a contribution of this thesis, and on methods for automated ontology evaluation (Sections 4.3-4.5). Furthermore, these automated methods require a reflection on the topic of reproducibility. This topic is examined in Section 4.6, in the context of LLM-aided knowledge engineering.

The historical survey reveals the persistent methodological tensions that have shaped the field's evolution and continue to influence contemporary AI-assisted approaches. The examination of traditional methodologies, in fact, exposes fundamental bottlenecks in requirements engineering and module implementation that have long hindered the scalability of ontology engineering. These limitations become particularly relevant when contrasted with the emergence of neural and LLM-based systems that promise to automate many of these traditionally manual processes, yet raise new questions about quality control, consistency, and human oversight. The transition from traditional to AI-assisted methods marks a paradigmatic shift in how we conceptualize the ontology engineering pipeline. Rather than viewing automation and human expertise as opposing forces, our investigation demonstrates how LLMs can serve as reactive engines within established methodological frameworks.

The Ontogenia methodology, presented in Chapter 4.3, exemplifies this synthesis, bridging the pattern-based foundations of eXtreme Design with the generative capabilities of LLMs. This approach preserves the theoretical soundness of traditional methodologies while addressing their scalability limitations through strategic

automation. The empirical investigation that follows this methodological development unfolds across three interconnected studies, each addressing a distinct aspect of LLM-assisted ontology engineering. The initial Ontogenia studies establish the fundamental feasibility of generating ontologies that meet functional requirements, demonstrating that LLMs can produce outputs at least comparable to novice human engineers when properly guided by metacognitive prompting and ontology design patterns. This foundation enables a more systematic comparative analysis that positions LLM capabilities against established human benchmarks, revealing both the strengths and persistent limitations of automated approaches.

The investigation proceeds with domain generalization studies that test the robustness of these findings across diverse knowledge areas, establishing the broad applicability of LLM-assisted methods while identifying domain-specific challenges. We then move on to examine the practical integration of LLMs within human-centered knowledge engineering workflows. The OE-Assist framework and associated user studies (§4.5) reveal the complex dynamics of human-AI collaboration in ontology evaluation, demonstrating both the potential for enhanced efficiency and the risks of automation bias. These findings confirm the previous studies' intuition that there needs to be a balance to harness LLM capabilities while preserving human judgment and expertise, providing concrete design principles for future human-AI collaborative systems in knowledge work.

However, the promise of LLM-assisted ontology engineering cannot be fully realized without addressing fundamental questions of reliability and reproducibility. The CoLLM framework emerges from this necessity, providing the first systematic approach to assessing consistency in LLM-assisted knowledge engineering. This framework addresses critical gaps in current research practices by establishing empirical methods for evaluating repeatability, update impact, and model outputs replicability across different LLM-based knowledge engineering tasks. The findings from this analysis reveal both encouraging levels of overall reproducibility and concerning brittleness to model updates, establishing clear guidelines for responsible deployment of these technologies. Throughout this progression, each component builds systematically toward a comprehensive understanding of how LLMs can be integrated into reproducible, human-in-the-loop pipelines for knowledge engineering automation. The chapter then aims to establish both the theoretical foundations and practical guidelines necessary for the responsible advancement of AI-assisted ontology engineering, and in the context of the thesis, will provide a backbone for the next one.

4.1 Traditional methodologies

The practice of ontology engineering has been shaped by a variety of methodological traditions. These long-standing methodologies can be grouped into two broad families. *Top-down* approaches begin with an explicit conceptual framework of reference, often guided by domain experts or by reuse of existing upper ontologies, and then formalise the resulting structure in a description-logic language. *Bottom-up* approaches invert the workflow: they analyse corpora or databases first and allow the schema to emerge from observed linguistic or statistical regularities.

This section reviews the principal representatives of each framework, highlighting their development workflows, collaboration models, and tool support, and identifying persistent bottlenecks that later AI-assisted methods aim to alleviate.

4.1.1 Top-down approaches

Ontologies are not only software components, they are also information artefacts. The quality of ontologies is indeed closely linked to their design, which must accurately describe a domain and fulfil a purpose that involves identifying elements considered in the same case, and undergo a life cycle. Despite their critical role in the operation of numerous infrastructures across both industry and academia, one of the main challenges preventing the widespread adoption of ontologies is the complexity of their development. In fact, this process requires interdisciplinary expertise and practical guidance (El Kadiri and Kiritsis, 2015).

In the realm of engineering ontology applications, a predominantly top-down construction approach is embraced. It involves leveraging existing ontologies or patterns to derive structure and applying it to available data, typically achieved through comprehensive domain analysis and interviews with engineers (Park et al., 2008).

As the Semantic Web gained increasing attention towards the late Nineties, researchers embarked on creating numerous Ontology Engineering Methodologies (OEMs) aimed at providing comprehensive and clear directions for ontology construction. There seems to be consensus among all OEMs on the essential activities required for ontology engineering: beginning with a feasibility study to assess the utility of an ontological approach for the problem at hand, followed by domain analysis to identify relevant domains and gather knowledge for formalisation, leading to the conceptualization phase where relationships among concepts are established, and culminating in the actual implementation through formal languages, resulting in a prototype ontology ready for maintenance and application. However, the specific approaches to these activities vary significantly. Each OEM adopts unique approaches for specific tasks, leading to a variety of classifications based on the organisation of workflows, the degree of collaboration, and the type of support provided. Workflow strategies among OEMs extend widely, from the linear steps of the waterfall model, which enforce a strict order from a feasibility study to the release of prototypes, to lifecycle approaches that treat ontologies as continuously evolving entities without a fixed development sequence, and agile methodologies that provide ontology engineers with a flexible set of actions to tackle as necessary.

The now stable trend towards cooperative development underscores the perception of ontology engineering as a process aimed at creating a really shared formal representation of a domain (see Section 2.1). The success of the involvement of experts of different knowledge highlights the essential role of domain experts and stakeholders in determining concepts and relationships for a more efficient creation of ontologies alongside ontologists, and really makes a difference. In fact, a key finding from Kotis, Vouros, and Spiliotopoulos (2020)'s survey of OEMs is that non-collaborative OEMs negatively affect the vitality, evolution, and reusability of ontologies. In contrast, collaborative and tailored OEMs, to positively influence the condition of developed ontologies, must be supported by comprehensive collaboration tools for the development, upkeep, and progression of ontologies in a modular fashion. Furthermore, OEMs vary in their emphasis, with some focusing on providing broad guidance for moving from informal domain representations to formal structures, and others on offering detailed, or authoring, methodologies support to help choose the most appropriate logical frameworks for knowledge modelling. In the following section, we revise main OEMs.

4.1.1.1 Early methodologies

The field of ontology engineering has evolved significantly since its inception, with early methodologies establishing foundational principles that continue to influence contemporary approaches. These frameworks emerged primarily in the late 1990s and early 2000s, responding to the growing need for systematic processes to develop knowledge representations across various domains. METHONTOLOGY (Fernández-López, Gómez-Pérez, and Juristo Juzgado, 1997) was one of the early structured methodologies, proposing a waterfall-like process with phases such as specification, knowledge acquisition, conceptualization, integration, implementation, evaluation, and documentation. Despite its structured approach, METHONTOLOGY introduced flexibility with overlapping phases, emphasising a comprehensive yet adaptable process for ontology development.

The Ontology Development 101 guide by Noy, McGuinness, et al. (2001) laid down a foundational framework pre-dating the widespread adoption of OWL. It described a traditional, sequential method comprising domain definition, reuse consideration, term enumeration, class hierarchy development, property definition, and instance creation. Although detailed in several steps, it lacked emphasis on requirements gathering and testing. DILIGENT (Pinto, Staab, and Tempich, 2004) was instead based on theories of argumentation, focusing on empowering domain experts in ontology engineering through continuous and distributed construction and updating of ontologies. It highlighted the importance of domain expert involvement and the collaborative nature of ontology development. This same participation of knowledge workers in the ontology lifecycle was advocated by the Human-Centred Collaborative Ontology Engineering Methodology (HCOME) (Kotis and Vouros, 2006). It emphasised a human-centred approach, supporting collaboration between domain experts and ontology engineers through direct and continuous interaction with conceptualizations. Ontology-Grounded Methods and Applications (DOGMA) (Jarrar and Meersman, 2009) and its extension DOGMA-MESS (De Leenheer and Debruyne, 2008) followed a data-driven, bottom-up approach to ontology engineering, enabling the modelling of shared ontologies in stakeholders' terminology. It introduced a collaborative and iterative development process driven by social knowledge conversion modes. Melting Point (Garcia et al., 2010), drawing from best practices in earlier methodologies, emphasised collaborative development by communities of practice without prescribing specific techniques, thus recommending that ontology developers consider those methods best suited to their particular situations.

4.1.1.2 Collaborative and agile methodologies

Agile OEMs have emerged as a response to the demand for swift and collaborative ontology development in decentralised environments. These methodologies highlight the significance of stakeholder collaboration and the vital role of domain experts in shaping the knowledge base and underlying conceptual framework. Furthermore, the flexible and iterative development of ontologies that allows rapid usage and consumption of ontologies by users or domain experts has become increasingly central after the first OEMs. RapidOWL (Auer and Herre, 2006) had already emphasized as a methodology the agile elicitation and structuring of knowledge to foster cooperation among stakeholders, ontology engineers, and domain experts. Its approach is distinctly non-prescriptive, granting participants the autonomy to select and sequence activities as they see fit, without any explicit dependency among them,

thereby enhancing the methodology's adaptability and responsiveness to project-specific needs.

Today, different studies continue to agree on the fact that agile OEMs efficiently scale down the involvement of engineers, directly addressing task-related hurdles such as knowledge retrieval, documentation creation, and conceptualization, while also shortening the learning curve attributed to their flexible structure (Vigo et al., 2014). Agile OEMs facilitate the process without necessitating stakeholders and domain experts to possess ontology modelling expertise. Instead, they recommend the use of supportive tools for identifying and conceptualising knowledge, emphasising the agile paradigm's focus on intervention areas, ultimately guiding ontology engineers toward prototype development.

A facet-based OEM for Geospatial Ontologies (FMCLGO) (Giunchiglia et al., 2012) adopted an approach for the development of large-scale geospatial ontologies. It focused on manual and automatic processes for identifying, categorising, and populating ontologies, suggesting a top-down methodology to ontology engineering.

Contrary to other methodologies that offer structured guidance for ontology engineering, the NeOn Methodology (Suárez-Figueroa, Gómez-Pérez, and Fernández-López, 2011) presents a flexible approach, proposing various paths for ontology development. Introduced in 2008, NeOn enriches the ontology engineering methodology landscape. Compatible with NeOn is the Grounding Ontologies with Social Processes and Natural Language (GOSPL) (Debruyne, Tran, and Meersman, 2013). It can be used within NeOn scenarios to operationalize the social-collaborative part of ontology work, i.e., how stakeholders negotiate and document definitions, reach consensus, and evolve artefacts, while NeOn provides the broader process guidance for specification, scheduling, reuse and integration.

NeOn has had particular success in that it encourages the collaborative creation of ontologies and defines the reuse and re-engineering of existing knowledge. A notable feature of the NeOn methodology is its comprehensive library of Ontology Design Patterns (ODPs) and the provision of tool-supported methodological guidance, known as eXtreme Design (Presutti et al., 2009), to facilitate ontology development. From that moment onwards, agile methodologies focusing on the re-use of ODPs (Blomqvist and Sandkuhl, 2005; Gangemi, 2005), such as eXtreme Design itself, SAMOD, and MOMo (Blomqvist, Hammar, and Presutti, 2016; Peroni, 2016; Shimizu, Hammar, and Hitzler, 2023), have become increasingly popular, reflecting the needs of many real-world settings by maximising the cognitive soundness, logical correctness and effectiveness of the derived ontological artefacts. For example, Rexplore and the Computer Science Ontology provide measurable improvements in discovery and analytics over scholarly data, illustrating how KE artefacts show real-world utility (Salatino, Osborne, and Motta, 2018).

Among the various agile OEMs, SAMOD (Peroni, 2016), the Simplified Agile Methodology for Ontology Development, stands out as it treats ontologies as evolving prototypes. It begins with a motivating scenario from which requirements are collected through competency questions, leading to the development of a model or "micro-ontology" that is iteratively refined and expanded until it meets all the ontology's requirements. This iterative process encourages the incorporation of additional complexity or scenarios to ensure the final model comprehensively answers all competency questions, symbolising the completed ontology.

Instead, in Poveda-Villalón et al. (2022) the Linked Open Terms (LOT) methodology specifically focuses on the alignment with industrial development and emphasizes the importance of precise documentation and the update of these methodologies over the years when new lessons were learned, with respect to existing methodologies such as XD or SAMOD.

Despite the inevitable differences among OEMs, all agile approaches show a shift towards more dynamic, collaborative, and user-engaged processes in ontology development. By prioritising flexibility, stakeholder involvement, and the use of established design patterns and methodologies, agile OEMs aim to make the ontology development process more accessible and responsive to the rapidly evolving requirements of diverse domains.

4.1.1.3 The eXtreme Design Methodology

Among agile OEMs, the eXtreme Design methodology stands out for its success. Since the very beginning, XD was defined as a collaborative, incremental, iterative method for pattern-based design. At the heart of XD's strategy are Ontology Design Patterns, which serve to address common and recurring challenges in design.

After the initial theoretical development, core principles were progressively established further alongside tool development, and in 2016 the methodology was matured. Multiple large-scale projects have now successfully used this methodology for more than a decade, such as the ArCo Cultural Heritage Knowledge Graph (Carriero et al., 2019), the Water Health Open Knowledge Graph (WHOW) (Lippolis, Lodi, and Nuzzolese, 2025), the Polifonia ontology network (Berardinis et al., 2023), and many others.

XD implements a structured 12-task iterative workflow that goes from user stories to tested ontology modules through systematic pattern reuse.

The process starts with an intensive effort to gather requirements, facilitated through close collaboration between the customer and design teams. This phase aims to define the project scope, objectives, and stakeholder roles and for collaborative requirement gathering it often employs the use of stories, a concept borrowed from agile software development, to clearly and effectively communicate the needs and expectations for the ontology.

Following the collection of requirements, these stories are then transformed into competency questions (CQs), which act as a natural language representation of the ontological commitments that will guide the development process. These questions are important in defining the scope and functionalities that the ontology is expected to support, ensuring that the design remains aligned with the intended use cases. These stories can be more or less refined and fine-grained depending on the use case at hand. For example, as shown in Table 4.1.1.3, in a music ontology about the use of bells, a user story could involve the definition of a persona, a goal, and a scenario: the first one describes who the person involved in the user story is, the second one their main aim as to why they need the resource, and then the third one the actual usage scenario. Examples of CQs could be: "Are there bells in the church/bell tower? How many bells are there? By whom (by which foundry) were they cast? Which is the material of the bell?" and others¹.

The inclusion of user stories and CQs as requirements has multiple advantages. First of all, it breaks complex narratives into atomic facts that allow then to transform concrete instances into ontological concepts. Furthermore, these requirements are

¹This example was derived from the Polifonia project https://github.com/polifonia-project/stories/blob/main/Keoma_Architect.

TABLE 4.1: User Story Example: Music Ontology for Bell Usage.

Component	Description
Persona	Keoma is an architect and a restorer
Goal	Keoma needs to find different types of information and data from different sources. His goal is to interconnect data between architectures and their use for sound performances, in order to proceed with protection actions. • sound practices • restoration
Scenario	Keoma has to elaborate a restructuring plan for various bell towers and wants to know what kind of interventions to implement on the wall structure respecting the characteristics related to the use of the bell cell as a component of the production of the sound of the bells, and respecting traditional local sound practices. Keoma therefore needs to know what kind of sound practices are performed in each bell tower, starting with the names of the Churches given in a list. Keoma begins to search for the location and the address of each church, also to understand if it is an urban, peri-urban or rural context. Keoma needs to know if in each of the churches the practices of sound are currently practiced by moving the bells manually or electrically. Furthermore, for static calculations he needs to know with which mounting system each bell is attached, its weight, its measurements and the material it is made of. Keoma also needs to know if there are sound documents of the musical performances carried out in the different bell towers, to document the use of the bells in different events. To do this, Keoma would have to resort to numerous searches: bibliographic sources, archival sources, field surveys and interviews with informed people. Thanks to Polifonia, Keoma can carry out these searches automatically, through a keyword search, retrieving these data in a short time.
Competency Questions	Are there bells in the church/bell tower? How many bells are there? By whom (by which foundry) were they cast? Which is the material of the bell?

testable. CQs in particular drive both ontology modeling and validation and allow for functional ontology testing (Gangemi et al., 2006).

After the requirements elicitation phase, there is a pattern matching phase in which recurring modeling problems within CQs are identified and CQs are matched to established generic use cases that represent common ontological challenges; these are then searched for pattern repositories.

ODPs can be selected in multiple ways. Pattern operations in fact involve either: specialization, so the creation of domain-specific subclasses from generic patterns; composition, by combining multiple ODPs to address the requirements; import, which directly includes the pattern with full dependency; or instantiation, creating instances following the pattern with sample data.

The ontology module development can apply both direct and indirect reuse approach; in any case, it has to ensure proper alignment and consistency across modules as well as document all modeling decisions and pattern applications.

Finally, each CQ can be converted into SPARQL queries that serve as unit tests. The validation phase goes beyond SPARQL, as CQs are verified against real scenarios, but also there is reasoning validation using logical inference. This is the CQ verification phase, in which each CQ is operationalised as a reproducible test case and executed against (i) the asserted ABox and (ii) the ontology's inferred model. Because each CQ is meant to be tested in isolation, the procedure functions as a true unit-testing regime for ontologies, mirroring software engineering best practices and enabling fault localisation at module level (Blomqvist, Seil Sepour, and Presutti, 2012).

As the process is inherently iterative, the methodology is designed for ongoing refinement. New CQs can be introduced as requirements evolve, and insights from feedback or validation may reveal the need to adjust or create additional modules. To govern these changes, XD incorporates robust version-control and release-management practices. Its modular architecture also allows individual modules to be updated or replaced independently, minimizing disruption to the broader system.

These core principles distinguish XD from other approaches: customer involvement ensures continuous domain expert feedback, collaborative programming practices are encouraged, test-driven development mandates unit testing before module integration, and the intensive use of ODPs provides proven modeling solutions. Therefore, rather than comprehensive domain coverage, XD emphasizes task-oriented design focusing on specific user requirements, in other words, prioritizes the need of the user. The methodology also has a consequence on cognitive load management, since patterns reduce design complexity while assuring quality by being proven solutions.

Before XD, pattern-based ontology engineering primarily focused on logical ODPs: research such as OntoCase showed how these logical patterns could be discovered automatically and injected into an ontology to learn missing classes or to enrich the axiom set (Blomqvist, 2009a). Logical-only support proved insufficient when it was necessary to reuse richer domain fragments, such as content patterns, track versions, and iteratively evolve modules. With XD, each ODP can be updated independently without breaking the rest of the design.

The development of the Water Health Open Knowledge Graph (WHOW-KG) presented by Lippolis, Lodi, and Nuzzolese (2025) exemplifies the XD methodology application in a real-world context and successfully illustrates the methodology's practical effectiveness. The project aimed at fostering the analysis and development

of innovative services concerning water consumption and quality, health parameters and dissemination of diseases through the development of the WHOW-KG. The project's collaborative approach involved 77 participants in a co-creation program across EU countries, embodying XD's emphasis on stakeholder involvement and distributed expertise. This foundation supported use case-driven development among the project partners, which had both technical expertise and environmental expertise, focused on three specific scenarios: marine water contaminants, water quality for human consumption, and meteorological extreme events, each representing distinct but interconnected aspects of water and health monitoring.

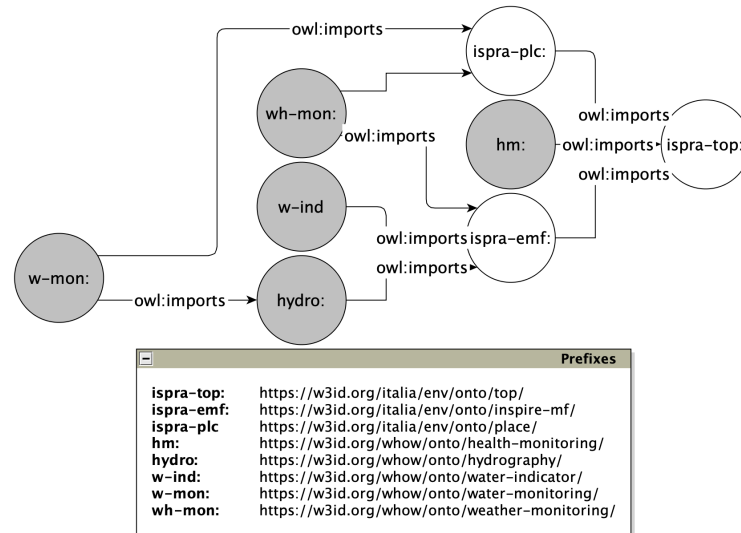


FIGURE 4.1: The WHOW ontology network.

As shown in Figure 4.1, the pattern-based design philosophy central to XD is, in fact, evident throughout the WHOW-KG architecture, with systematic reuse of established ODPs including the PartOf ODP² for expressing water basin relationships, the Stimulus-Sensor-Observation ODP (Janowicz and Compton, 2010) for structuring monitoring data, and the Indicator ODP³ for representing water quality metrics. This pattern reuse facilitated the creation of a modular architecture comprising a network of five interconnected ontologies: Hydrography⁴, Water Monitoring⁵, Water Indicator⁶, Weather Monitoring⁷, and Health Monitoring⁸, each designed to minimize coupling while maximizing internal cohesion.

The implementation's adherence to XD's testing principles is demonstrated through systematic SPARQL-based validation, where competency questions derived from user stories were then converted into formal SPARQL queries and executed against data to ensure the ontology network's effectiveness in addressing stakeholder requirements, ultimately resulting in a knowledge graph containing over 265 million RDF triples distributed across multiple endpoints.

From the application of XD to the WHOW-KG and other projects, the success of the methodology is apparent, and in fact, empirical validation studies on the matter

²<http://ontologydesignpatterns.org/wiki/Submissions:PartOf>

³<https://schema.gov.it/lodview/onto/Indicator>

⁴<https://github.com/whow-project/assets/tree/main/ontologies/hydrography>

⁵<https://github.com/whow-project/assets/tree/main/ontologies/water-monitoring>

⁶<https://github.com/whow-project/assets/tree/main/ontologies/water-indicator>

⁷<https://github.com/whow-project/assets/tree/main/ontologies/weather-monitoring>

⁸<https://github.com/whow-project/assets/tree/main/ontologies/health-monitoring>

confirm this claim. In fact, it has been shown that XD provides helpful means to organize the design process with drastic decrease of frequent mistakes through testing focus (Blomqvist et al., 2010). Furthermore, the study shows significant reduction in modeling errors, improved ontology quality metrics, and better task coverage compared to traditional approaches, though these benefits come with learning curve requirements.

Alongside learning curve requirements, one of the limits of XD is scalability and adoption challenges, since collaboration is resource-intensive especially in terms of time and in terms of expert resources, but also there is an increasing difficulty for pattern discovery as repositories grow. Also, XD is set within broader ontology engineering challenges, when comparative analysis suggests that no methodology enjoys consensus among the community (ElHassouni and Qadi, 2021). XD in particular requires additional debugging and matching capabilities not inherently part of the core methodology, needing supplementary approaches for concrete implementation.

4.1.1.4 Modular ontology modeling after eXtreme Design

The use of patterns is still central to ontology development. Building upon XD, MoMo extends it with graphical modeling approaches and tooling support (Shimizu, Hammar, and Hitzler, 2023).

MoMo's contribution lies in making ontology engineering accessible to non-experts through visual modeling paradigms while preserving the modular, pattern-based principles proven effective in XD.

Similarly to XD, MoMo operationalizes modular ontology development in ten steps. The methodology begins with use case description and CQ gathering, then proceeds through key notion identification and pattern selection before moving to graphical module design and integration. This process distinctively emphasizes graphical schema diagrams as primary modeling artifacts rather than post-facto documentation. Schema diagrams are created for individual modules and then modules are integrated into unified diagrams. The diagram-first approach enables non-expert participation in ontology modeling while deferring formal OWL until the final step.

Thus, the methodology's theoretical foundations are the same as XD, which emphasize the four core principles of modular development, pattern-based reuse, collaborative knowledge elicitation, and human-centered design.

Furthermore, MoMo adds a suite of tool support. It provides comprehensive support through CoModIDE⁹, enabling drag-and-drop graphical editing, built-in ontology design pattern libraries, and automatic OWL generation from visual representations. In a user study with 21 subjects, the authors have seen using CoModIDE improves task completion rate while reducing task completion time, compared to using standard Protégé (Shimizu, Hammar, and Hitzler, 2020).

MoMo addresses the bottleneck of limited readily available ontology design patterns through modular ontology design libraries: curated, internally consistent collections of well-documented patterns that can be programmatically queried and accessed. The methodology implements a OPLa (Ontology Pattern LAnguage) annotations to maintain module provenance metadata, enabling tracking of pattern origins and modifications.

However, this and recent developments that address XD and ODP-based design's infrequent updates and insufficient focus on precise documentation (Poveda-Villalón et al., 2022), appear to have been largely addressed in the ever increasingly

⁹<https://comodide.com/>

publication of patterns and the possibility to derive them in a bottom-up way. In the next chapter, we will investigate this approach more in detail.

4.1.2 Bottom-up data-driven approaches

The objective of bottom-up knowledge engineering pertains to the augmentation and refinement of knowledge graphs through the use of open information extraction (OIE) methodologies. In bottom-up approaches, differently to top-down approaches where the ontology or schema is designed first, and then data are fitted to it, the process begins with data acquisition; the ontology emerges from analysing the data. In this section, we aim to provide a concise exposition of extant OIE techniques, accompanied by an analysis of their respective advantages and limitations.

4.1.2.1 Traditional Open Information Extraction

OIE is the task of extracting relational tuples directly from unstructured text without requiring a predefined relation schema. Before the explosion of neural networks, most strategies used for OIE systems relied on the structure of the text, for example, rule-based system, learning-based system, or clause-based system. Rule-based methods use linguistic patterns or rules to identify the semantic class of a relation phrase. For instance, REVERB (Fader, Soderland, and Etzioni, 2011) integrates two constraints, a syntactic constraint and a lexical one. The syntactic constraint imposes restrictions on relation phrases to adhere to specific patterns, mitigating errors such as the inadvertent concatenation of distant words. Complementarily, the lexical constraint facilitates the refinement of extractions by reducing overspecified outputs. These constraints yield substantial improvements in both precision and recall when compared to earlier extractors such as TEXTRUNNER (Yates et al., 2007).

Learning-based methods employ supervised or semi-supervised learning techniques to train a classifier on pre-defined relations. For instance, OLLIE (Schmitz et al., 2012) represents a learning-based Open Relation Extraction (ORE) system that acquires unlexicalized pattern templates from bootstrapped training data. OLLIE adeptly extracts relational tuples from text without necessitating a pre-specified vocabulary. It surpasses earlier systems by capturing relations mediated by nouns, adjectives, and various syntactic forms, while also integrating contextual cues from the sentence into its extractions. A clause-based OIE system identifies clauses in a natural language sentence and then extracts relations and their arguments from each clause. For example, Angeli, Premkumar, and Manning (2015) propose an approach to produce relation triples from natural language sentences, leveraging the linguistic structure to split sentences into self-contained clauses, and then apply natural logic inference to extract the maximally specific arguments for each candidate triple. Another proposed method, called ClauseIE (Del Corro and Gemulla, 2013), handles complex sentences with multiple clauses, nested clauses, and non-clausal modifiers. From the perspective of the source of the information, most of these OIE systems require shallow features, which could span from lexical, syntactical, to hierarchical dependencies.

Bottom-up knowledge exploration systems complement OIE by guiding sense-making before formalisation. Aemoo (Nuzzolese et al., 2016) operationalises Encyclopedic Knowledge Patterns mined from Wikipedia's link structure to select, organise, and visualise salient, entity-centric summaries over Linked Data, effectively supporting facet discovery and CQ elicitation without a predefined schema. In the same research line, FRED transforms natural-language text into OWL/RDF graphs,

enforcing ontological constraints during extraction and thus bridging flat OIE tuples with pattern-based design (Draicchio et al., 2013; Gangemi et al., 2017). Together, these systems exemplify bottom-up, data-driven workflows that anticipate today's AI-era pipelines. Here, the schema emerges from data, while discovery and enrichment operate over the live Semantic Web through services that locate and inspect ontologies at web scale (d'Aquin and Motta, 2011).

4.1.3 Towards the AI era

So far, we have traced the evolution of traditional and agile ontology engineering. From the review, we can conclude that while these frameworks provide robust, pattern-based workflows that require little annotated data, they struggle with long-range dependencies and unseen patterns and performance is heavily tied to handcrafted rules, creating significant bottlenecks in requirements engineering and module implementation.

This persistent challenge of scalability and automation directly motivates the turn towards AI-assisted techniques: topic classification pipelines operating over ontologies show the practical gains of KE at scale (Osborne et al., 2016; Salatino, Osborne, and Motta, 2020). The next section will explore how to leverage machine learning and large language models to address these very limitations.

4.2 AI-assisted extraction methodologies

Recent advances in representation learning have substantially broadened the range of techniques available for pattern extraction and knowledge graph construction. Early neural architectures improved upon rule-based systems by learning distributed representations and optimising end-to-end for span labelling or sequence generation. However, these approaches remain fundamentally discriminative and operate largely in isolation from formal ontology engineering. The more recent LLMs (see §3.0.4.3) provide a faster, accessible alternative: these systems can read unstructured text, satisfy symbolic constraints, and produce ontology-conformant axioms within a single generative framework. The research community has in fact responded to the surge of LLMs via dedicated tracks on LLM-enabled knowledge engineering at major Semantic Web conference venues such as ESWC, EKAW, and ISWC, alongside journal ones¹⁰. The present section first reviews the neural baselines, then surveys the emerging body of LLM-driven work for ontology engineering tasks, covering entity linking, requirements generation, ontology learning, generation and evaluation.

4.2.1 Neural methods

With advancements in neural networks and pre-training methodologies, there arises an opportunity to delve into deeper semantic representations, such as embeddings (see §3.0.4.2), moving beyond reliance solely on surface-level features like Part-of-Speech tags. This evolution enables an improved understanding and extraction of information from textual data. Overall, information extraction can be defined as

¹⁰ESWC: <https://2024.eswc-conferences.org/>; EKAW: <https://event.cwi.nl/ekaw2024/cfp.html>; ISWC: <https://iswc2024.semanticweb.org/>; Semantic Web Journal: <https://www.semantic-web-journal.net/blog/special-issue-large-language-models-generative-ai-and-knowledge-graphs>.

two problems, the discrimination (classification) problem and the generation problem. Considering the input information, these methods extensively use semantic information. They commonly rely on word and sentence representations derived from embedding models (e.g. ELMo) or pre-trained language models.

4.2.1.1 Discriminative neural approaches

Within a sequence-to-sequence framework, a Transformer-based language model receives unstructured text as an input sequence and autoregressively generates an output sequence that linearly encodes subject–predicate–object triples. By learning this mapping end-to-end, the model simultaneously internalises entity recognition, relation classification, and triple aggregation, eliminating the need for hand-engineered pipelines.

Sequence-to-Sequence Models emerged with Neural OIE. In 2018, the first encoder-decoder approach with attention mechanisms for unknown tokens was implemented. This approach established neural paradigms despite challenges with noisy training data (Cui, Wei, and Zhou, 2018).

In later years, BERT-based systems achieved breakthrough performance with IMoJIE (Kolluru et al., 2020a) and Multi²OIE, which supported multilingual OIE (Ro, Lee, and Kang, 2020).

Span-based neural models also advanced the field with SpanOIE, the first span-based model for n-ary OIE using BiLSTM with span representation (Zhan and Zhao, 2020). Recent neural systems include OpenIE6 (Kolluru et al., 2020b), which outperforms baselines due to the use of iterative grid labeling architecture treating OIE as 2-D grid labeling task, and QuORE, which implements query-based framework for single/multi-span relation extraction with template-based query mapping and dynamic sub-module selection (Yang et al., 2022).

4.2.1.2 Generative Sequence-to-Sequence approaches

Generative methodologies that employ the sequence-to-sequence framework for constructing knowledge graphs are noted for their adaptability and versatility across diverse tasks. In essence, these approaches conceptualize the extraction process as a generation problem, wherein the sequential generation of linearized entities and relations within a triplet occurs autonomously through autoregressive mechanisms. UIE (Lu et al., 2022) is a unified text-to-structure generation framework aimed at addressing diverse information extraction tasks while enabling the generation of targeted structures and the acquisition of general IE capabilities from multiple knowledge sources. DEEPSTRUCT (Wang et al., 2022) aims to enhance the structural comprehension capabilities of language models in structure prediction tasks. By pre-training language models on task-agnostic corpora to generate structures from text, DEEPSTRUCT facilitates the transfer of acquired knowledge to downstream tasks and exhibits superior performance across 21 out of 28 evaluated datasets among 10 structure prediction tasks, without necessitating task-specific architectures or data augmentation.

The Linked Open Knowledge Extractor (LOKE) (McCusker, 2023) addresses fundamental alignment issues between traditional OIE outputs and existing KGs, finding that the alignment of OIE results with existing KGs to be inadequate. Using GPT models with prompt engineering for Wikidata construction, LOKE-GPT outperforms baselines, although it over-generates triples compared to the reference set due to overall triple scarcity in the original set.

OIE research thus shows how far we can get by treating knowledge acquisition as a surface-level problem: identify text spans and output triples. However, OIE-related work tends to encounter three structural gaps that ontologies must fill. First of all, OIE terms rarely match the canonical classes and properties needed in a curated ontology. Secondly, triples neither capture nor enforce constraints such as disjointness, cardinality, or property domains. Finally, flat triples ignore higher-order context such as temporal scope or provenance that an ontology can encode explicitly.

These gaps make OIE an ideal launch-pad but an insufficient destination for knowledge engineering. The latest sequence-to-sequence and prompt-based systems already blur the line between extraction and generation systems, which suggest a natural next step of leveraging LLMs for knowledge engineering tasks. For instance, ontology-guided tripled extraction demonstrates significant advancement in the work by Cauter and Yakovets (2024). The proposed approach addresses the challenge that large-scale KG construction remains unfeasible since it requires significant human-expert involvement: by using only 20 semantically similar examples with a Llama-based LLM, and performance comparable to previous methods was achieved. The following subsection surveys that emerging body of work.

4.2.2 LLM-enhanced ontology engineering

LLMs have pushed knowledge engineering beyond the focus of OIE. Because an LLM can both read (in that, it is massively trained on) natural language and reason over symbolic constraints, it can now serve as a unifying substrate for ontology engineering, a spectrum of tasks that stretches from schema refinement to full ontology generation.

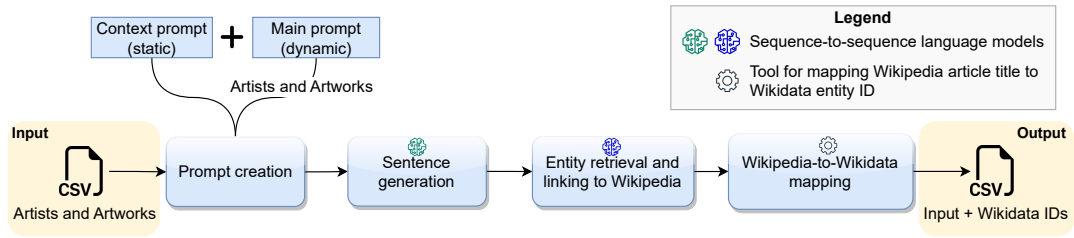
In general, existing work on the use of LLMs for the ontology domain has involved different tasks such as ontology augmentation (Zaitoun et al., 2023), ontology matching (Hertling and Paulheim, 2023), and ontology alignment (He et al., 2023). In this section, we survey how recent work makes use of LLMs for ontology engineering tasks. Specifically, we focus on entity linking, requirements engineering, ontology learning and population, and ontology generation.

4.2.2.1 Entity linking

Entity linking was one of the first examples of how LLMs could move beyond text generation into KG engineering. Classic Entity Linking pipelines achieve high recall, but they struggle when surface forms are sparse (e.g. names in minority languages) or when the local context is too thin. Prompting an LLM lets it exploit encyclopaedic background knowledge and multi-step reasoning, often flipping low-confidence or out-of-vocabulary cases into correct matches.

In particular, some approaches aim to improve embedding-based alignment methods by enriching alignment seeds (Chen et al., 2024; Zhang et al., 2023b), while other studies prompt LLMs to detect equivalent entities, often within a retrieval-then-prediction framework to reduce computational inference costs. ChatEA (Jiang et al., 2024) confirm the viability of this two-stage retrieve, then reason pattern on benchmark KGs. Similar approaches are introduced in Yang, Cheng, and Zhang (2024) and Jin et al. (2025).

In Lippolis et al. (2023), LLMs have been exploited to link the art-centric ART-GRAPH KG to Wikidata. The goal was to find for each artist and artwork in the graph, the corresponding Wikidata entity so that the two graphs could be integrated and gaps could be filled, such as information about genre or missing images. The

FIGURE 4.2: *pArtLink* architecture diagram.

workflow outlined in Figure 4.2 proceeds in two consecutive stages. It begins with Wikidata Entity Search (WES): a combined SPARQL / ElasticSearch query issues up to ten candidate Wikidata identifiers for every ARTGRAPH label (whether an artwork or an artist) optionally constraining the search by creator names when those are provided. Any mentions that remain unresolved after this initial retrieval are then handed off to *pArtLink*, a LLM post-filter built on Llama 2. In a single prompt, the model must choose the most plausible candidate, justify its decision in natural language, and return the special token `cannotLink` whenever the evidence is too weak to support a confident match.

Applied to the full corpus, WES alone succeeded in recovering Wikidata IDs for 14,982 artworks and 2,029 artists. The *pArtLink* stage contributed a further 76 difficult artist links that the search layer had missed, raising overall artist coverage by 3.7 percent. A manual audit of a selected sample across five annotators, with 95% percentage agreement for both artists and artworks cases, places the precision above 92 percent, with the majority of residual errors attributable to genuine name ambiguities rather than model hallucinations. In a similar way, EARAG, proposed by Yang et al. (2025), is a method that integrates a retriever for identifying candidate equivalent entities and employs a prompting strategy that enables LLMs to determine and justify entity equivalence, though its integration point is closer to RAG, as retrieval constrains the candidate pool before passing on to the LLM.

This trend of studies in Entity Linking show that LLMs enable a new paradigm where ontology engineering becomes a generative rather than purely extractive task, providing the computational substrate necessary for capturing the kind of relational structures and context that implicit meaning requires.

4.2.2.2 Automated ontology requirements engineering and testing

Given the centrality of requirements in OEMs, automated requirements generation is a long-standing task (Gangemi et al., 2022; Malheiros and Freitas, 2017). Recently, significant effort has recently proposed tools to specifically automate CQ generation according to different scenarios and needs, using LLMs. These systems differ in the type of input they expect, the prompting techniques adopted, and their intended use within the ontology engineering pipeline.

OntoChat (Zhang et al., 2024; Zhao et al., 2024) is a conversational-based framework that enables CQ generation starting from user stories. The tool supports iterative refinement of questions and is designed to facilitate participatory ontology engineering. NeOn-GPT (Fathallah et al., 2024) is a NeOn-inspired method (see §4.1.1.2) which generates CQs from requirements by combining structured prompt templates with ontology metadata. The system incorporates filtering and testing mechanisms to improve question quality and domain relevance. AgOCQs (Antia and Keet, 2023) is an automatic corpus-based CQ generation tool that uses fine-tuned LLMs to generate CQs from domain-specific corpora, such as scientific articles. It applies lexical,

syntactic, and semantic filters to ensure that generated CQs are well-formed, answerable, and reflect the intended semantics of the ontology. Pan et al. (2024) introduce a retrieval-augmented generation approach to CQ generation. The approach first retrieves relevant ontology content or background knowledge from scientific papers, which is then passed to an LLM to generate context-aware CQs. The system also provides semantic similarity scoring to assess the match between generated and reference questions. Rebboud et al. (2024b) investigate the use of LLMs to generate CQs given an existing ontology, comparing six different models and multiple prompt settings to evaluate how well LLMs can support knowledge engineers in this task, and assess CQ quality by measuring similarity to ground-truth questions using Sentence-BERT embeddings. Alharbi et al. (2024a) introduce the RETROFIT-CQs approach to generate candidate CQs from ontology triples using GPT-3.5 and GPT-4. Di Nuzzo et al. (2024) propose a pipeline that combines LLMs and Knowledge Graphs for automated CQ generation from domain standards. Their method retrieves entity and community level summaries from KGs to prompt LLMs.

As shown in the comparative Table 4.2, all of these works mainly focus on the generation of requirements to be used for ontology design or enhancing the quality of already existing requirements. In the next section, we review ontology learning and population tasks.

4.2.2.3 Ontology learning and population

Alongside requirements, ontologies can be generated automatically. This task of automating ontology generation, e.g., from natural language text, is commonly referred to as *ontology learning*. Initially, the field used classic NLP techniques such as POS-tagging in combination with lexical patterns or frames, to detect occurrences of ontological constructs in text (Cimiano, 2006). However, with methods such as those presented in Petrucci, Rospocher, and Ghidini (2018), deep learning has also made significant contributions to this field. Additionally, early language models have been applied for various subtasks, such as extending ontologies and mapping concepts to top-level ontologies (Lopes et al., 2023).

Many solutions for generating ontologies from text have focused on the Ontology Learning and Population (OL&P) task, as shown for example in the work of Al-Aswadi, Chan, and Gan (2020) and Cimiano (2006). Examples of such methods include Text2Onto (Cimiano, 2006), OwlExporter (Witte, Khamis, and Rilling, 2010), and the approach proposed by Tanev and Magnini (2006). Most of these methods employ machine learning techniques and are therefore data-hungry and subject to long training phases. Other approaches to OL&P use lexico-syntactic patterns (Hearst, 1992), or hybrid lexical-logical methods (Völker, 2009). Additionally, there are methods for transforming CQs into SPARQL queries, a task that complements the creation of OWL ontologies from text (Wisniewski, 2018). In contrast with such approaches, FrODO (Gangemi et al., 2022) was developed as a tool that does not depend on any training data to automatically construct a domain ontology from CQs. However, it did not rely on any language model and datatype properties, inverse functional datatype properties and disjoint axioms were not covered by the tool.

In this context, LLMs offer a natural solution to these outlined challenges. In the following chapter, we reframe what was traditionally called ‘ontology learning’ as ‘ontology generation’, reflecting the shift toward generative methodologies.

4.2.2.4 Ontology generation with LLMs

The LLMs4OL approach by Babaei Giglou, D'Souza, and Auer (2023) introduced a comprehensive framework that uses LLMs for automated ontology generation. This paradigm investigates whether LLMs can effectively apply their language pattern capturing capabilities to the task, which involves automatically extracting and structuring knowledge from natural language text. The authors conducted comprehensive evaluations using zero-shot prompting methods across nine different LLM model families for three main ontology learning tasks: term typing, taxonomy discovery, and extraction of non-taxonomic relations. Their results demonstrated that foundational LLMs are not sufficiently suitable for ontology construction that requires high degrees of reasoning skills and domain expertise, though when effectively fine-tuned they can serve as suitable assistants for alleviating the knowledge acquisition bottleneck in ontology construction. However, the approach primarily focused on extracting relations among ontology classes and instances rather than complete ontology generation.

Through an extensive review of inference making capabilities in Large Language Models, Saba (2023) proposed that applying a bottom-up reverse engineering strategy can enable the discovery of the underlying ontological structures present in natural language text.

Other preliminary work (Mateiu and Groza, 2023) has used fine-tuned GPT models to translate restricted natural language sentences into DL axioms. However, such specific statements do not represent realistic ontology requirements or scenarios.

Furthermore, tools have appeared to support the practical integration of OWL with LLMs, e.g. in He et al. (2024b), but such tools do not provide any specific guidance for the ontology generation task. The NeOn-GPT pipeline investigated in Fathallah et al. (2024) applied the principles of the NeOn methodology to a generative AI setting. However, that work only evaluates one single complex prompting technique, and evaluates the pipeline only on one single ontology generation example.

In LLMs4Life (Fathallah, Staab, and Algergawy, 2024), the NeOnGPT approach is extended to the life sciences domain. However, implementation details remain unclear, especially for what concerns the evaluation criteria. Instead, what is needed goes beyond new isolated proposals to explore, evaluate, and contrast different prompting techniques and pipelines to identify strengths and weaknesses, on more realistic and previously unseen tasks. The extent to which manual intervention is required and evaluation criteria are especially important in this line of work. In Reboud et al. (2024a), the ontology conceptualization task is proposed to be evaluated according to ontology evaluation criteria such as accuracy, completeness, and conciseness of the generated ontology as well as logical consistency. Unlike previous attempts which concerned single-prompt pipelines on single examples, Ontogenia explicitly studies context engineering techniques pattern reuse factors across conditions, reports costs and settings, and formalizes repeatable CQ-based tests (Section 4.3). In particular, in Saeedizade and Blomqvist (2024) and Lippolis et al. (2024), different methods and ontology evaluation approaches for ontology generation from requirements have been tested, obtaining results which are at least comparable to the quality of human novice modellers (for more details on these two studies, see Section 4.2.2.4).

Despite LLMs' well-known drawbacks, such as hallucinations, studies show (Ali et al., 2023; Hanna and Levic, 2023) they have the potential to outperform existing systems, with only a handful of examples and a well-designed prompt. However,

none of the proposed methods have so far treated the specific task of generating both single and combined ontology modules from real-world ontology requirements.

Context engineering techniques for ontology generation. Context engineering stands out for its simplicity in implementation and adaptability with respect to fine-tuning, offering an approach to enhancing knowledge engineering processes¹¹.

In Saeedizade and Blomqvist (2024) and Lippolis et al. (2024), the authors have explored and evaluated various context engineering techniques and identified ones that seem suitable for different aspects of ontology generation. In particular, research has explored subtask-decomposed prompting, i.e., in this case the promising CQbyCQ method from Saeedizade and Blomqvist (2024), as well as Metacognitive Prompting (MP), based on Chain Of Thought (CoT) (Wei et al., 2022) adopted in Lippolis et al. (2024).

In Saeedizade and Blomqvist (2024), the authors conducted the first systematic study of LLMs' ability to assist ontology engineers, experimenting with several state-of-the-art models including GPT-4, GPT-3.5, and open-source alternatives. Their methodology incorporated different popular context engineering techniques and introduced the CQbyCQ method, a sub-task decomposition technique where the LLM models only one competency question at a time, with outputs subsequently merged with previously generated ontology components. Their findings revealed that GPT-4 was the only model capable of providing ontology suggestions of sufficient quality, with CQbyCQ achieving performance comparable to novice ontology engineering students. This work established important benchmarks for evaluating LLM-assisted ontology development and highlighted both the potential and current limitations of context engineering approaches in specialized domains.

In the case of Lippolis et al. (2024), CoT, known for the specific ability to successfully improve the performance of different tasks and granting decrease in hallucinations (Chu et al., 2023), is specialized into MP, whose aim is to encourage self-evaluation through the introduction of a series of steps.

Starting from the studies of Brown et al. (2020), Wei et al. (2022), Wang et al. (2023c), and Wang et al. (2023a), where the response abilities of LLMs was stated to significantly improve when using CoT. MP has been shown to consistently outperform existing prompting methods in both general and domain-specific NLU tasks (Wang and Zhao, 2023). Unlike CoT, MP incorporates a systematic series of structured, self-aware evaluations to assess the reasons behind outputs, not only the way it has to be produced. Following Wang and Zhao (2023), the MP framework consists of five main stages: (1) understanding the input text, (2) making a preliminary judgment, (3) critically evaluating this preliminary analysis, (4) reaching a final decision accompanied by an explanation of the reasoning, and (5) evaluating the confidence level in the entire process. This approach mirrors human metacognitive processes, often defined as "thinking about thinking", and has demonstrated superior performance across ten NLU datasets (Wang and Zhao, 2023).

¹¹Hereon, we employ the term "context engineering" rather than "prompt engineering" to reflect the systematic, multi-component approach required for ontology engineering applications. While prompt engineering focuses on crafting individual textual instructions for LLMs, context engineering concerns the systematic optimization of information payloads (Mei et al., 2025; Karpathy, 2025) including domain knowledge, ontology design patterns, competency questions, and structured data integration. This distinction is particularly relevant for ontology generation, which requires comprehensive management of domain knowledge, memory systems, and multi-step reasoning processes rather than simple query optimization.

MP has been tested and applied to different domains. In ontology engineering, Lippolis et al. (2024) introduced Ontogenia, which combines MP with ODPs for ontology generation showing promising results when compared to other prompting techniques. We will present Ontogenia more in detail in Section 4.3. Similarly, in mathematical problem solving, MP has been used to extract metacognitive knowledge in the form of skill annotations, which can then improve reasoning by providing pertinent in-context examples (Didolkar et al., 2024). Furthermore, MetaRAG extends this approach by combining retrieval-augmented generation with metacognition, enabling systems to identify and fix inadequacies in initial cognitive responses (Zhou et al., 2024b).

4.2.3 Neuro-Symbolic methods

There is growing agreement that combining symbolic reasoning with neural learning is necessary to cover the breadth of intelligent behavior (Lake et al., 2017; Marcus, 2020; Marra et al., 2024; Yu et al., 2023). Symbolic systems excel at explicit rule-based reasoning, while neural networks are based on statistical learning from raw data. The complementary strengths of these paradigms have led many scholars, beginning with Valiant (2003)'s proposal and extending through the more recent arguments of Marcus and Davis (2019) and Kautz (2022), to call for neuro-symbolic hybrids. H. Kautz offers a six-fold taxonomy mapping the design space for neuro-symbolic hybrids.

The first pattern simply wraps neural networks in symbolic input and output layers, where symbols are vectorized, processed neurally, then re-translated back to symbols. Standard NLP models like Word2Vec exemplify this approach. The second pattern retains symbolic systems in control while delegating specialized perceptual tasks to neural subroutines, as seen in AlphaGo's combination of symbolic Monte-Carlo tree search with neural board evaluation.

A third configuration treats neural and symbolic paradigms as equal partners without hierarchy. The Neuro-Symbolic Concept Learner combines neural vision modules with symbolic reasoners, each contributing unique capabilities. The fourth possibility places neural systems center-stage but trains them on symbolically structured data.

The fifth pattern compiles symbolic knowledge directly into network topology or weights. Logic Tensor Networks translate logical formulae into tensors, making reasoning emergent from network structure through ordinary gradient descent. Finally, the sixth pattern reverses the second's hierarchy: neural systems remain in control but invoke symbolic modules for explicit reasoning when needed. Kautz argues this dual-process architecture may prove most fruitful, combining neural adaptability with symbolic transparency and reliability.

Together, these six patterns chart the main routes by which contemporary researchers are attempting to fuse learning and reasoning. Each route exploits a different balance of power between neural and symbolic components, and each entails its own advantages and trade-offs. The prevailing view is clear: the future of AI almost certainly belongs to technologies that can both acquire knowledge from experience and manipulate that knowledge with the precision of a shared consensus on a logical representation.

4.2.3.1 Logic Augmented Generation

Logic Augmented Generation, as presented by Gangemi and Nuzzolese (2025), represents a Type 2-3 neurosymbolic system (Kautz, 2022) (§4.2.3) that dynamically generates domain-specific knowledge graphs and ontology-based heuristics to guide LLM reasoning, rather than relying on static knowledge bases or simple knowledge retrieval like other hybrid approaches. Unlike purely symbolic approaches that lack flexibility in handling novel concepts, or purely neural approaches that lack explainability in their reasoning processes, LAG bridges this gap by using structured knowledge representations to guide neural generation while maintaining justifiability. It addresses the limitations of existing approaches that typically focus on single tasks or modalities. Furthermore, by generating explicit knowledge graphs, the approach provides interpretable capabilities that current LLM-based methods lack.

LAG conceptualises LLMs as potential Reactive Continuous Knowledge Graphs (RCKGs), which can dynamically adapt to diverse inputs by extending and contextualising Semantic Knowledge Graphs (SKGs) that work as base models. LAG exploits the duality between implicit knowledge generated during training and explicit knowledge, externalised in the form of structured knowledge graphs through generation and prompting. RCKG extraction involves a three-step process, i.e. (i) mapping multimodal signal to natural language, (ii) converting natural language into an SKG, and (iii) extending the SKG with tacit knowledge based on multiple heuristics (the extended knowledge graph as XKG).

Step (ii) is carried out with AMR2FRED, which converts Abstract Meaning Representation (AMR) parses into OWL/RDF graphs using FRED patterns, yielding SKG-ready triples that ground subsequent heuristics.

In this frame of reference, SKGs ensure logical consistency, enforce factual boundaries, and foster interoperability, while LLMs process unstructured data and provide contextual insights on implicit knowledge on demand.

In line with recent evidence that performance gains hinge as much on prompt adaptation as on model capacity (Jahani et al., 2024), LAG treats ontologies as stable but dynamic heuristics. Ontology-specified roles and constraints are compiled into parameterized prompt schemas with user-visible slots and a per-model rendering layer, so prompts evolve with context and model updates while remaining traceable to the SKG. This approach balances standardization (shared terminology, logical checks) with adaptability (task- and model-specific phrasing), mitigating the brittleness of static prompt libraries.

A study by De Giorgis, Gangemi, and Russo (2025) indeed shows the usefulness and accessibility of the approach via a user-based interface¹². Such an interface currently operationalises 11 domain-independent heuristics: Presuppositions, Conversational Implicatures, Factual Impact, Image Schemas, Metonymic Coercion, Moral-value-driven Coercion, Symbolic Coercion, Event Sequences, Causal Relations, Implied Future Events and Implied Non-events, each encoded as a reusable prompt-template that takes the base AMR2FRED graph as context and returns only the newly inferred triples, all anchored to pre-existing nodes. For every heuristic the system stores (i) the prompt, and (ii) the expected formal signature of the generated triples. In this sense the interface does not merely visualise the enriched graph but implements a lightweight continuous-integration loop for ontology modules: each time the user accepts or edits a heuristic result, the corresponding test is executed again over the asserted and the inferred model, and any discrepancy is logged.

¹²<https://arco.istc.cnr.it/itaf/>.

From a three-tier evaluation on a publicly available corpus of image-derived scenarios, the work confirms logical soundness and high results on surveys about the heuristics with respect to a vanilla Graph-RAG pipeline.

LAG shows multiple advantages over existing hybrid pipelines. First of all, because each heuristic is applied only to the graph fragment relevant to its anchoring term(s), the system avoids the combinatorial explosion typical of unconstrained LLM-generation. Furthermore, every inferred triple keeps an explicit trail, and while this cannot be considered in itself explainable, it can be considered interpretable, which is unattainable with retrieval-augmented generation. Finally, the framework is multimodal, accommodating speech transcript, video captions, or sensor logs, provided the upstream AMR parser can handle the modality.

4.2.4 The missing link: systematic LLM integration in ontology engineering

This chapter has shown that research on AI-assisted extraction and ontology construction has progressed from surface-level triple mining to full-fledged ontology generation and neurosymbolic enrichment. Neural baselines (§4.2.1) deliver strong span detection and relation labelling, but remain limited to flat, task-specific schemas. They neither enforce ontological constraints nor provide reusable, modular knowledge structures. LLM-centric pipelines (§4.2.2) widen the scope towards the automation of many knowledge engineering tasks, such as entity linking, CQ design, and ontology generation. Yet, issues of hallucination control, logical soundness, and systematic evaluation persist (Rebboud et al., 2024a; Saeedizade and Blomqvist, 2024). Neurosymbolic architectures such as LAG (Gangemi and Nuzzolese, 2025; De Giorgis, Gangemi, and Russo, 2025) illustrate how explicit KG layers in addition to heuristic context engineering can guide LLMs, yielding interpretable, testable extensions of a base ontology. Their success also exposes open questions on heuristic coverage, incremental versioning, and multi-modal alignment.

Together these strands chart a clear trajectory: from pipeline extraction to interactive, constraint-aware generation. What is still missing is a unifying methodology that (i) leverages LLMs to automate eXtreme Design-related tasks, (ii) supports modular ontology extension, and (iii) comes with comprehensive, shared, measurable evaluation criteria.

The gaps identified in existing approaches, and in particular, the need for systematic LLM integration with established methodologies like XD, and comprehensive evaluation frameworks, directly motivate our development of Ontogenia. This methodology represents a bridge between the pattern-based, human-centric traditions of ontology engineering and the generative capabilities of modern language models.

4.3 Ontogenia: empirical investigations into LLM-based ontology generation

Building on the limitations identified in early LLM-based approaches to ontology engineering, Ontogenia was developed as a method that integrates metacognitive prompting with ontology design patterns for one of the most important steps in ontology design methodologies: the formalisation of an ontology (or module thereof) based on a specific set of requirements.

Following the MP technique described in Section 4.2.2.4, the Ontogenia methodology for ontology generation is composed of five phases, which have been defined as an iterative and incremental process. Building on XD’s pattern-based approach (Section 4.1.1.3), Ontogenia adapts the iterative workflow by integrating metacognitive prompting with ODP selection. The final phase involves empirical testing once evaluation methods have been established. In this section, we propose two studies, a preliminary one and a comparative one, showing Ontogenia’s performance.

4.3.1 The Ontogenia methodology: a preliminary study

Ontogenia’s capabilities have been first explored in a preliminary study that applied the method to a specific narrow use case from an existing dataset, with the methodology detailed as follows.

4.3.1.1 Methodology

Experimental design overview. We designed a controlled experiment to evaluate whether Ontogenia’s approach produces higher-quality ontologies than baseline methods. Our experimental pipeline consists of four key components: (1) dataset selection and preparation, (2) prompt engineering with metacognitive stages, (3) systematic ontology generation across different configurations, and (4) multi-dimensional evaluation.

Dataset and ODPs selection. The dataset under consideration for the experiment is derived from a recognized gold-standard used for testing and benchmarking in the field of competency questions research and ontology querying, detailed in Potoniec et al. (2020). We selected the African Wildlife Ontology as our test case because it provided sufficient domain complexity to test pattern integration, included well-defined competency questions for evaluation, and allowed comparison with existing benchmarks.

The definition of the Ontology Design Patterns to reuse starts from the list of Content Ontology Design Patterns listed in the ODP website¹³. From these, nine have been manually selected and included in a dataset to be dynamically inputted to the prompt: AgentRole¹⁴, AquaticResources¹⁵, Classification¹⁶, Climatic Zone¹⁷, Collection Entity¹⁸, PartOf¹⁹, Linnaean Taxonomy²⁰, SpeciesEat²¹. These have been chosen because of the domain and CQs pertaining to animal behavior, habitats, and ecological relationships.

In particular, the *AgentRole* ODP is instrumental in understanding the various roles agents play in their ecosystems. *Aquatic resources* provides insight into the aquatic habitats and the species that inhabit them. *Classification* is crucial for categorizing animals based on established biological taxonomy, offering a structured approach to animal study. *Climatic zone* aids in correlating animal species with their

¹³<http://ontologydesignpatterns.org/wiki/Submissions:ContentODPs> and <https://github.com/odpa/patterns-repository>.

¹⁴<https://github.com/odpa/patterns-repository/tree/master/AgentRole>.

¹⁵<https://github.com/odpa/patterns-repository/tree/master/AquaticResources>.

¹⁶<https://github.com/odpa/patterns-repository/tree/master/Classification>.

¹⁷<https://github.com/odpa/patterns-repository/tree/master/ClimaticZone>.

¹⁸<https://github.com/odpa/patterns-repository/tree/master/CollectionEntity>.

¹⁹<https://github.com/odpa/patterns-repository/tree/master/PartOf>.

²⁰<https://github.com/odpa/patterns-repository/tree/master/LinnaeanTaxonomy>.

²¹<https://github.com/odpa/patterns-repository/tree/master/SpeciesEat>.

preferred climatic conditions. The *Collection Entity* ODP allows for the examination of animal groupings and their interactions. *Part Of* sheds light on the anatomical and functional aspects of animals and plants in relation to their environment. *Linnaean Taxonomy* offers a classical approach to understanding the hierarchical classification of animals. Lastly, *Species Eat* is essential for studying the food chains and dietary habits within animal communities. Collectively, these ODPs can provide a comprehensive foundation for answering the targeted competency questions, enabling a thorough exploration of animal-related topics in a systematic and informed manner.

Procedure definition. Because it is fundamental to guide the LLM towards a specific definition of ontology features, a nine-step procedure has been defined to include the core ontology elements in the generated ontology.

Figure 4.3 is an extension of the illustration proposed by Wang and Zhao (2023) about the mapping between the phases of human metacognition, the operational phases of MP, and the phases of the human metacognition method, tailored for ontology design. In the figure, we showcase the steps added to the MP method to enable it within the context of ontology design through intermediate phases.

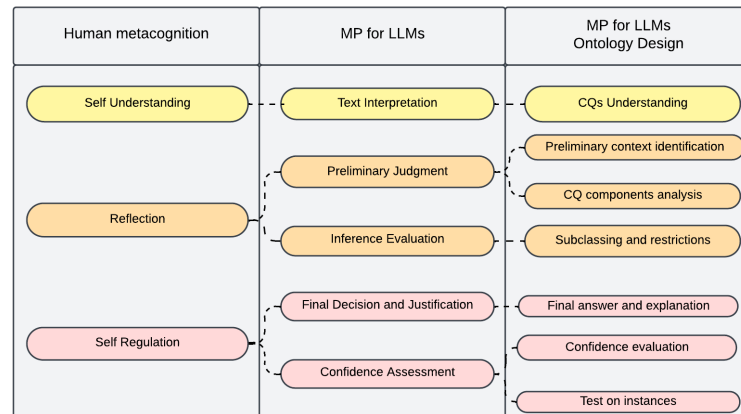


FIGURE 4.3: Extension of the figure in Wang and Zhao (2023) and mapping between human metacognitive processes, MP phases in LLMs, and metacognitive phases for ontology design using LLMs.

The resulting procedure is mapped to the five steps as shown in Table A.1 and further details on the methodology are presented in A.1.

Prompt design. The prompt design is meant to incorporate information about the procedure, eventual previous output, competency questions and patterns to be also

added dynamically to the prompt on the basis of specific needs. The patterns in this study are all combined into one single file. A specification to not repeat itself and not send comments was added in order to refine the output. The final prompt template can be seen below:

Read the following instructions: $\{procedure\}$. Basing on the procedure, and following the previous output: $\{previous\_output\}$, design an ontology that comprehensively answers the following competency questions: $\{CQs\}$, using the following ontology design patterns: $\{combined\_patterns\}$. Do not repeat classes, object properties, data properties, restrictions, etc. if they have been addressed in the previous output. When you're done send me only the whole ontology you've designed in OWL format, without any comment outside the OWL.

4.3.1.2 Evaluation setup

The definition of the evaluation is twofold. On the one hand, it involves an expert ontology engineer that analyzes the produced ontologies in terms of essential requirements such as required classes and object properties and usage of restrictions. On the other hand, it involves the Ontometrics service²² and the OOPS! Ontology Pitfall Scanner (Poveda-Villalón, Gómez-Pérez, and Suárez-Figueroa, 2014).

We use *GPT-4 Turbo* via the API (model: *gpt-4-1106-preview*²³) as our backbone model, with greedy decoding. Because GPT has a token limit in output, we came up with a division of competency questions to be given one group at a time, and each time the previous output is provided in order not to have a repeating of classes and properties.

We conducted four trials to evaluate the effect of different inputs on ontology generation. Trial 1 used competency questions with a generic prompt, Trial 2 added ontology design patterns, Trial 3 involved only competency questions and MP, and Trial 4 combined competency questions with MP and patterns. We thus test four configurations in a 2×2 factorial design. Factor 1: MP vs. Standard Prompting. Factor 2: ODPs vs. No Patterns. This yields four test conditions: MP+Patterns, MP only, Patterns only, Neither MP nor Patterns. We tested both the original and thematically grouped questions by GPT. The total computation cost was \$2.10. Experiment details are documented in a Github repository²⁴.

4.3.1.3 Results

Table 4.3 shows the metrics obtained using the OntoMetrics service for each case considered in the experiment. It can be seen that in our experiment the adoption of MP favors a richer formalisation. This is also evident from Table 4.4 that shows usage of a set of different types of axiom types across the test cases.

The tables also show the pitfalls found by OOPS!. For what concerns basic metrics, Ontometrics shows a larger number of axioms when a pattern is used, along with a higher number of classes and object properties. Data properties are instead a weaker point, despite their use having been specified in the procedure.

According to the qualitative analysis by the ontology expert, while the LLM successfully identifies necessary classes and relevant subclasses, the generated ontologies exhibit numerous intrinsic and domain-related issues. Particularly problematic

²²<https://ontometrics.informatik.uni-rostock.de/ontologymetrics/>.

²³<https://openai.com/blog/new-models-and-developer-products-announced-at-devday>.

²⁴https://github.com/dersuchendee/Ontogenia/blob/main/ontology_design.log.

is the pairing of classes and properties. Properties like `eats` often possess overly specific domains and ranges, leading to the creation of unrelated properties such as `eatsPlant`, `eatsAnimal`, and `eatsPlantPart`. While simple restrictions in class definitions are generally correct, the classification of animals by diet consistently falls short.

TABLE 4.3: How Ontometrics base metrics vary between various test cases and the reference ontology. For each case, it is indicated whether the patterns or the MP have been used.

Ontometrics	Case1 No pattern No MP	Case2 Pattern No MP	Case3 No pattern MP	Case4 Pattern MP	Reference Ontology
Axioms count	49	119	64	118	108
Logical axioms count	26	74	36	76	56
Class count	14	17	14	21	31
Object property count	8	11	8	14	5
Data property count	0	2	3	2	0
Properties count	8	13	11	16	5
Individual count	1	19	0	11	0
DL expressivity	ALCROI	AL(D)	ALC(D)	ALCI(D)	SRI

TABLE 4.4: Use of types of OWL axioms in various test cases. For each case it is indicated whether the patterns or MP were used.

Axiom type	Case1 No pat- tern No MP	Case2 Pattern No MP	Case3 No pat- tern MP	Case4 Pattern MP
owl:Ontology	Yes	Yes	Yes	Yes
owl:Class	Yes	Yes	Yes	Yes
owl:ObjectProperty	Yes	Yes	Yes	Yes
owl:DatatypeProperty	No	No	Yes	Yes
rdfs:domain	Yes	Yes	Yes	Yes
rdfs:range	Yes	Yes	Yes	Yes
rdfs:subClassOf	Yes	Yes	Yes	Yes
rdfs:subPropertyOf	No	No	No	No
owl:disjointWith	No	No	Yes	Yes
owl:equivalentClass	Yes	No	Yes	No
owl:Restriction	Yes	No	Yes	Yes
owl:imports	No	Yes	No	Yes

4.3.1.4 Discussion

The results of this preliminary study open many avenues for discussion. The first one is derived from the evaluation methodology adopted. While quantitative metrics are useful, the results also show that an expert’s insight is needed to explore error patterns and boundary cases, and assess whether improvements are meaningful in practice. In fact, Ontometrics shows the adoption of MP favors a richer formalisation, along with a set of different types of axiom types across the test cases. At the same time, there are important limitations common to all the cases, such as the absence of property hierarchical relationships.

The OOPS!-based analysis reveals the nature of these limitations. Some issues are common to all the cases, as the lack of annotations (P08) and inverse relationships (P13)²⁵. All the cases except Case 3 contain at least a property that is defined

²⁵The lack of license information (P41) is also common to all the cases, but this is not an information to be expected from the LLM.

with more than one domain or range (P19). This happens when the LLM generates multiple times the definition of an object property, with somewhat different domain/range values. It seems that the “intent” of the LLM would be to define the property over the union of the referenced classes, mirroring a common beginner’s error in RDFS modelling. In the cases using patterns (2 and 4) an “untyped class” (P34) and “different naming conventions” (P22) are found. These are both due to the erroneous of an object property (hasPart) imported from a pattern but used as it was a class. Furthermore, in Case 1, no disjoint axioms are used (P10) and there are a couple of properties missing explicit domain/range declaration (P11)²⁶. The ontology obtained in Case 3 is the only one featuring an ontology element, the Plant class, unconnected from the rest of the ontology. Nevertheless, Case 3 is the one having less pitfalls because patterns may have introduced noise.

Apart from finding structural errors, the expert analysis shows the existence of numerous superfluous classes and properties that could, instead, be merged into a single one. This tendency toward overly specific properties may partially explain the consistently higher property counts observed across all LLM-generated ontologies compared to what might be expected for comprehensive modeling of the domain. This shortfall in the example ontology chosen to test is, however, also partly due to the ambiguity of terms like “carnivore” in biological contexts, where strict logical constraints are challenging to establish.

It is also necessary to discuss the comparison with the reference ontology, which has deliberately not been carried out with semantic similarity metrics. While the reference ontology provides a baseline, direct quantitative comparison may be misleading given the fundamental differences in modeling approaches between human experts and LLM-generated structures, making multiple assessment methods a more reliable evaluation method.

These findings highlight the necessity for further research into the collaboration between ontology design teams and LLMs, but also underscore the need for more systematic comparative evaluation of context engineering techniques in collaboration with Linköping University. Building on these preliminary insights, we designed a controlled comparative study to refine and re-assess Ontogenia and another established approach, CQbyCQ (Saeedizade and Blomqvist, 2024), using standardized evaluation frameworks.

4.3.2 A comparative analysis of context engineering techniques

In this second study, Ontogenia was refined and experimented as a technique to be compared to the refined CQbyCQ technique by Saeedizade and Blomqvist (2024) (§ 4.2.2.4), called MemorylessCQbyCQ. In this study, a new dataset and new evaluation measures are defined, proposing a standardized framework for the implementation and evaluation of ontology generation.

4.3.2.1 Preliminaries

Superfluous Element. Given an ontology O , a set of competency questions and the set of corresponding validation SPARQL queries, if a named class, object property or data property is not used in any SPARQL query, it is considered a superfluous element of O .

²⁶Domain and range of those two properties are actually in part inferable because, errors aside, they are meant to be defined in relationship to other properties.

Minimal Ontology Module. Given an ontology O and a competency question CQ_i , O_i is a minimal ontology module for CQ_i if there are no superfluous elements of O with respect to CQ_i .

Minor Issue. For an ontology O and a competency question CQ_i , if O includes all necessary elements except for only one object property or only one data property, and adding this single element to O would make CQ_i modelled, this is considered a minor issue in modelling CQ_i . This definition is based on the intuition that one such missing link is rather easy to add, even by a novice modeller, given that all classes are present.

4.3.2.2 Methodology

The initial phase of the work involved manually creating a benchmark dataset using available ontologies accompanied by their corresponding requirements for CQ verification (Blomqvist, Seil Sepour, and Presutti, 2012) (§ 4.1.1.3). Two methods for ontology generation were then introduced, namely Independent Ontology Generation and Incremental Ontology Generation, and adapted existing prompting techniques in our experiments to guide LLMs in generating ontologies. In this first part of the study, we primarily employ *GPT-4*, identified as the best-performing LLM in earlier comparative analyses (Frey et al., 2024a; Saeedizade and Blomqvist, 2024). Additionally, we independently compare *GPT-4* with *OpenAI o1-preview* and *Llama-3.1-405B-instruct-16b* using the dataset proposed in Saeedizade and Blomqvist (2024).

Dataset creation. In order to evaluate the generated ontologies using CQ verification (Blomqvist, Seil Sepour, and Presutti, 2012), a dataset of CQs and user stories was developed, along with their corresponding minimal ontology modules. The process involved selecting a set of ontologies for the study, extracting CQs and stories, and subsequently extracting modules from these chosen ontologies in order to have minimal modules for easier evaluation and possible future extensions of the work.

The dataset combines two main types of sources: extracted datasets and manually created datasets. The extracted datasets consist of CQs and stories derived from existing resources, where superfluous elements were removed to identify minimal modules. In contrast, the manually created datasets were created as a part of a controlled module engineering process. In this case, we include the dataset present in Saeedizade and Blomqvist (2024) called “SemanticWebCourse”, where master’s students with a computer science background but no prior ontology design experience were tasked with creating semantic web solutions. Each student group submitted an initial solution, revised it based on teacher feedback, and resubmitted a final version to pass the assignment, resulting in two distinct solutions per group. However, as previous work questioned the generalisability of the LLMs generating ontologies (Saeedizade and Blomqvist, 2024), in this work, we enriched existing data with more ontology sources, primarily ones used in real-world projects.

An ontology was included in the dataset according to the following criteria: if it included (i) a set of competency questions and (ii) a set of corresponding user stories. These criteria were motivated by the eXtreme Design methodology Presutti et al., 2009 (§ 4.1.1.3), which takes into consideration CQs and user stories as fundamental requirements of ontologies. As a result, we selected a total of ten ontologies, with 100 distinct CQs, and 29 different user stories from four real-life semantic web projects and three educational ones. More specifically, these ontologies have CQs assigned to specific minimal ontology modules with different user story distributions.

The main goal of developing the dataset was to support the generation of ontologies using LLMs. The CQs and user stories provide input for prompting techniques in the ontology generation process. The minimal ontology modules contained in the datasets provide ontologists with a gold standard for the expected output of the LLMs and are used to assist the ontologists' assessment, potentially also serving as useful resources for future work on fine-tuning LLMs.

To create the minimal ontology modules for each CQ the following steps were used: First, we manually checked to see if there were duplicate CQs and removed them. Then, from the ontologies, we removed superfluous elements so each module contained the minimum necessary classes and properties to effectively model the CQ concerning the ontology story. Finally, these modules were cross-checked and evaluated independently by two of the authors of this study.

Dataset composition. The dataset is composed of SIMPLE and COMPLEX CQs.

For instance, "Who built and/or renovated an organ?" is classified as SIMPLE as it involves two classes (Person, Organ) and one object property (built/renovated): $\exists x, y : \text{Person}(x) \wedge \text{Organ}(y) \wedge (\text{built}(x, y) \vee \text{renovated}(x, y))$. Conversely, "What was the disposition of the organ at a specific point in time?" is COMPLEX due to multiple classes and properties: $\exists x, y, z : \text{Organ}(x) \wedge \text{Parthood}(y) \wedge \text{TimeInterval}(z) \wedge \text{isWholeIncludedIn}(x, y) \wedge \text{hasTimeInterval}(y, z)$.

These have been manually divided into four categories as defined in Saeedizade and Blomqvist (2024): Data Property Modelling (10 CQs) and Object Property Modelling (25 CQs) are Simple CQs, while Reification (62 CQs) and Restrictions (3 CQs) are Complex CQs. The CQs, stories and corresponding ontologies are taken from Polifonia²⁷, Onto-DESIDE²⁸, WHOW (Lippolis, Lodi, and Nuzzolese, 2025), IKS²⁹ and the "SemanticWebCourse" Saeedizade and Blomqvist (2024) consisting of three ontology stories and 15 CQs each.

As an additional baseline for the experiments in this paper we used the same set of student solutions of the "SemanticWebCourse" as in Saeedizade and Blomqvist (2024). In Saeedizade and Blomqvist (2024), the baseline for evaluation was the first submission and the last submissions of student groups, and because it is the current state of the art in ontology generation, we use the same baseline in this paper for comparison. The intuition behind using this baseline is to compare the modelling abilities of LLMs to novice ontology engineers, with and without expert feedback. We did not compare to the original Ontogenia work Lippolis et al., 2024 (see Section 4.3.1) as, given its noted shortcomings, the technique had been revised and adapted for decomposed prompting.

Context engineering techniques. The context engineering techniques are based upon those described in Section 4.2.2.4. These previous techniques were therefore refined and evaluated on the benchmark dataset.

Memoryless CQbyCQ. Memoryless CQbyCQ processes one CQ at a time, using the ontology story to guide the LLM with a prompt in generating an ontology model for each CQ. It then merges the outputs into a single ontology. Unlike the CQbyCQ method, the LLM does not have access to previously generated ontologies and other CQs while modelling a specific CQ, which reduces the input context size of the LLM by around 60%. In fact, in Saeedizade and Blomqvist (2024) it was shown that long

²⁷<https://polifonia-project.eu/>

²⁸<https://ontodeside.eu>

²⁹<https://cordis.europa.eu/project/id/231527>

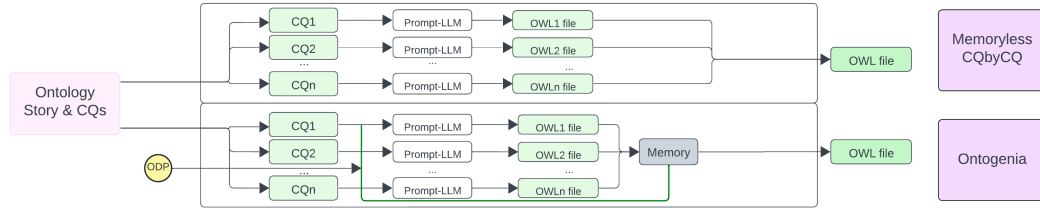


FIGURE 4.4: Illustration of Memoryless CQbyCQ (top part) and Ontogenia (bottom part).

context can result in distraction of the LLM. Hence, the intuition for removing the memory part is that slight overlaps between partial solutions will be easier to resolve and correct by even a novice ontology engineer, than completely irrelevant or inconsistent solutions, which could be the result of distracting the model. As a result, each CQ is independently modelled, and then all the resulting models, representing CQs, are merged at the end. This prompting technique guides an LLM, gives it an ontologist persona, and introduces Turtle syntax for defining classes, properties, reifications, and other ontology engineering features. It also includes a story section that outlines ontology requirements and a CQ, followed by common pitfalls in ontology development using LLMs, such as producing an empty output or engaging in conversation.

Ontogenia technique. Ontogenia was first defined in the preliminary study by Lippolis et al. (2024) (see Section 4.3) and has been further refined for this work. Like the CQbyCQ method, this technique guides an LLM by instructing it to be an ontology engineer, and defines basic guidelines for effective ontology formalization. The model processes one CQ at a time. When prompted to model a set of CQs, it models and merges the generated ontology at each step and provides it in the context. The technique has been refined to include a user story in the prompt. The set of ODPs remains as input, but only generic ones are chosen in this study to maintain domain agnosticism while preserving context limitations (AgentRole, Classification, Collection Entity, PartOf), and possibly previous output. The procedure is also the same as the original one with the addition of further instructions to handle better restrictions and reification.

Similarities and differences between Memoryless CQbyCQ and Ontogenia. Although the methods are similar in nature, and use the same testing setup to ensure comparability, there are some notable differences. Ontogenia explicitly requests labels, comments, inverse relationships and individuals to be provided. Furthermore, Ontogenia requires the injection of Ontology Design Patterns from the ODP repository (Blomqvist and Sandkuhl, 2005) and makes use of MP.

Unlike Ontogenia, Memoryless CQbyCQ includes common pitfalls to ensure they are avoided in the ontology output and attempts to reduce the context size.

Figure 4.4 shows an overview of these two prompting techniques.

Ontology Generation Methods Here, we present the two ontology generation methods used in our experiments, as shown in Figure 4.4: Independent and Incremental Ontology Generation.

Independent Ontology Generation

In this method, each CQ with its corresponding ontology story is fed into an LLM using a prompting technique to generate the corresponding ontology. This ensures that each CQ is treated as a standalone unit, allowing for a focused assessment of the ontology generation process on a per-question basis. By isolating each CQ,

it becomes possible to analyze the performance of the prompting techniques without interference from interdependencies or complexities that may arise in multi-CQ contexts.

Incremental Ontology Generation

Here, similar to previous work Saeedizade and Blomqvist, 2024; Lippolis et al., 2024, all CQs of a story are fed to an LLM at once, and a single OWL ontology is expected as output. Unlike the independent approach, where each CQ is modelled in isolation, this technique integrates solutions to multiple CQs to produce a cohesive minimal ontology module representing the complete story and its requirements. This method can be carried out either by merging the outputs for each CQ into a single ontology (Memoryless CQbyCQ) or by incorporating the output of each CQ incrementally into the prompt and joining it at each iteration (Ontogenia).

4.3.2.3 Evaluation setup

Experimental setup. The experiment consisted in running the Independent Ontology Generation method with *GPT-4-1106* with both prompting techniques across the entire dataset. For the Incremental Ontology Generation method, we used both MemorylessCQbyCQ and Ontogenia with three different LLMs (*GPT-4 1106*, *o1-preview*, and *Llama-3.1-405-Instruct-16b*) solely on the dataset introduced in Saeedizade and Blomqvist (2024) of three semantic web course stories and 45 CQs. As a baseline for comparison we used their results (Saeedizade and Blomqvist, 2024), i.e. their best solutions produced using *GPT-4* with CQbyCQ, and the recorded scores of the students' submissions.

In order to provide a comprehensive evaluation, each of the solutions produced through these methods for the Independent and Incremental Ontology Generation experiments has been evaluated by the proportion of modelled CQs, with two ontology engineers cross-checking each other's judgments. In the event of any conflicting assessment, they engaged in discussions to resolve the disagreement. In the case of the Incremental Ontology Evaluation experiment, also standard ontology metrics through the OOPS! tool (Poveda-Villalón, Gómez-Pérez, and Suárez-Figueroa, 2014), as well as a structural analysis to evaluate the rate of superfluous elements and a qualitative expert evaluation, were carried out. Due to lack of resources and time, we were not able to perform this evaluation on the complete result set. For the LLMs, we used the default hyperparameters of *o1-preview* and *Llama-3.1-405-Instruct-16b*; for *GPT-4*, temperature and penalty are set to zero as recommended in Saeedizade and Blomqvist (2024). Figure 4.5 illustrates the three evaluation steps.

Standard ontology metrics. The first step in evaluating the generated ontologies is applying standard evaluation metrics. We chose the OOPS! tool (Poveda-Villalón, Gómez-Pérez, and Suárez-Figueroa, 2014), which was also used in the Ontogenia preliminary study (§ 4.3.1), due to its coverage of common modelling mistakes and best practices. In the results, we report only the number of critical issues related to each method, as they are the only ones crucial to correct, while other pitfalls are not certain to actually represent modelling flaws in the domain (Poveda-Villalón, Gómez-Pérez, and Suárez-Figueroa, 2014).

Proportion of Modelled CQs. In this step, for each CQ, we evaluate whether it is modelled in accordance with the definition provided in Section 4.3.2.1. Additionally, we adopt a relaxed interpretation of this criterion by disregarding minor issues, as outlined in Section 4.3.2.1. Finally, the proportion of CQs deemed to be modelled,

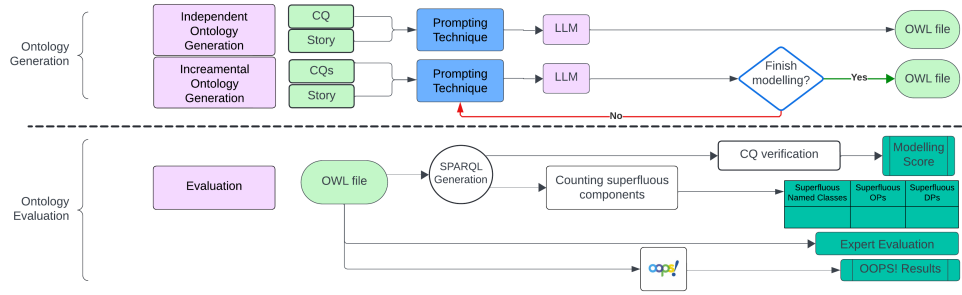


FIGURE 4.5: Illustration of the two ontology generation settings (top) and the four evaluation steps for assessing the generated ontologies (bottom). The top setup generates an ontology concerning only one CQ, which is evaluated individually. The second setup generates an ontology covering multiple CQs associated with one story, which is then evaluated. At the bottom, the four ontology evaluation settings are shown: OOPS!, the proportion of modelled CQs, statistics of superfluous elements and expert evaluation.

both with and without accounting for minor issues, is used to compute an overall score representing the proportion of the total list of CQs being sufficiently covered.

As an example, in Figure 4.6, part A (top), a CQ such as “Who is the author of a book?” is shown to be correctly modelled, as all the elements required to write a SPARQL query are present in the model (dotted box). In part B, for the same CQ, an object property to connect Book and Author (or either connecting it to Person or a new data property) is missing. However, adding “authorOf” would resolve the issue. This is considered a minor error, as the ontology modeller can easily add this element.

Structural Analysis. We manually evaluated each OWL file using three novel criteria with respect to superfluous elements 4.3.2.1, i.e., superfluous named classes, object properties, and data properties. We report the rate of superfluous elements by dividing the number of superfluous elements by the total count of that element type, and compared them to the rates in Saeedizade and Blomqvist (2024).

For example, in modelling the CQ “Where did a certain festival take place?”, the ontology may have two classes (e.g., “Location” and “Festival”) and one object property (e.g., “take_place”). However, multiple mentions of each class and property may exist with different spellings or styles in the generated ontology. Determining the number of superfluous components involves a two-step manual process. Firstly, if two or more components were found to refer with the same meaning to the same entity or relationship, all but one were classified as superfluous. Subsequently, the remaining classes and properties were used to formulate SPARQL queries for each CQ in the narrative. Then, the classes and properties that were not mentioned in any SPARQL query were also considered to be superfluous. For example, in the scenario where $n + 1$ classes pertain to a single concept (e.g., “location” and “place” referencing one concept to model the location of a festival), n extra representations have been modelled, including superconcepts of a used class when not present in the SPARQL query.

In Figure 4.6, part A, for the given CQ, there is a class “Person” and properties such as “name” or “wrote” that are not used in the SPARQL query (top-right corner). Since they (outside of the dotted box) do not appear in the SPARQL, they are referred to as superfluous elements.

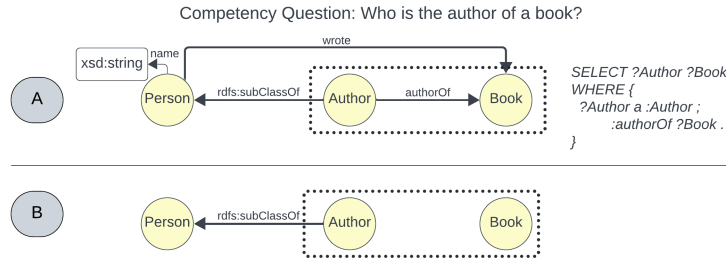


FIGURE 4.6: Analysis of an ontology for a CQ. Part A shows a correctly modelled CQ, ensuring all necessary elements for a SPARQL query are present, but contain superfluous elements. Part B shows a minor issue where a data property is missing.

Expert Qualitative Analysis. The expert qualitative analysis was conducted by two experienced knowledge engineers, not involved in the practical ontology generation task, who carefully and independently evaluated the outputs generated by the best performing open- and closed-source LLMs, *Llama-3.1-405-Instruct-16b*, and *OpenAI o1-preview*. The experts were instructed to assess the LLM-generated ontology files as if they were students’ course submissions and to provide feedback accordingly. Therefore, they focused on: (i) assessing the general quality of the output with dedicated comments on usability, completeness, and accuracy of the produced output, and mentioning any ontology errors found; and (ii) assessing the ontology with respect to the adequacy of the modelling solutions for each CQ³⁰, with a substantial agreement (Cohen’s kappa: 0.61).

Each metric evaluates a specific aspect of the LLM-generated ontology, as no single metric is sufficient to provide a comprehensive assessment. For instance, Llama produces many superfluous elements, hence it also has a reasonable likelihood of randomly modelling certain simple CQs correctly. By penalizing superfluous elements, we prevent LLMs from generating too many unnecessary classes and properties; however, this approach may also discourage the generation of taxonomies or more generalized modelling, which is a disadvantage. By combining all metrics, we can achieve an holistic evaluation of ontology generation.

4.3.2.4 Results

In this section, we start by examining the results of the OOPS! resource on the incrementally generated ontologies. Then, we compare the two prompting techniques introduced in this paper by going in-depth into the modelling results with respect to the assessment of the proposed ontology generation methods. More specifically, we measure the proportion of modelled CQs for all experiments and analyze the structure of the generated ontologies concerning superfluous elements. For the Incremental Ontology Generation method, we focus on the “SemanticWebCourse” dataset since it contains 15 CQs associated with one story while others have approximately 2 on average and choose the best performing open- and closed-source LLMs to compare across both previous work (Saeedizade and Blomqvist, 2024) and the results of

³⁰The “not adequate” assessment combines two judgments by the KE experts, i.e. a clear “no” where the CQ is not modelled, and a “maybe” category where the ontology simply does not allow for accurate assessment of the CQ, for instance, due to usability issues, naming, etc.

Independent CQ Generation. In this way, we can test the LLMs’ capability to generate an ontology incrementally. Finally, we present the results of the assessment by the two KE experts who analyzed the outputs.

Standard Ontology Metrics. After running OOPS! on the ontologies generated by the Incremental method for the “SemanticWebCourse” stories (Hospital, Music, and Theatre), the resulting pitfalls are shown in Table 4.5. Overall, there are a few critical pitfalls for specific combinations of LLMs and prompting techniques. The most common flaw is having multiple domains or ranges, which is in line with our other results discussed further in the following sections (see Section 4.3.2.4). In this case, the domain or range (or both) of a property (object or data property) is defined by more than one `rdfs:domain` or `rdfs:range` axiom. In OWL, multiple `rdfs:domain` or `rdfs:range` axioms are allowed, but are interpreted as a conjunction. Therefore, they are equivalent to the construct `owl:intersectionOf`, which could generate many unwanted inferences or even inconsistencies in the ontology. Some issues with inverse relations are also noted, as well as missing declarations of namespaces. Lastly, *OpenAI o1-preview* produced the fewest critical issues when using Memoryless CQbyCQ, whereas Llama-3.1-405-instruct-16b generated, with both methods, files with the most critical pitfalls.

TABLE 4.5: Pitfalls after OOPS! pitfall scanning for Incremental Ontology Generation. Results are merged by the three stories in the “SemanticWebCourse”. Llama* refers to *Llama-3.1-405-instruct-16b*. For the CQbyCQ method, only GPT-4 was used Saeedizade and Blomqvist, 2024.

		CQbyCQ	MemorylessCQbyCQ			Ontogenia		
		GPT-4	GPT-4	Llama*	o1	GPT-4	Llama*	o1
P05	Wrong inverse relationships	0	1	25	0	2	7	5
P06	Cycles in a class hierarchy	0	0	2	0	5	0	11
P19	Multiple domains or ranges	0	23	32	1	4	15	0
P29	Wrong transitive relationship	0	0	0	1	0	0	0
P37	Ontology not available	3	2	0	0	0	0	0
P39	Ambiguous namespace	3	2	0	1	1	0	0

Proportion of Modelled CQs.

Independent Ontology Generation Results: The evaluation of the Memoryless CQbyCQ and Ontogenia techniques on *GPT-4* shows that the proportion of correctly modelled CQs was 0.91 and 0.84, respectively (0.94 and 0.89 by ignoring minor issues), with significantly lower scores for complex CQs (0 and 0.66 respectively). These results show that by overlooking minor issues both prompting techniques are effective for modelling single CQs in isolation.

Incremental Ontology Generation Results: The results of the Incremental Ontology Generation are presented in Figure 4.7. Ontogenia with *OpenAI o1-preview* is the best-performing combination with respect to this evaluation criterion. Compared to the students’ submissions for the “SemanticWebCourse”, Ontogenia and Memoryless CQbyCQ with any LLM exceeded the proportion of CQs being modelled by students by a noticeable margin. Also, Memoryless CQbyCQ performed better than the CQbyCQ technique in Saeedizade and Blomqvist (Saeedizade and Blomqvist, 2024), possibly suggesting that using partially generated ontology models as data “in memory” for LLMs, which includes previously generated responses to CQ modelling, actually may diminish their performance. This result highlights the importance of reducing the input context size for LLMs, which is consistent with

the findings in Saeedizade and Blomqvist (2024). The proportion of correctly modelled CQs is analyzed across four categories: datatype property (DP), object property (OP), reification (Reif.), and restrictions (Rest.) for the Theatre, Music, and Hospital Story. We compared *GPT-4*, *Llama-3.1-405B-instruct-bf16* and *OpenAI o1-preview*. The evaluation shows that Ontogenia with *OpenAI o1-preview* achieves the highest number of correctly modelled CQs. Overall, both Ontogenia and Memoryless CQbyCQ surpass the students’ last submissions and the state of the art in ontology generation (Saeedizade and Blomqvist, 2024). Moreover, *Llama-3.1* in both prompting techniques had shortcomings in modelling reification CQs.

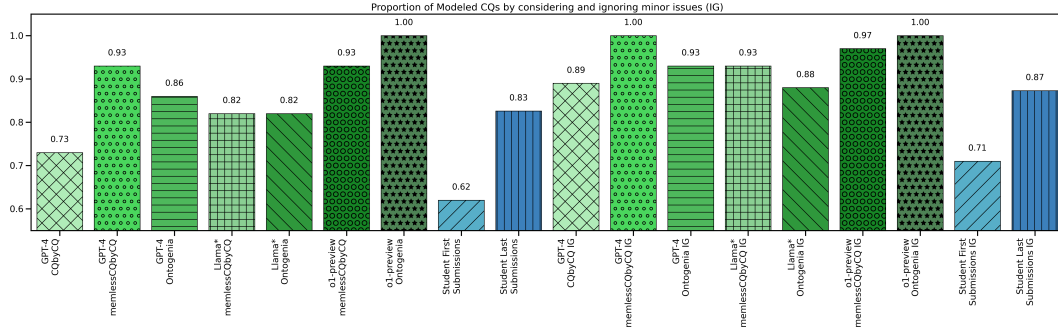


FIGURE 4.7: Scores for “SemanticWebCourse” from the outputs for the different prompting techniques compared with students’ submissions according to the proportion of the CQs that were accurately modelled. ‘IG’ indicates results when minor issues are ignored.

Llama* refers to Llama-3.1-405B-instruct-bf16.

Structural analysis. Table 4.6 presents our structural analysis regarding the superfluous elements. The results, compared to those in Section 4.3.2.4, show that the Memoryless CQbyCQ and Ontogenia, despite their performance measured in other criteria, yield a significant number of superfluous elements within the ontology. This underscores the importance of applying additional criteria for evaluating ontologies constructed using LLMs. CQbyCQ with *GPT-4* produces the fewest superfluous elements compared to our newly introduced prompting techniques, while the Memoryless CQbyCQ approach yields comparable results. Conversely, *Llama-3.1-405B-instruct-bf16* generates numerous superfluous elements across both techniques and all stories, with a rate close to 40%, indicating that this model tends to produce many superfluous classes and properties.

TABLE 4.6: Comparison of the three prompting techniques, using the Incremental Ontology Generation method, for generating superfluous elements across different ontology domains (Theatre, Music, and Hospital). Each cell represents the proportion of superfluous elements relative to the total number of elements of that type (%). CQbyCQ was only tested with GPT-4 due to its low CQ-coverage. Rates under 15% are bolded (best performance), and rates over 50% are in red (worst performance).

Prompting Technique	Superflu. Classes			Superflu. Obj. Properties			Superflu. Data Properties		
	Theatre	Music	Hospital	Theatre	Music	Hospital	Theatre	Music	Hospital
CQbyCQ (GPT-4)	0	14.3	8.6	0	4.2	6.6	12.5	28.5	18
Ontogenia (GPT-4)	0	19.2	13.5	4	38.9	29.6	45.5	55.5	22.2
Ontogenia (Llama)	47.1	32.1	37.5	40.6	45.9	38.3	16.7	0	100
Ontogenia (o1)	16.7	12.5	27.3	29.4	55.2	37	28.6	66.7	45.5
MemorylessCQbyCQ(GPT-4)	25.7	31.2	38.9	17.4	20.9	16.2	48	40.9	46.1
MemorylessCQbyCQ(Llama)	45.6	42.3	23.9	20.5	25.6	41.7	60	60.7	13.6
MemorylessCQbyCQ(o1)	29.3	50	28.6	14.8	0	7.4	10	0	28.6

Expert qualitative analysis. After analyzing the results presented in Tables 4.5 and 4.6 and bar charts in Figure 4.7, we conclude that *OpenAI o1-preview*, especially with Ontogenia, seems to produce less errors overall than *GPT-4*, and provides a better trade-off between different error types, notwithstanding all the generated files showed incorrect domains and ranges. Therefore, *Llama-3.1-405B-instruct-bf16* and *OpenAI o1-preview* are confirmed as the best open-source and closed-source LLMs, respectively, across both previous experiments (Saeedizade and Blomqvist, 2024) and the ones proposed in this work. The former’s performance was not impressive when evaluated by experts however: the ontologies generated by *Llama-3.1-405B-instruct-bf16* show structural flaws, including inconsistent naming, redundant classes, and overlapping domains/ranges. The taxonomy shows circular references and poor property organization. Additional, key issues include malformed cardinality restrictions, lack of comments/labels, and misaligned namespaces. Apart from superfluous elements, flat property hierarchies, and inconsistent axiomatisation, are visible. Table 4.7 shows, for each selected story, the resulting percentage scores from the adequacy assessment of the qualitative evaluation. Different results are calculated as an average between the two. The full qualitative analysis can be seen on Github³¹.

TABLE 4.7: Adequate CQ modelling (%) by Llama and o1 models for the selected stories.

Story	Model	Llama-3.1-405	o1-preview
Music	MemorylessCQbyCQ	0.6	0.9
	Ontogenia	0.86	0.96
Theatre	MemorylessCQbyCQ	0.66	0.73
	Ontogenia	0.63	1.0
Hospital	MemorylessCQbyCQ	0.63	0.73
	Ontogenia	0.66	1.0

4.3.2.5 Discussion

In this section, we discuss the results with respect to the overall results alongside the multidimensional evaluation proposed.

Overall results. According to the findings, and similarly to previous work (Saeedizade and Blomqvist, 2024; Lippolis et al., 2024), the use of an LLM yields promising

³¹<https://github.com/dersuchendee/Onto-Generation>.

results in supporting ontology engineering processes. In the Independent Ontology Generation, Memoryless CQbyCQ and Ontogenia perform well with Single Data Property CQs and Single Object Property CQs, but worse on generating Reifications and Restrictions. For the Incremental Ontology Generation method, Ontogenia and Memoryless CQbyCQ outperformed previous methods, including the students' solutions, with Ontogenia combined with *OpenAI o1-preview* yielding the highest percentage of modelled CQs. However, the structural analysis reveals that with respect to previous work, the two prompting techniques generate more superfluous elements and yield critical pitfalls according to OOPS!. These structural issues are also noted by the experts in their evaluation, which found many flaws in the Llama-generated ontologies related to the usability and understandability of the solutions. While o1-generated ontologies are overall comparable to a student in the case of MemorylessCQbyCQ and even better in the case of Ontogenia.

Comparison of LLMs. By comparing the results of the LLMs in Section 5.4.3.6, we can observe that *o1-preview* is the most promising LLM for ontology generation and GPT-4 may serve as its replacement. However, while Llama exhibits similar performance to GPT-4 in correctly modelling CQs, it produces a considerable number of superfluous elements and critical OOPS! pitfalls and performs the worst according to the expert evaluation. Thus, it may still not be possible to create an ontology engineering co-pilot using open models, and hence the issue of cost and the risks related to the black-box nature of proprietary models, remain.

Impact of the context size. Comparing the scores of Memoryless CQbyCQ with other techniques shows that reducing the input context size of *GPT-4* significantly enhances the output score of this model despite producing some superfluous elements. These results highlight a trade-off between computational efficiency, cost, and output quality. If quality is a priority but resources are limited, Memoryless CQbyCQ balances efficiency and speed, though it may require additional post-processing to refine the output.

Multidimensional evaluation. Our multidimensional evaluation demonstrates that expert assessments align closely with both the standard and structural metrics used in this study. This suggests that experts, perhaps implicitly, rely on similar criteria when evaluating ontologies, resulting in evaluations that comprehensively address most aspects. In particular, if we think of finding a correspondence of expert feedback with existing tools, OOPS flags structural and logical anti-patterns such as missing labels and comments, taxonomic loops, and properties with multiple domains or ranges. The superfluous-components check highlights redundancy, for example unnecessary or duplicate classes and properties, which often co-occur with OOPS alerts when loops or ambiguous labels inflate the vocabulary. CQ verification tests functional adequacy and exposes gaps where requirements are not modelled, which frequently trace back to weak or absent axioms and to misaligned domains and ranges. Finally, some problems such as poorly defined or missing axioms and generally bad or missing taxonomy cut across all checks, reinforcing that automated reports and CQ tests are complementary and still need expert judgement. Therefore, our results highlight the need for multidimensional and qualitative evaluation for a holistic assessment of LLM-generated ontologies.

Superfluous elements. An issue noted in LLM-generated ontologies involves the creation of superfluous properties with the same domain and range or multiple classes and properties that could be considered equal. For example, `employedSince` and `employmentStartDate` are generated for the same CQ, but only one is needed. This raises questions about superfluous elements: their number, why they are generated by the LLM, and their consequences. At the same time, having superfluous elements (in moderate numbers) may be considered less important than having an unmodeled CQs or more complex errors. Thus, our prompting techniques can be considered more complete and more usable than previous work (Saeedizade and Blomqvist, 2024), which was less accurate (Table 4.6), but more concise. In our envisioned setting of a future ontology engineering co-pilot, superfluous elements can be identified, e.g. through OOPS!, and manually removed or avoided with an even more elaborate prompt. As our approach aims to assist rather than replace ontology engineers, we believe this is acceptable.

4.3.3 Ontology generation using Large Language Models

The findings of the two studies presented in this section open multiple avenues for discussion. In the first preliminary Ontogenia study, it has been shown that while the use of MP and Ontology Design Patterns in LLM-driven ontology generation shows promise for richer formalization and a higher complexity of generated structures, significant issues persist. These include the lack of proper property hierarchies, annotation errors, and incorrect domain/range assignments, indicating that the current state of LLM-driven ontology tools may not yet be ready as definitive files for real-world applications. Furthermore, the generation of multiple superfluous elements are noted, causing the produced ontologies to be redundant in conceptualization. However, the results also provided valuable insights that stimulated further discussion and drove refinements in both the methodology and implementation of LLMs in ontology design, in the comparative study.

Therefore, Ontogenia as a context engineering technique has been refined in the second work presented in this section, and it has been enhanced with respect to the inclusion of more instructions about ontology engineering procedures and additional key concepts to refine ontology quality. This technique has then been compared a novel one based on the one proposed in Saeedizade and Blomqvist (2024). Furthermore, a dedicated dataset for these experiments was developed. The refined Ontogenia technique resulted the best performing one with respect to the other setups in the original study according to CQ verification.

The results of this second study showed that the proposed prompting techniques proved promising to support the generation of ontologies that meet a predefined set of requirements, improving the proportion of modelled CQs, surpassing previous approaches and novice ontology modellers. However, some issues persist: challenges like multiple domains or ranges and other pitfalls were highlighted both by the OOPS! pitfall scanner and the expert evaluation. Both prompting techniques often generated superfluous elements, which (in low numbers) are not detrimental to ontology usability, compared to other major errors involving wrong axiomatisation and mistakes in the taxonomy. Regarding to what extent LLMs can be used to support the generation of ontologies that meet a predefined set of requirements, we conclude that current commercial models, such as *OpenAI o1*, can certainly perform on par with non-experts. Again, for what concerns evaluation, we also conclude that it is not sufficient to use one single criterion, nor simple automated evaluation

metrics to detect errors and compare performance. While tools like OOPS! can automatically detect certain issues (missing labels/comments, taxonomy loops, and multiple domains/ranges), more elaborate manual methods are needed to identify other critical problems. Specifically, expert assessment is required to detect superfluous components (unnecessary and duplicate classes/properties), verify proper competency question modeling, evaluate axiom quality, and assess taxonomy structure. This multi-faceted approach combining automated tools with expert review reveals that ontology quality issues span from basic structural problems to more complex semantic and design flaws. The results of these studies combined highlight recurring challenges in LLM-based ontology generation that we identify as superfluous element generation and domain/range specification errors.

These studies also bear limitations. There is a need for more evaluation criteria to assess the utility and generalizability of generated ontologies. Our ontologies, drawn from different domains, are primarily single-module, limiting their reusability across different systems. Moreover, our methods do not adequately prevent or eliminate superfluous elements, as shown by model output repetitions. Furthermore, the use of patterns for the Ontogenia method needs to be further investigated, especially considering the automatization of their choice. All in all, however, the need for a standardized technique for ontology generation that also takes into account newer models such as reasoning ones is fundamental. Additionally, further tests are needed to mitigate the potential leakage effect and bias in LLMs. One way to achieve this is by applying the method to new use cases in different domains. In the next section, we investigate these steps.

4.4 Reasoning-driven ontology generation for specific domains

In the previous section, among the conclusions that were drawn, there was the need to perform additional experiments on both (i) more real-life domains and (ii) having a mechanism that can choose ODPs on-premise to investigate their efficacy better in ontology development. Indeed, the studies shown in Section 4.3 suggest a potential for using LLMs in ontology engineering, where LLMs offer a promising addition by automating key aspects of ontology construction through their language understanding and reasoning capabilities.

Building on these earlier studies that showed the feasibility of LLM-supported ontology generation, we extend the investigation by suggesting a new context engineering technique for reasoning models and examining the generalizability of automated ontology generation across diverse and more specific domains. Neither of these aspects were so far addressed in previous work.

We call reasoning LLMs those models explicitly optimized to produce stepwise, deliberative solutions on complex tasks—through training signals that reward intermediate reasoning (process supervision) and inference-time methods that explore and verify multiple reasoning paths—often coupled with tool use and planning (Lightman et al., 2023; Wei et al., 2022; Yao et al., 2023; Yao et al., 2022). In this chapter we focus on two such families—OpenAI’s *o1* (incl. *o1-preview*) and DeepSeek *R1*—which are positioned as reasoning-centric models (OpenAI, 2024; DeepSeek-AI, 2025). See also Huang and Chang (2023) for a broader survey.

By leveraging a new, curated dataset comprising 95 carefully curated competency questions paired with corresponding user stories, from specific knowledge domains, our approach systematically tests whether these models can generate good

ontology drafts based on requirements that capture domain-specific needs and whether the insertion of ODPs can improve the performance.

4.4.1 Domain-specific ontology generation

In this study, we propose a new, standardized ontology generation technique on different domains, specifically developed for reasoning models³².

4.4.1.1 Methodology

In this section, we outline the methodology employed in this study. We begin by describing the dataset creation process and its statistical properties, followed by the explanation of the prompting technique used in this work. Finally, we present the pipeline for ontology generation.

Dataset creation. To evaluate ontologies generated by LLMs with reasoning capabilities, we constructed a dataset comprising 95 CQs with 12 ontology stories across 6 distinct domains. Each dataset entry includes: (i) a CQ, (ii) a user story (providing the contextual background), (iii) the binary label of the difficulty level of the CQ (easy or hard), and (iv) the domain of the ontology. Table 4.8 shows statistics for the dataset.

Dataset composition. The CQs were extracted from four sources: *Onto-DESIDE*³³ (domain: Circular Economy), the *Polifonia Project* (domains: Music and Events) (Berrardinis et al., 2023), the *WHOW Project* (domain: Water and Health) (Lippolis, Lodi, and Nuzzolese, 2025) and *AquaDiva* (domains: Microbe Habitat and Carbon and Nitrogen Cycling) (Algergawy et al., 2024). In particular, these derive from carefully curated large-scale projects, with the fourth one being a semi-automatically annotated dataset. Part of these derive from the annotated dataset already developed in the study of Section 4.3.2.

CQ classification. To ensure balanced and comparable evaluations, each CQ was classified into one of two categories, “Easy” and “Hard”, based on the complexity required for its formal representation. An LLM, *o1-preview*, was used to assign these labels by estimating the difficulty level of classes and properties needed to model the CQ. Specifically, if a CQ required at most 2 classes and 1 property (either a data or an object property), it was classified as “Easy”; otherwise, it was classified as “Hard”. The labels were then checked manually by an ontology engineer. This labelling strategy was implemented to maintain a consistent distribution of CQ difficulties across all domains.

³²Data, code and other material for this study is available at this link: <https://github.com/dersuchendee/Domain-OntoGen>.

³³<https://ontodeside.eu>.

TABLE 4.8: Domain Statistics for Competency Questions.

Domain	Total CQs	Easy	Hard
Circular Economy	16	5	11
Music (Bells, Organs, Tunes)	16	10	6
Events	18	8	10
Microbe Habitat	15	8	7
Carbon and Nitrogen Cycling	15	4	11
Water and Health	15	7	8
Total	95	42	53

Dataset annotation. The labelling of the dataset for the verification of CQs has been carried out manually by two ontology engineers. Following the error classification outlined in previous work (Lippolis et al., 2025b)(§ 4.3), the output ontologies were considered to be modelled, not modelled, or not modelled due to a minor error.

Context engineering technique. The context engineering technique employed in this study follows a few-shot methodology (§3.0.4.3), with some examples, building on and distinguishing it from previous approaches which mainly employed a decomposed prompting technique (Lippolis et al., 2024; Lippolis et al., 2025b; Saeedizade and Blomqvist, 2024). This choice is motivated by the observation that reasoning models have been shown in Lippolis et al. (2025b) and Section 4.3 to perform effectively, and model cards state they work best with minimal instructions rather than being explicitly guided to think step by step³⁴. After trials with and without a few-shot example, the performance of the few-shot has proven more accurate than the other. The prompt is structured into three key components:

- **Instruction:** A description of the task (ontology generation) and its objective (ontology must model the input CQ).
- **Example:** A Turtle syntax sample ontology model corresponding to a CQ and its associated ontology story, illustrating an ideal response.
- **Input:** A designated section where a new CQ and ontology story are inserted, prompting the LLM to generate an ontology model accordingly.

Ontology generation method. We used the Independent Ontology Generation method described in Lippolis et al. (2025b) and in Section 4.3.2.4. In this approach, each CQ and its associated ontology story are provided to an LLM through a prompt to generate the corresponding ontology. This method treats each CQ as an independent unit, enabling a precise evaluation of the ontology generation process on a per-question level. By handling CQs in isolation, the impact of the settings can be assessed.

4.4.1.2 Evaluation setup

In this section, we describe the experimental setup carried out for the evaluation. To run the experiments, we used Azure API to call *o1-preview* and the original DeepSeek API with default hyperparameters. The prompt with a fixed example (few-shot) is appended with ontology story and CQ and sent to the API to get the resulting output file. The results are saved in a Turtle file.

³⁴<https://platform.openai.com/docs/guides/reasoning-best-practices>.

LLMs. As already suggested earlier, we have selected two large language models with reasoning capabilities (§4.4): DeepSeek R1 and OpenAI o1-preview. This selection was motivated by the desire to perform a comparative analysis between the current state-of-the-art closed-source and open-source LLMs. The OpenAI *o1-preview* model was accessed through the Microsoft Azure API, whereas DeepSeek was used according to the specifications detailed in its official documentation³⁵.

Context engineering technique. The prompt used in this work is available in the Appendix (A). Initially, several prompt configurations were evaluated, including versions with and without examples (i.e., few-shot and zero-shot settings). The results demonstrated that prompts incorporating examples consistently outperformed those without, justifying their use. Furthermore, the impact of varying the complexity of the examples was examined in a separate experiment involving multiple ontology generation tasks. In the interest of conciseness, the detailed results of these additional experiments are not reported here.

Hyperparameters. The default hyperparameters for the LLMs, including temperature and penalty, were employed throughout this study. This means that since reasoning models require a temperature of 1 as default, this had to be selected, hindering more deterministic outputs.

4.4.1.3 Results

Overall, both the *o1-preview* and DeepSeek R1 models exhibited high and comparable accuracy. Initially, nine and ten CQs remained unmodelled for the *o1-preview* and DeepSeek R1 models, respectively, out of a total of 95 CQs. However, when minor modelling issues are excluded, the number of unmodelled CQs decreases to eight for the *o1-preview* model and five for the DeepSeek model.

Table 4.9 provides a breakdown of mistakes on both easy and hard CQs by domain. Regarding the *o1-preview* scores, performance was consistent across all domains, suggesting that the ontology generation method is broadly applicable rather than being limited to a specific domain. With slight differences, although DeepSeek generally demonstrated similar consistency, the scores for the Events domain were noticeably lower, indicating a need for further investigation into the model’s performance in that particular domain.

TABLE 4.9: Count of errors in modelling for each LLM, divided into Easy and Hard competency questions grouped by ontology domain.

Category	o1-preview					DeepSeek				
	Not modelled		Contains a minor issue		Total	Not modelled		Contains a minor issue		Total
	Easy	Hard	Easy	Hard		Easy	Hard	Easy	Hard	
Circular Economy	0	0	0	2	2	0	0	0	0	0
Music	0	1	1	0	2	2	0	0	0	2
Events	0	0	1	1	2	1	2	0	2	5
Microbe Habitat	0	0	0	0	0	0	0	1	0	1
Carbon and Nitrogen Cycling	0	0	0	1	1	0	0	0	1	1
Water and Health	0	0	0	2	2	0	0	1	0	1

4.4.1.4 Discussion

Our results reveal a performance trade-off between the two models. Specifically, while *o1-preview* tends to commit more “minor” errors, DeepSeek is prone to a higher incidence of incomplete modellings. When we consider cases with “minor” mistakes

³⁵<https://api-docs.deepseek.com/>.

equivalent to complete modellings, both models demonstrate similar performance across the domains. This finding suggests that differences in error types may not translate into significant disparities in overall outcomes.

Domain generalizability. Our work extends previous findings by Lippolis et al. (2025b)(§ 4.3) by showing that reasoning models working for a specific domain can generalize effectively across other datasets. Therefore, the improvements achieved in earlier studies are not limited to a single domain, affirming the broader applicability of the developed techniques.

Reasoning models architecture. One possible explanation for the similar outputs observed from both *R1* and *o1-preview* is their reasoning capabilities. It is plausible that these models benefit from mutual fine-tuning or shared architectural strengths. The collective evidence points to a significant enhancement in reasoning with respect to models with non-reasoning capabilities, which appears to underlie their comparable performance.

Difficulty levels of the requirements. Furthermore, the analysis indicates that both models handle hard and easy CQs with similar performance. This observation challenges the common assumption that increased question complexity necessarily results in lower scores. Instead, our findings suggest that the models are robust enough to process questions of varying complexity without a marked performance penalty. Nonetheless, this is limited to only one evaluation criterion.

Error analysis. Finally, a closer look at the errors reveals that the majority occur solely within the Events domain. The corresponding CQs in this category are much shorter than those in other domains, but there are multiple user stories that serve as requirements. This pattern could suggest that longer, more descriptive CQs could provide additional context that aids the models in processing and reasoning, leading to improved performance, especially if multiple scenarios are provided. Therefore, the lack of standardization in requirements is also fundamental when dealing with automated techniques.

Overall, these insights highlight the importance of both model architecture and the design and selection of requirements in achieving balanced and generalizable performance. One thing left unaddressed is how the system would behave with the insertion of ODPs in this loop, an automatic one so the person should not manually insert them.

4.4.2 RAG-Based Ontology Extension with ODPs

In this study, we experiment with ontology extension with patterns. This experiment investigates whether an LLM can automatically select and reuse the most relevant ODPs at run-time basing on the actual requirement, rather than relying on a hand-curated list. We extend the previous workflow with a lightweight Retrieval-Augmented Generation (RAG) layer that surfaces candidate patterns and injects them into the prompt. RAG is a technique that couples a generator (the LLM) with a retriever over an external knowledge source so the model conditions its output on retrieved evidence rather than relying only on parametric memory (Lewis et al., 2020c). Concretely, a query triggers retrieval of relevant items (documents,

passages, or structured artefacts), and the generator conditions on them during decoding, which supports better factuality, provenance, and easy knowledge updates by modifying the index rather than the model (Gao et al., 2023). In our setting, the retriever returns candidate ODPs and concise pattern summaries; the LLM then uses these as schema building blocks to propose ontology extensions that reuse the retrieved ODPs instead of inventing ad-hoc structures.

4.4.2.1 Methodology

The methodology builds on three stages: (i) corpus acquisition and preprocessing, (ii) vector store construction, and (iii) retrieval-augmented ontology generation.

Corpus acquisition and pre-processing. At start-up the system clones the latest version of the ODP repository from GitHub³⁶ and caches it locally. For every pattern the Markdown description is parsed to extract *Name*, *Intent*, *Domain(s)*, *Elements*, and any associated *Competency Questions*. The resulting JSON catalogue is persisted for auditability.

Vector store construction. The *Intent* field (falling back to the pattern name if missing) is embedded with `text-embedding-3-small` and indexed in an in-memory FAISS store. The intents are short, well-formed phrases, so a lightweight embedding is sufficient to capture topical similarity without paying the latency and cost of larger models. `text-embedding-3-small` provides stable vectors with good semantic coverage for sentence-length inputs³⁷, which makes it a practical choice for frequent re-indexing of the ODP repository. For indexing we adopt FAISS, a widely used library for efficient nearest-neighbor search in dense vector spaces, optimized for memory and speed and proven in large-scale IR scenarios³⁸. Because pattern intents often contain distinctive terminology, we combine embedding similarity with a light lexical prior based on metadata completeness and domain term overlap. This hybrid scoring follows established practice that dense semantic matching benefits from a complementary sparse signal when queries are short and vocabulary is informative³⁹.

RAG for ODPs. Given a competency question (CQ) and its optional user story, derived from the already developed dataset of Section 4.4.1, the retriever returns the top- k patterns (default $k = 8$). Any accompanying `.owl` or `.ttl` artefacts are truncated to 3 000 characters to respect the token limitations. A single LLM call then produces a Turtle fragment, which is isolated from the chat transcript via a regular-expression post-filter.

4.4.2.2 Evaluation setup

Each output was assessed by a single ontology engineer using the CQ-verification protocol from Section 4.3.2.1. We compare three configurations: a domain-specific baseline, our framework paired with DeepSeek *R1*, and our framework with retrieval-augmented generation using *o1-preview*.

³⁶<https://github.com/odpa/patterns-repository>

³⁷<https://platform.openai.com/docs/models/text-embedding-3-small>

³⁸<https://faiss.ai/>

³⁹For background on sparse–dense complementarity in retrievers see, e.g., <https://docs.haystack.deepset.ai/docs/retrievers>

4.4.2.3 Results

Integrating ODPs with OpenAI’s *o1-preview* model produced the most comprehensive coverage of the competency questions: 87 were fully modelled, 5 were only partially satisfied, and 3 were unsupported. This performance surpasses both the original approach and the DeepSeek variant of study Section 4.4.1.

TABLE 4.10: Automatic Pattern Retrieval via RAG.

Setting	CQs modelled (%)	Minor issues (%)	Not modelled (%)
IndOntoGen (o1)	90.5	8.4	1.1
IndOntoGen (DeepSeek)	89.5	5.3	5.3
IndOntoGen + RAG (o1)	91.6	5.3	3.2

4.4.2.4 Discussion

The RAG layer yields a modest but meaningful gain in coverage, and converts several “minor” gaps into fully modelled CQs. Qualitatively, the LLM now reuses at least one pattern in nearly every answer, confirming that the retrieval signal is influential. The main drawback is occasional pattern noise: mis-aligned domains, redundant imports, or half-applied restrictions when a retrieved pattern is only partially relevant. A lightweight post-filter or a user approval step could mitigate these issues. Still, a more comprehensive evaluation on a wider set of ontologies is needed to confirm this preliminary assessment.

4.4.3 Establishing design principles for pattern-based, domain-specific ontology generation

The studies outlined in this section aim to advance our understanding of reasoning-driven ontology generation in several key directions. First, we established a standardized few-shot prompting technique specifically tailored for reasoning models, demonstrating that minimal instruction paired with concrete examples outperforms more complex decomposed prompting approaches for models like DeepSeek’s *R1* and OpenAI’s *o1-preview*. Second, through systematic evaluation across six diverse domains using a carefully curated dataset of 95 competency questions, we demonstrated that reasoning-capable LLMs exhibit robust cross-domain generalizability, with both models achieving consistent 85-90% coverage rates regardless of domain complexity or CQ difficulty level.

The comparative analysis between *R1* and *o1-preview* revealed complementary error patterns: while *o1-preview* tends toward minor modeling issues that are easily correctable, *R1* exhibits higher rates of incomplete modeling but fewer systematic errors when successful. This finding suggests that the choice between open and closed reasoning models may depend more on downstream workflow requirements than on absolute performance differences, though the differences in this data are very subtle.

A critical limitation identified in previous sections was the reliance on manually curated ontology design pattern selection. The proposed RAG-based extension addresses this gap by demonstrating that automatic, context-aware pattern retrieval can improve ontology coverage while maintaining quality standards. The system successfully injects relevant patterns in nearly every generation task, though challenges remain with domain misalignment and partial pattern application. These

findings establish the feasibility of fully automated pattern selection, but highlight the need for more complex filtering mechanisms and experimentation on a wider set of data.

Collectively, these results establish reasoning-driven generation as a viable approach for domain-specific ontology construction, with clear pathways for further automation through intelligent pattern integration.

4.5 Ontology evaluation with LLMs

LLMs have changed the ontology engineering landscape. What should not change, however, is the framework that we should use to evaluate these tasks. Having established that LLMs can generate ontologies with the Ontogenia method (§4.3), and that we can use them for both domain-specific ontology generation and potentially even for automatic ODP selection (§4.4.1), after having used manual evaluation in all these previous setups as a part of a holistic evaluation framework, the natural next question is how to evaluate these artifacts systematically. This motivated our works on LLM-assisted evaluation, which we will present in this chapter.

Ontology evaluation would be one step after ontology generation. The benefits lie in the fact that automating the evaluation would require less people to do so. This section reports three independent studies that evaluate ontologies with LLMs along three complementary axes: (i) Terminological adequacy (does the ontology cover domain terms?), (ii) Functional adequacy via CQ-grounded inference/modeling (does an ontology enhance LLM reasoning and CQ verification?), and (iii) Human-in-the-loop effectiveness (to what extent do LLM suggestions assist engineers evaluate ontologies in CQ verification?).

4.5.1 Early approaches to ontology evaluation

The first mention of CQs for ontology development cycles can be traced back to 1996, where Uschold and Gruninger described CQs as ontology requirements bridging the gap between informal questions and formal ones (Uschold and Gruninger, 1996). In this view, CQs are natural language queries an ontology must be able to answer to evaluate the expressiveness of the ontology and constrain the ontology design search space (Noy and Hafner, 1997). Later, Vrandečić and Gangemi (2006) proposed treating ontology requirements as unit tests, turning each CQ into a query whose results on the ontology are checked against expected outcomes. This work introduced CQ-driven validation as a distinct evaluation dimension, termed *functional evaluation*, related to the intended use of an ontology and separate from structural analysis (Gangemi et al., 2006). This evaluation kind is opposed to the structural one concerning schema quality and to usability-profiling concerns such as labels, documentation, and metadata completeness.

The importance of CQs was further cemented in main agile ontology engineering methodologies, as seen in Section 4.1.1.2, like METHONTOLOGY (Fernández-López, Gómez-Pérez, and Juristo Juzgado, 1997) and eXtreme Design (Blomqvist et al., 2010), which included CQs as part of the requirement specification and verification phases. Building on this foundation, Blomqvist, Seil Sepour, and Presutti (2012) proposed a methodology for ontology testing that uses CQs as requirements (see Section 4.1.1.3). They showed that reformulating CQs into SPARQL queries effectively identifies mistakes. In fact, CQ verification has been used for over a decade

in several projects (Carriero et al., 2019; Lippolis et al., 2025b; Litovkin et al., 2020; Saeedizade and Blomqvist, 2024; Zhang et al., 2024; Zhang et al., 2025).

At the same time, manually writing these queries is perceived as tedious work, pointing directly to the automation opportunity our work addresses. With the advent of LLMs, several approaches have emerged to assist various ontology engineering tasks. LLMs have been employed to evaluate domain-specific CQs, as demonstrated by Di Nuzzo et al. (2024) on criteria such as relevance, clarity, and depth. While promising, this work lacked comprehensive qualitative evaluation on a larger dataset (and comparison between LLM-generated metrics and human judgements) involving human engineers. In another work, Tufek et al. (2024) reported using GPT-4 to automatically convert CQs and other textual requirements into SPARQL queries for ontology validation. Similarly, Alharbi et al. (2024d) mentioned as a future direction of the work that LLMs could assist in determining whether a CQ truly aligns with an ontology's knowledge scope by translating CQs into SPARQL queries. Researchers have also begun exploring LLMs for ontology evaluation. Tsaneva, Vasic, and Sabou (2024) tested GPT-4 in the chat interface as an evaluator for ontology axioms compared to human experts. However, this kind of ontology evaluation is structural, not functional, and was carried out on one toy ontology only. Benson et al. (2024) investigated GPT-4's ability to generate and evaluate ontology class definitions in the chat interface, following the Basic Formal Ontology. Despite suggesting that human-in-the-loop refinement shows potential for using GPTs as productivity tools for ontology-related tasks, the work was not tested on a larger-scale benchmark, but rather on selected example classes. Furthermore, the paper addresses structural correctness but not functional evaluation.

Thus, despite promising results, the use of LLMs for ontology evaluation, especially functional adequacy, remains underdeveloped. In this context, Garijo et al. (2024) leave the categorisation of existing resources on LLM use for ontology evaluation blank, highlighting this gap. Furthermore, current evaluation approaches for LLM-generated ontologies face significant methodological challenges. As shown in Section 4.3.2, Lippolis et al. (2025b) proposed a comprehensive evaluation method, structural and functional, for LLM-generated ontologies that still, similar to Alharbi et al. (2024c), relies heavily on manual components requiring the intervention of a domain expert. Similarly, Silva et al. (2024) developed LLM-based ontology generation techniques that depend on manual verification processes.

Rebboud et al. (2024a) propose evaluation criteria including accuracy, completeness, and conciseness of generated ontologies measured against a ground truth. However, these criteria introduce two critical limitations: (i) they necessitate either a gold reference taxonomy or a comprehensive domain corpus: resources that are so far unavailable or incomplete, and (ii) they fail to explicitly incorporate ontology requirements in the evaluation process.

Also, a fundamental tension exists between expecting LLM evaluation to provide definitive answers versus using it as a decision support tool to assist human judgment. This distinction becomes crucial when establishing evaluation methodologies that balance automation with human-in-the-loop oversight.

Another important aspect in this field concerns the quality of requirements that seems to significantly impact ontology generation success (Lippolis et al., 2025a) (§ 4.3.2). Therefore, methods to evaluate ontologies against their intended functional requirements need to account for the quality of those requirements, especially when derived from multiple projects. To sum up, using LLMs as evaluators in ontology engineering is only now beginning to emerge.

4.5.2 Terminological coverage evaluation

As a first study in ontology evaluation, we use LLMs for ontological terminological coverage, where an ontology network includes the analysis of its ability to cover domain-specific terminology with respect to other available ontologies (Roche, 2012). In this sense, terminological coverage is seen as an ontology alignment problem between a selected, automatically extracted vocabulary and the target ontologies.

There are multiple ways to extract such a vocabulary: starting from a predefined set of documents, Carriero et al. (2020a), this has been done with Rapid Automatic Keyword Extraction (RAKE); Demaidi and Gaber (2019) used a custom keyword extraction tool, TONE, while Zaitoun, Sagi, and Hose (2023) focused on a domain-tuned named-entity-recognition-model. As LLMs have been proved to absorb, infer, and reproduce information from self-supervised training on large quantities of text (Patil and Gudivada, 2024), we expand the approach of Carriero et al. (2020b) not only to two different RAKE algorithms (nlk: Sharma, 2023 and RAKE-keyword Prashant, 2023), but also to three different LLMs: GPT (in the Chat, version *GPT-4*), LLaMa (version *llama-3-sonar-large-32k-online*), and Claude (version *3-haiku*).

4.5.2.1 Ontologies under consideration

The study was applied in the context of the WHOW-KG (see Section 4.1.1.3), where we have taken as a reference the Directive 2000/60/EC establishing a framework for Community action in the field of water policy (European Union, 2000). The ontologies to compare the results with have been selected basing on the domain relatedness and on the availability of the ontology network, along with descriptions, labels, or comments of it, online. They are DOCE (NEMO-UFES, 2023), aimed to produce and analyse water quality data of the Doce River Basin, and SAREF4WATR (ETSI, 2023), the SAREF extension for the water domain. Furthermore, we included the ENVO ontology (Buttigieg et al., 2013), a general-purpose representation of environments, environmental processes, ecosystems, habitats, and related entities.

4.5.2.2 Methodology

The methodology incorporates five distinct approaches for vocabulary extraction. The traditional automated keyword extraction techniques include two implementations of the Rapid Automatic Keyword Extraction algorithm: the NLTK (Sharma, 2023) and RAKE-keyword (Prashant, 2023) derived from the corresponding GitHub repositories. These implementations represent established approaches to automated keyword extraction that rely on statistical and linguistic features to identify relevant terms.

The LLM-based vocabulary generation component of our methodology incorporates three distinct large language models accessed through their respective chat interfaces. *GPT-4* was accessed through OpenAI's ChatGPT web interface, representing the current state-of-the-art in conversational AI capabilities. *LLaMA* was utilized through Perplexity's implementation of the *llama-3-sonar-large-32k-online model*, providing access to Meta's large language model architecture. *Claude-3-Haiku* was employed through Anthropic's interface, offering an alternative LLM perspective on

vocabulary generation. The selection of these three models ensures broad coverage of different LLM architectures and training approaches⁴⁰.

For the LLM-based vocabulary extraction, we developed a prompting strategy that provides the models with the complete text of the EU Water Framework Directive and explicitly requests the generation of comprehensive lists of water-related terminology. The prompts were developed to elicit domain-specific vocabulary while avoiding bias toward particular ontological structures or terminological frameworks. Each model was presented with identical prompting instructions to ensure consistency in the vocabulary generation process (see the Appendix for the supplementary material.).

The coverage assessment process was implemented using the Terminological and Ontological Coverage (TOCo) package, which was specifically developed for this evaluation and subsequently made available as an open-source Python package⁴¹. The package automates the vocabulary extraction process for each method, processes the target ontologies to extract their terminological content including class labels, property names, annotations, and comments, and calculates coverage metrics by determining alignment between extracted vocabularies and ontological concepts. The evaluation metrics focus primarily on terminological coverage percentage, calculated as the proportion of extracted vocabulary terms that achieve successful alignment with concepts in each target ontology. Supplementary metrics include weighted coverage scores that incorporate term frequency and relevance factors.

4.5.2.3 Results

TABLE 4.11: Measurement Results.

Measurement	WHOW Ontology Network	DOCE	SAREF4WATR	ENVO
Rake-nltk	7.22%	3.73%	5.38%	18.79%
Rake-keyword	6.28%	3.17%	4.01%	13.92%
GPT	69.93%	39.89%	54.85%	58.15%
LLaMa	10.03%	10.41%	0%	16.24%
Claude	58.40%	44.54%	38.66%	81.09%

The results in Table 4.11 show that the WHOW Ontology Network exhibits broader terminological coverage compared to the other two ontologies related to water monitoring, i.e. DOCE and SAREF4WATR, which represents our core domain of interest. Conversely, ENVO provides a slightly better terminological coverage than WHOW for most of the measurement applied.

4.5.2.4 Discussion

The result outlined in Table 4.11 is not recorded consistently as in cases where GPT is used, WHOW outperforms ENVO in terms of terminological coverage. This is

⁴⁰The LLM-based runs in this early study were conducted via consumer chat interfaces rather than the APIs due to budget and access constraints at the time. This choice limits strict reproducibility and control over decoding parameters, session state, and model updates, and it may introduce interface-level effects such as unseen system prompts or retrieval settings in hosted online modes. To mitigate these risks we refer to the published article about this work: Lippolis, Lodi, and Nuzzolese (2025), where we record, in the supplementary material, the exact interface, provider, model name as displayed by the UI, the full prompts and responses, and the list of keywords.

⁴¹<https://github.com/dersuchendee/terminological-ontological-coverage>

because ENVO is a large, general-purpose ontology in the context of the environmental domain with a high likelihood of containing keywords relevant to the water directive, thus not exactly comparable to the purposes of the WHOW project. In contrast, the WHOW Ontology Network is specifically designed to address selected use cases, focusing on water monitoring while also serving as a foundation for the knowledge graph with integrated data. Therefore, this automated evaluation performs differently at various levels of granularity.

4.5.3 Inference and modeling evaluation

LLMs enable additional ontology evaluation methodologies beyond terminological analysis. These include assessment of ontological utility through measurement of inference capability enhancement, integration within retrieval-augmented generation frameworks (see §4.4.2) for inconsistency detection, and application to competency question verification processes, particularly for functional evaluation of ontological reasoning support. In this section, we show a preliminary examination of how this could be carried out.

4.5.3.1 Ontology and dataset under consideration

To carry out inference and modeling evaluation, we chose the BDI ontology, which is an ontology about Beliefs, Desires and Intention⁴². The BDI ontology models an agent's mental reality via (i) mental states (Belief, Desire, Intention) as endurants tied to propositions, justifications, and time, and (ii) mental processes—Belief/Desire/Intention formation and Planning—as perdurants that generate and evolve those states. It reuses DUL-aligned ODPs (EventCore, Situation, TimeIndexedSituation, Provenance, BasicPlan) to connect agents, world states, goals, and plans. Causal links capture the BDI pipeline, intention–plan specification, and external triggers. Formal DL and FOL axioms enforce disjointness, cardinalities, and functional roles, enabling temporal and provenance-aware querying over state histories. Requirements are organised by the competency questions table (Agents and Mental Entities; World and Mental States; Dynamics; Goals and Planning; Temporal Reasoning).

In this section, we show how the BDI ontology can provide the symbolic backbone necessary to augment a Large Language Model to a neuro-symbolic agent by improving explainability, interpretability, and cognitive plausibility of its agency. To this extent, we reuse the Microsoft Locations and Times of Task Execution (MS-LaTTE) dataset, developed by Jauhar et al. (2021). It gathers more than 10,000 tasks along with likely locations and times of day at which it is completed. We filtered the dataset to include only annotations with high inter-annotator agreement, meaning judges must have a consensus about the inclusion of a specific location and time, applying a threshold of 0.70 for Location annotations and 0.65 for Time annotations (the lower threshold for Time reflects the greater subjectivity noted in the original paper). This filtering method ensured we retained only tasks with reliable annotator consensus.

4.5.3.2 Methodology

For the experimental setup, we randomly choose 47 tasks from the filtered dataset upon which we designed two main tests to evaluate both the inference and modeling capabilities of LLMs, specifically assessing their ability to handle the 15 CQs.

⁴²<https://github.com/fossr-project/ontology/blob/main/bdi-v0.2.rdf>

As a result, we obtain 13 subtests (9 inference tests and 4 modelling tests). In two cases, we augment the tasks to add information on external events and a planning process, which are present in the CQs but not contemplated in the original dataset. The experiments are performed on *GPT-4o* with the OpenAI API, always in two configurations: with and without the ontology in the prompt⁴³.

4.5.3.3 Results

Inference accuracy. The first test concerns inference accuracy, which evaluates the model’s ability to detect logical inconsistencies, contradictions, and perform reasoning about agent mental state with respect to a gold standard,

In roughly a quarter of the inference subtests (2 out of 9), the ontology-augmented approach uncovered inconsistencies that the plain model overlooked. In the other cases, the two approaches led to the same conclusions. Furthermore, the LLM follows the ontology modelling correctly. Below, we show two examples of the task and the results for the Location Contradiction Detection subtest, in which the prompt lists three everyday to-do items and asks the system whether each task–location pair is logically sound.

TABLE 4.13: Location Contradiction Detection subtest data.

Task ID	Task Description	Location	Time
12693542	Check into hotel	Home	WE-morning
10169250	Buy new water filters	Work	WD-evening
5030401	Workflow meeting	Public	WD-afternoon

We then prompted the LLM both with and without the ontology. Both outputs correctly identified that checking into a hotel while at home is impossible, but the ontology-augmented LLM additionally detected that buying water filters while at work also represents a contradiction. The triples generated by the LLM are shown below:

```
@prefix bdi: <https://w3id.org/fossr/bdi/> .
```

```
:Task_12693542 a bdi:Intention ;
  bdi:hasProposition "check into hotel" ;
  bdi:atLocation :Home .
```

```
:Task_10169250 a bdi:Intention ;
  bdi:hasProposition "buy new water filters" ;
  bdi:atLocation :Work .
```

```
:Task_5030401 a bdi:Intention ;
  bdi:hasProposition "workflow meeting" ;
  bdi:atLocation :Public .
```

```
:Belief_ErrandContext a bdi:Belief ;
  bdi:hasProposition "Personal errands should occur
```

⁴³Material along with further details is available on the GitHub repository⁴⁴

in a personal context (home or store)".

```
:Conflict_HotelHome a bdi:Justification ;
  bdi:justifies :Task_12693542 , :Belief_ErrandContext ;
  bdi:hasProposition "Hotel check-in cannot happen at home" .

:Conflict_ErrandAtWork a bdi:Justification ;
  bdi:justifies :Task_10169250 , :Belief_ErrandContext ;
  bdi:hasProposition "Buying water filters
  at work violates context norm" .
```

Furthermore, the test shows that the output correctly models CQ5 and CQ8.

Modeling abilities. The second test concerns modelling completeness, which instead shows if the answer explicitly grounds its analysis in the ontology, so if it generates appropriate RDF triples. In both cases, we also check if it answers the CQs.

In the modelling test as well, the LLM produces plausible and correct ontology triples from the task-based data. Below, we show one example of modelling for this test, which mirrors the modeling of CQ8 and CQ10 (category of Dynamics of Mental States).

In this scenario, the agent receives a push-notification from their banking app at 10:15 AM, stating that “Ghadeh has requested \$250 via Zelle”. Upon receiving this external event, the agent begins deliberating whether to fulfil the request and how to do so. This deliberation leads to the formation of an intention to send the payment, which is subsequently recorded as a candidate task in the agent’s plan manager: “Task ID 11439167: *pay Ghadeh* (suggested for weekday afternoon)”.

The BDI ontology allows us to explicitly trace this causal and cognitive path from event to mental process to intention to plan through formalization in Turtle syntax.

```
@prefix bdi: <https://w3id.org/fossr/bdi/> .
@prefix ex: <http://example.org/bdi-demo/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
ex:Evt_2025-03-02T10:15-ZelleRequest
  a bdi:ExternalEvent ;
  rdfs:label "Ghadeh requests $250 via Zelle" ;
  bdi:occursAt "2025-03-02T10:15:00-05:00"^^xsd:dateTime ;
  bdi:triggers ex:Deliberation_1 .

ex:Deliberation_1
  a bdi:IntentionFormation ;
  bdi:isProcessOf ex:AgentA ;
  bdi:generates ex:Intention_PayGhadeh .

ex:Intention_PayGhadeh
  a bdi:Intention ;
  bdi:isIntentionOf ex:AgentA ;
  bdi:fulfills ex:Goal_SettleDebt ;
  bdi:hasProposition ex:Prop_PayGhadeh .

ex:Goal_SettleDebt
```

```

a                                bdi:Goal ;
bdi:hasProposition ex:Prop_PayGhadeh .

ex:Prop_PayGhadeh
a                                bdi:Proposition ;
rdfs:label                      "Transfer $250 to Ghadeh via Zelle" .

ex:Task_11439167
a                                bdi:Plan ;
bdi:hasProposition ex:Prop_PayGhadeh ;
bdi:suggestedTime  "WE-afternoon" ;
bdi:isPlanOf       ex:AgentA .

```

4.5.3.4 Discussion

The experimental results reveal differential performance improvements across the various competency question categories, suggesting that ontological augmentation provides variable benefits depending on the type of reasoning required. The successful demonstration of CQ3 and CQ4 from the World and Mental States cluster, alongside CQ7 and CQ9 from the Dynamics of Mental State Formation category, indicates that ontology-augmented LLMs particularly excel in tasks requiring explicit grounding of mental states to world conditions and causal reasoning about mental processes. This pattern suggests that ontological structures are most beneficial when the reasoning task involves complex relational dependencies and temporal causality, rather than simple factual retrieval or classification tasks. A systematic analysis of which competency question categories consistently benefit from ontological augmentation could inform future ontology design priorities and help identify the specific types of reasoning that most require formal symbolic support.

Furthermore, from these tests, we can note that although the current BDI ontology deliberately does not explicitly model intention–intention conflicts, it still enables the surfacing of internal contradictions when incompatible intentions are grounded in mutually exclusive beliefs or goals, basing on the context. This design choice is deliberate, as it keeps the core vocabulary flexible, in that agents are allowed to possess mutually incompatible intentions before committing, and it is the reasoner (or a Logic-Augmented Generation layer, see Section 4.2.3.1) that surfaces the contradiction basing on the context, not the ontology axioms themselves.

The generated RDF triples also suggest that, as already shown in previous sections, LLMs possess an emergent understanding of formal ontological structures. Within a LAG-based framework, the precise use of ontological predicates in contextually appropriate ways indicates that the LLM successfully internalized the semantic relationships encoded in the BDI ontology without explicit training on RDF syntax or ontological modeling principles. However, the consistency and reliability of this modeling capability across different domains and ontological complexities remains an open question that warrants systematic investigation, particularly regarding the model’s ability to assess and maintain semantic coherence when working with more abstract or less well-represented ontological concepts. Overall, the studies on LLM-assisted on ontology evaluation presented so far are preliminary and have shown that, first, terminological coverage differs by ontology and method; second, ontology-augmented inference/modeling can surface contradictions missed by

a plain model in roughly a quarter of subtests and produce CQ-aligned RDF, indicating emergent but still preliminary modeling competence. In the next section, we discuss an experiment on assisted ontology evaluation, which will partly answer our questions.

4.5.4 Assisted ontology evaluation

LLMs present opportunities to serve as tools for assisted ontology evaluation, providing automated assessment capabilities that can complement traditional evaluation methodologies which, as we have seen (Section 4.1), tend to be manual. The foundation for such assisted evaluation requires the construction of appropriate evaluation datasets that capture the specific assessment criteria and domain requirements. OE-Assist is a framework for ontology evaluation with two complementary settings: an automatic mode that compares LLM-predicted YES/NO labels against a gold standard, and a semi-automatic mode where ontology engineers make the final decision after seeing LLM-suggested labels and SPARQL queries.

4.5.4.1 Dataset creation

Our primary dataset, called OntoEval, consists of 1,393 ontologies, both human-curated ones from European projects (57) and LLM-generated ontologies (1,336), each paired with a corresponding CQ, user story and annotations. Including potentially imperfect, LLM-generated ontologies is intentional: real-world evaluation tools must handle noisy or partially correct models, and such ontologies can provide controlled but realistic variation. OntoEval ontologies, stories and CQs are mostly derived from multiple recent works (Lippolis et al., 2025a; Lippolis et al., 2025b; Saeedizade and Blomqvist, 2024) (see Section 4.3), in which ontologies were manually annotated for their evaluation with respect to the CQ Verification method into three labels: YES, NO, and NO MINOR. The latter refers to the definition by Lippolis et al. (2025b) and Saeedizade and Blomqvist (2024) (see 4.3.2.1) and indicates that only a single property is missing from the ontology to achieve complete CQ verification. In the automatic evaluation of this work, we treat NO MINOR as NO. Furthermore, CQs were assigned a difficulty rating on a binary scale, SIMPLE or COMPLEX, based on the presence of at most two classes and one object property for the SIMPLE label, otherwise it was considered COMPLEX (see §4.3.2.2).

Ontologies and their requirements were drawn from multiple sources to ensure a variety of domains and modelling patterns, and in particular, four large-scale European projects, one educational project with three ontologies, and three LLM-generated ontologies from recently peer-reviewed work:

1. *Polifonia Project* (domains: Music and Events)(Berardinis et al., 2023);
2. *WHOW Project* (domains: Water and Health) (Lippolis, Lodi, and Nuzzolese, 2025);
3. *Onto-DESIDE Project*⁴⁵ (domain: Circular Economy);
4. *IKS* (Interactive Knowledge Stack for content management platforms)⁴⁶;

⁴⁵<https://ontodeside.eu>

⁴⁶<https://cordis.europa.eu/project/id/231527>

5. “SemanticWebCourse”, used in Saeedizade and Blomqvist (2024) and Lippolis et al. (2025b), consisting of three ontology stories about the music, theatre and hospital domain and 15 CQs each;
6. AcquaDiva (domains: Microbe Habitat and Carbon and Nitrogen Cycling)(Algergawy et al., 2024);
7. LLM generated ontologies in the works by Lippolis et al. (2025a), Lippolis et al. (2025b), and Saeedizade and Blomqvist (2024).

The LLM-generated ontologies were generated by the following LLMs: *GPT-4* family (versions: *GPT-4-0613*, *GPT-4-1106*, *GPT-4-Turbo 2024-04-09*, *GPT-4-2024-05-13*), *Llama-3.1-405B-instruct*, *o1-preview*, and *DeepSeek-V3*. As shown in previous sections, in these works, a sample was evaluated, among other measures, by a panel of expert ontology engineers who found that while LLMs introduce distinctive modelling patterns, they do not necessarily underperform compared to novice human modellers. These ontologies were thus designed specifically to allow a controlled comparison between human-generated and LLM-generated artefacts using the same input requirements.

OntoEval-small. From the OntoEval corpus, we constructed a balanced subset for controlled experimentation, *OntoEval-small*. To achieve this, we employed a combination of hard and soft constraints. As hard filters, we excluded ontology models addressing only minor issues (i.e., resulting in a binary YES/NO annotation) to make sure labels were unambiguous. Additionally, we retained only the CQs rated as COMPLEX to have a more challenging evaluation. To ensure domain diversity, we limited the selection to a maximum of three CQs per project. LLM correctness served as a soft filter, meaning this filter was used as a sampling guide rather than a strict rule, such that the small dataset would have statistics similar to that of the full dataset with respect to correct and incorrect LLM predictions. We also sampled an approximately equal number of LLM-generated and human-curated ontologies and made sure that the LLM-generated ontologies were generated by different LLMs. Lastly, we selected a similar number of CQs that were modelled and not modelled. To facilitate user evaluation, each original story was condensed into a single-line summary using OpenAI’s *o1-preview* model in a few-shot setting.

Dataset statistics. Table 4.14 summarises key descriptive statistics for *OntoEval* and *OntoEval-small*. For each dataset, we report the total number of CQs and the number of correctly modelled CQs. Most CQs in OntoEval are modelled and therefore the dataset is imbalanced (1,204 YES and 189 NO). We also include the number of CQs classified as SIMPLE and COMPLEX according to the criteria outlined in Section 4.5.4.1, along with the domain coverage of the ontologies and their provenance. The “not modelled CQs” were not generated by automated negative-sampling techniques but were instead drawn from planned EU project deliverables or manually annotated by ontology experts. Furthermore, the ontologies are of different sizes, which are computed by the number of axioms in OntoMetrics (Lantow, 2016), ranging from 15 to 567 axioms (mean: 132 axioms, median: 44 axioms, STD: 164 axioms).

TABLE 4.14: Comparison of the full *OntoEval* corpus for automatic ontology evaluation and the *OntoEval-small* subset for semi-automated evaluation.

Metric	OntoEval	OntoEval-small
Total CQs	1,393	20
Modelled CQs	1,204	10
Difficulty: SIMPLE	725	0
Difficulty: COMPLEX	135	20
Domains	33	6

4.5.4.2 Methodology and experimental setup

Overall methodology. OE-Assist—structured around semi-automatic and complementary automatic setups with ontology engineers—aims both to determine whether integrating AI-assisted suggestion can improve quality assessment and to explore the extent to which LLMs can independently evaluate ontologies as standalone agents. In this section, we describe the methodology of automatic ontology evaluation and the semi-automatic method.

Automatic ontology evaluation. The automatic evaluation is designed as a label-based assessment mechanism. In this phase, a prompt is introduced to an LLM to analyse whether a CQ is adequately modelled within a given ontology and an ontology story. The evaluation is structured to classify ontology coverage without human intervention. The components of the prompt are the task description, CQs, a user story, and an ontology. The expected outputs are a binary class (YES: CQ is modelled, NO: CQ is not modelled) and a SPARQL query for CQ verification.

Semi-automatic ontology evaluation. In the semi-automatic evaluation, we investigate the extent to which LLMs can support ontology engineers in the ontology evaluation process. This phase employs the same inputs and expected outputs as the automatic approach. The LLM’s suggested label (YES or NO) and the SPARQL query are provided to the user as suggestions. Given an ontology, its ontology story, and the LLM’s suggestions, the user is responsible for determining whether each CQ has been accurately modelled within the ontology or not.

Prompt design. We propose a prompt⁴⁷ that (i) describes the ontology evaluation task, (ii) provides some examples (out of the *OntoEval-small* dataset, cf. Section 4.5.4.1) in a few-shot style, with both positive (YES) and negative (NO) cases, (iii) includes the user story, CQ, and ontology, and (iv) instructs the model to: (a) Return a Boolean label (YES or NO) indicating whether the CQ is correctly modelled, and (b) generate a SPARQL query to verify the CQ. If the model returns NO, it is also instructed to supply a partial SPARQL query highlighting the available portions of the ontology related to the CQ.

Automatic ontology evaluation We automatically assess the correctness of each ontology using a set of three different LLMs on the *OntoEval* dataset and comparing the LLM YES/NO judgments against the gold standard.

⁴⁷Due to space constraints, the full prompt specification is available on GitHub: <https://github.com/dersuchendee/OE-Assist/tree/main/prompts>.

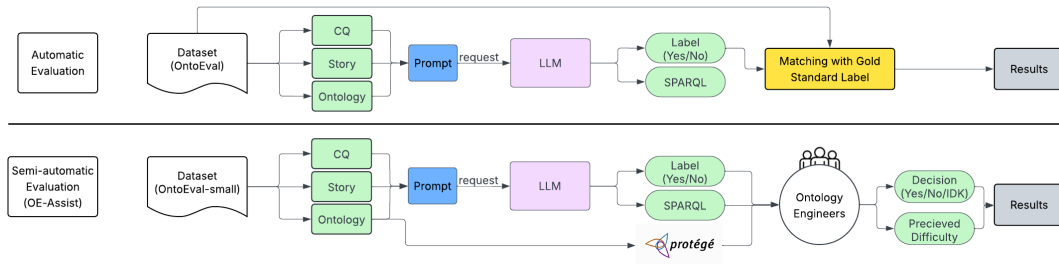


FIGURE 4.8: Overview of OE-Assist: (i) Automatic: predicted labels are compared to gold standards to compute performance metrics. (ii) Semi-automatic: Users aided by LLM suggestions assess whether a CQ is modelled.

Models under test. We run the prompt through three different LLMs: OpenAI *o1-preview*, *GPT-4o-0513*, and OpenAI *o3-mini* to represent a range of model capabilities and architectures. *o1-preview* was included as it is a state-of-the-art reasoning model with advanced capabilities for complex tasks. *GPT-4o-0513* is a widely-used foundation model with established performance across knowledge-intensive tasks. *o3-mini* was included to test whether smaller, more efficient models could perform adequately on ontology evaluation tasks. Each model is run three times, except *o1-preview* on the full OntoEval dataset and accessed via Azure API, with at least five hours between each run to avoid caching, using the default hyperparameters for *o1-preview*. Temperature and penalty are set to 0, along with keeping default hyperparameters, for others.

Answer extraction. Although the prompt requests both the SPARQL query and the binary label, for this evaluation, we parse only the label to compute Macro-F1 and Accuracy by comparing it to the gold standard.

Semi-automatic ontology evaluation This section presents the semi-automatic evaluation of ontologies. In this study, each participant was assigned 20 ontology evaluation tasks in a limited time. Half of these tasks included suggestions generated by an LLM, whereas the remaining tasks did not feature such assistance. The order of the tasks was changed between participants to mitigate the impact of learning effects on the results. The ontology stories, CQs, and the LLM-generated suggestions (if present) were shown using a basic user interface where ontology engineers could select a final assessment decision. A second screen provided access to the Protégé application with the ontology connected with the task pre-loaded.

Participants. Nineteen ontology engineers participated in this evaluation. They were recruited through academic networks and represented roles ranging from research assistants to professors and affiliated with eight different international institutions with industrial or academic experience in semantic web technologies. Participants were divided into two experience groups (*non-expert* vs. *expert*) with professors—including assistant and associate professors—classified as *experts*, and all others as *non-experts*. While we acknowledge this distinction tends to simplify expertise, we opted for professors as a conservative and externally verifiable criterion over self-reported years of experience, which could yield inconsistent groupings. Seven participants completed the evaluation in an online setting, while the remaining twelve participated in person.

LLM selection. The LLM employed to generate suggestions (label and SPARQL) was *o1-preview* as it was the best performing one in the automatic ontology evaluation phase of the experiment.

Interface. A prototype of a web-based interface was developed to present LLM suggestions and requirements along with the ontology, as illustrated in Figure 4.9. The interface was made available online for the duration of the experiment and was displayed side-by-side with Protégé to facilitate comparison. On the left, participants could browse the OWL ontology file in Protégé, while on the right they interacted with the evaluation interface. The interface displayed, for a given ontology, the current CQ and story, optional LLM suggestions, and response controls to answer if a CQ is modelled or not with the perceived difficulty on a Likert scale from 1 to 5. In this way, participants could carry out ontology evaluation in two modes: An assisted mode (with LLM suggestions) and an unassisted mode (without LLM suggestions).

Procedure. Participants completed two different sets of tasks (20 in total) under both conditions *Assisted* (10) and *Unassisted* (10) in a single session of approximately one hour. Prior to the main task, each participant completed a 10-minute brief tutorial, which included worked examples intended to familiarise them with the dual-window setup and the response workflow. Time was allotted for clarifying questions. Each condition used a dual-screen setup, Protégé and the interface, and was assigned a maximum time window of 20 minutes; CQs remaining at the end of this interval were recorded as skipped. Where participants were uncertain about a particular CQ, they could skip it by selecting “I don’t know” (IDK) via the interface. They could also skip each question if it was too difficult or if they could not understand a CQ. The expected sequence of actions was as follows: (i) participants were to open the ontology file in Protégé (without any plugins) and, after reading the CQ, the ontology story, and the suggestions (if available), (ii) they would select their answer (either “Yes”, “No”, or “IDK”) and finally, (iii) indicate how difficult they found the task using a 1–5 Likert scale, as shown in Figure 4.9a (Assisted) and Figure 4.9b (Unassisted). Upon reaching the time limit or completing all tasks, participants switched to the alternate condition and proceeded with the remaining 10 questions. The sole distinction between conditions was the availability of LLM suggestions in the Assisted setup. A think-aloud protocol was applied throughout the session, encouraging the participants to explain and comment on their actions and thoughts, to facilitate observational note-taking by the experiment observers. To conclude, participants completed a 10-item survey adapted from the System Usability Scale (SUS, Brooke et al., 1996).

In our setup, we recognise that each ontology evaluation task could have different difficulty, so we made sure each CQ appeared equally in the assisted mode and unassisted mode. Also, the order of Assisted mode and Unassisted mode, whether it appears in the first half of the experiment or the second, is changed between participants to maintain a balanced distribution.

4.5.4.3 Evaluation

Automatic ontology evaluation. For OntoEval, we report macro-averaged F1, which balances the majority (YES) and minority (NO) classes; for OntoEval-small, which is balanced, we report Accuracy and compare all LLMs with a uniform-random baseline (Acc=0.50, F1=0.42).

CQ: When is the level of a chemical substance recorded in a water body?
 Scenario: Context is about monitoring water quality by recording chemical substance levels to investigate potential links between contaminated water sources and adverse public health outcomes.
 LLM suggested answer: Yes

```
SELECT ?chemicalSubstance ?waterBody ?date
WHERE {
  ?cIR a :ChemicalLevelRecording .
  ?cIR :hasRecordingDate ?date .
  ?cIR :recordsChemical ?chemicalSubstance .
  ?cIR :recordedIn ?waterBody .
}
```

Is the CQ modelled?

Yes No I don't know

How difficult was the CQ for you to solve?
 (1: Very easy to 5: Very hard)

1 2 3 4 5

(A) Assisted version in the interface.

CQ: When is the level of a chemical substance recorded in a water body?
 Scenario: Context is about monitoring water quality by recording chemical substance levels to investigate potential links between contaminated water sources and adverse public health outcomes.

Is the CQ modelled?

Yes No I don't know

How difficult was the CQ for you to solve?
 (1: Very easy to 5: Very hard)

1 2 3 4 5

(B) Unassisted version of the interface.

FIGURE 4.9: Comparison of assisted and unassisted settings in the interface.

Semi-automatic ontology evaluation. To assess the semi-automatic evaluation, we define each user’s accuracy in setup (assisted vs not assisted) m as $\text{Acc}_{u,m} = C_{u,m} / (C_{u,m} + I_{u,m})$, where $C_{u,m}$ and $I_{u,m}$ are the counts of correct and incorrect answers (“IDK” ones are not counted here). We then compute the overall mode accuracy by averaging over N users: $\text{Acc}_{\text{setup}} = \frac{1}{N} \sum_{u=1}^N \text{Acc}_{u,\text{setup}}$, and define $\Delta\text{Accuracy} = \text{Acc}_{\text{assisted}} - \text{Acc}_{\text{unassisted}}$. We also report the standard deviation of $\text{Acc}_{u,m}$ in each setup. To further analyse the relationship between LLM assistance and task correctness, we conduct a chi-square test of independence. Since each CQ was attempted multiple times by different users across both assisted and unassisted modes, we treat each answer as an independent observation and construct a 2×2 contingency table of assistance condition (present vs. absent) and answer correctness (correct vs. incorrect).

We used paired two-tailed t-tests to compare each user’s accuracy (and separately perceived difficulty) in the assisted vs. unassisted conditions, reporting the t-statistic and p-value. We assessed the relationship between perceived difficulty and accuracy using Pearson’s correlation coefficient, testing significance at $\alpha = 0.05$. Where relevant, we computed Cohen’s d to quantify the magnitude of paired differences (e.g. in perceived difficulty or accuracy) (Becker, 2000).

Finally, to investigate the potential impact of LLMs’ correctness on user decision-making, we subdivide the reported results based on whether the LLM provides a correct suggestion or an incorrect suggestion. This approach allows us to discern how the varying degrees of LLMs’ accuracy affect the overall effectiveness of the semi-automatic ontology evaluation methodology.

4.5.4.4 Results

In this section, we present the results of the automatic and semi-automatic ontology evaluation. We start with the automatic ontology evaluation on OntoEval and then select the best performing model to be used in the semi-automatic setting, assessing the users’ performance in the various settings.

Automatic ontology evaluation. For the automatic evaluation of ontologies, as shown in Table 4.15, *o1-preview* overall outperforms the other LLMs on both datasets. The only weakness of *o1-preview* was over human-curated ontologies, resulting in a 0.05 lower score than *o3-mini*. Since all LLMs have a low standard deviation (STD) and the cost of running the experiment on *o1-preview* is high, we ran this model only once.

TABLE 4.15: Performance of LLMs in the automatic ontology evaluation, compared to the random-guess baseline. Bold shows the best score on each column.

Model	Macro-F ₁ (on OntoEval)			Accuracy (on OntoEval-small)
	EU projects,	LLM Generated,	All	All
Random baseline	0.41	0.42	0.43	0.50
GPT-4o-0513	0.34 ±0.02	0.48 ±0	0.48 ±0	0.55 ±0
o3-mini	0.51 ±0.02	0.56 ±0.01	0.58 ±0.01	0.72 ±0.02
o1-preview	0.46	0.66	0.66	0.75 ±0.05

Semi-automatic ontology evaluation. In this section, we present the results for the semi-automatic ontology evaluation, focusing on the differences in accuracy and perceived difficulty in the assisted compared to the unassisted mode. We also examine how the presence of suggestions from LLMs, whether correct or incorrect, shape these outcomes. In addition, we evaluate the relationship between perceived difficulty and correctness. Finally, we discuss the observed learning curve and the questionnaire that users filled in, highlighting how their responses inform the broader evaluation of ontologies and the role of LLMs within this process.

Accuracy in the assisted compared to the unassisted setting. Using the metrics described in Section 4.5.4.3, we calculated accuracy in both the assisted and unassisted settings. We then computed the difference (Δ Accuracy) for each condition, taking into account whether the suggestion provided was correct or not. By analysing the relationship between these variables, we assessed whether the presence of LLM suggestions had a statistically significant effect on correctness.

As accuracy scores are illustrated in Figure 4.10b, when the LLM provided correct suggestions, we observed a significant average improvement of +13% (from 70.46% to 83.18%), supported by a paired t-test ($t = 2.04$, $p = 0.047$). However, when the LLM suggestions were incorrect in the same task, performance significantly declined with an average change of -28% (from 71.93% to 43.86%), confirmed by a t-test ($t = -2.09$, $p = 0.043$). Since correct LLM suggestions occurred more frequently overall, there was a total improvement of +0.04% (0.72 to 0.76) for LLM-assisted users (see Figure 4.10b). However, this difference was not statistically significant. In addition, *o3-mini* and *o1-preview* have a similar accuracy to users' performance in both modes.

Using a chi-squared test, we found statistically significant associations for both correct ($\chi^2 = 4.12$, $p = 0.042$) and incorrect ($\chi^2 = 4.61$, $p = 0.0318$) LLM suggestions. These findings underscore that the presence of LLM input, within the experimental setup, for each CQ, indeed affects decision outcomes.

Perceived difficulty in assisted compared to unassisted mode. Participants judged the tasks as slightly less difficult when LLM suggestions were present (mean difficulty = 2.52, STD = 0.56) compared to when they were absent (mean difficulty = 2.75, STD = 0.46), where mean difficulty represents the average difficulty rating across all CQs on a Likert scale, and STD indicates how much those ratings varied from one CQ to another. This difference is small (Cohen's $d \approx 0.22$), indicating a

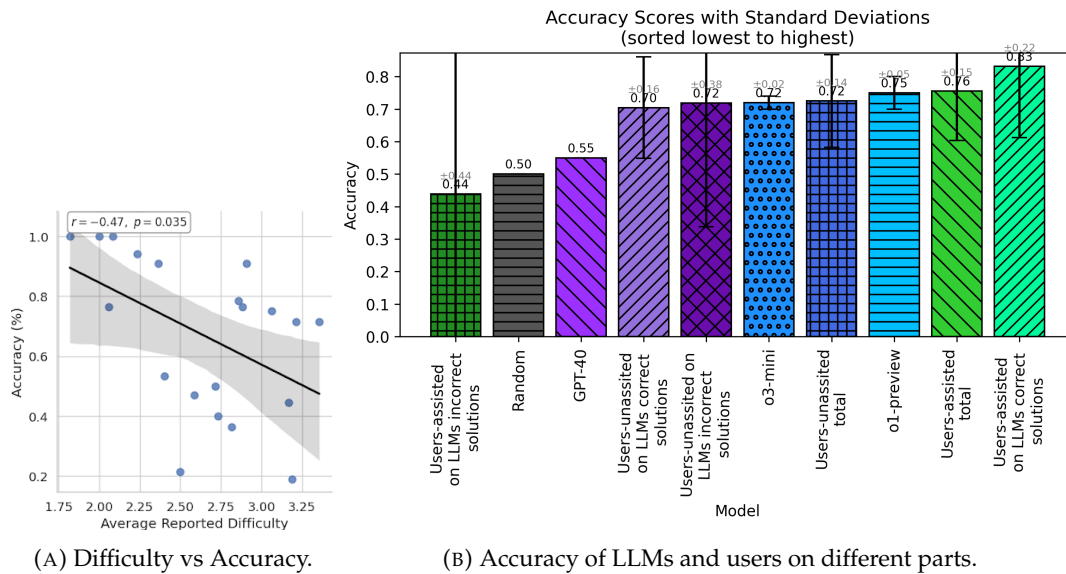


FIGURE 4.10: Results overview. a) Relationship between difficulty and accuracy: easier perceived tasks are solved more accurately; (b) Comparative accuracy of the LLMs (orange) and human evaluators (green: assisted and blue: unassisted), with user responses further stratified by LLM correctness. STD is in grey.

modest effect size; in behavioral research, d values around 0.2 typically reflect small but meaningful differences. Note that this reflects variation across CQ-level averages rather than individual tasks.

As expected, experts consistently rated tasks as easier than novices, with the average difficulty dropping by 9% for experts (from 2.67 to 2.44) and 5% for novices (from 2.84 to 2.69) when assisted. Thus, both groups experienced slightly lower perceived difficulty with suggestions, but the expert group consistently found the tasks easier overall. For what concerns “I don’t know” responses, users settled on “I don’t know” almost exclusively when they perceived the question as hard (mean difficulty=3.9 vs 2.4 for other responses). Furthermore, if the accuracy of the CQ was low, there was also a higher number of “I don’t know” responses. Instead, ontology size in terms of number of axioms does not play a role in pushing people towards being undecided ($r=0.01$, $p=0.8$): more likely, their decision is driven by their subjective difficulty towards, for instance, unclear phrasing or domain-specific CQs.

Relationship between perceived difficulty and accuracy. Figure 4.10a depicts Pearson’s correlation between perceived difficulty and accuracy in our setup. There is a moderate negative correlation ($r = -0.47$, $p = 0.035$) between perceived difficulty and accuracy, indicating that as users perceived tasks to be harder, their accuracy tended to decrease. This correlation seems to be explained by the tasks in which the LLM gave incorrect suggestions: users found those tasks harder and were also less likely to solve them correctly. Furthermore, when the tasks were analysed separately by the correctness of the LLMs’ suggestions, the difficulty–accuracy link was no longer significant (LLM correct: $r = -0.23$, $p = 0.44$; LLM incorrect: $r = +0.21$, $p = 0.69$).

Learning effect curve by expertise. As participants progressed, they rated each successive task as easier, indicating a learning effect. Participants who began in the assisted mode showed almost no change when they later switched to the unassisted

mode (-0.04 accuracy), whereas those who began unassisted improved substantially after switching to assistance (+0.17 accuracy). Expertise further modulated this pattern: non-experts gained little when assisted first, yet improved modestly without assistance (+0.06). In contrast, experts improved sharply in the unassisted block (+0.29) and only slightly with assistance (+0.06). Across all users, accuracy rose in the second half for 40% of CQs solved without LLM support but for just 15% with support. A paired-samples test approached significance ($p \approx 0.08$), and the effect size (Cohen's $d \approx 0.45$) indicates a medium learning advantage in the unassisted condition—but definitive statistical confirmation requires a larger sample.

Questionnaire.

In the post-task questionnaire, 63% of the participants strongly agreed that the LLM suggestions were useful in solving the task, 74% felt they could learn to use the system setup quickly, and, 21% found either the suggestions, the suggested label, or the SPARQL queries distracting.

4.5.4.5 Discussion

This section discusses our key findings and situates them within the broader context of automated and semi-automated ontology evaluation.

Automatic Ontology Evaluation. From our results, *GPT-4o* hovers around the random baseline, whereas *o1-preview* delivers +0.23 F1 over random on automatically generated ontologies and +0.15 on human curated ones. Similar improvement was seen for *o3-mini*, suggesting that reasoning LLMs can be helpful for ontology evaluation, especially given the steady upward trend in LLM performance. These results imply that *o1-preview* and *o3-mini* can already serve as an automated first-pass layer flagging potential errors before human review; for instance, they could accelerate curation in large, rapidly updated ontologies.

Semi-automatic ontology evaluation. Our findings show that while LLM-assisted ontology evaluation improves the accuracy by 13% when the suggestions are correct, the approach exhibits a significant decline when the user is supplied with incorrect suggestions. Overall however, these opposing effects largely cancelled each other out, resulting in no statistically significant net change in overall accuracy. Given that the *o1-preview* model's suggestions used in the experiment had a 70% accuracy, we speculate that future LLMs, with higher suggestion accuracy rates, will enable end users to realize a correspondingly greater enhancement in task performance. One may also consider other ways to mitigate the effects of wrong LLM suggestions, such as pre-checking suggested SPARQL queries against the ontology before showing the suggestion to the user, or including some kind of confidence assessment of the suggestions.

Also, LLM suggestions reduced the perceived difficulty (mean = 2.52 vs. 2.75). One downside of the assisted condition, however, is that it may hinder learning over time. The modest learning gains in unassisted settings suggest that task complexity from the user's point of view is higher without automated support (accuracy improved of 40% for CQs without LLM suggestions in the second half of the experiment vs. 15% with suggestions). This learning effect was shaped by task order: users who started with assistance received early support and thus performed equally well in the later unassisted phase as those who learned through doing. As a result, overall learning gains appear neutralised, since early assistance seems to substitute the effort that would otherwise lead to similar improvement.

Ontologies contained in OntoEval-small had different sizes. However, we found there is no correlation between the size of ontologies in terms of axiom number and

incorrect answers. However, we note that even the largest ontology used is small by real-world standards. Taken together, these results imply that LLM support seems to be valuable mainly when its precision is high and that future versions of the prototype should pair assistance with dedicated safeguards—and perhaps dedicated learning phases—to avoid over-reliance on LLMs and preserve users’ skill growth, but also that in realistic scenarios, automated suggestions could become even more valuable.

Feedback during the experiment. Looking at users’ behaviour performing the tasks revealed that when working with the assisted setting many users started the evaluation from the SPARQL queries, to find and match the named classes and properties with the ones mentioned in the CQ. A second group of users operated in reverse: first exploring the ontology through Protégé to establish a conceptual understanding, and only after that they proceeded to study the suggested SPARQL query.

These contrasting workflows suggest different methods of problem-solving: one favoring a code-first approach that uses the ontology as confirmation, and another prioritizing conceptual understanding before proceeding with CQ verification. In line with these strategies, in the questionnaire feedback, users who took the code-first path appear to have leaned heavily on the system’s inline suggestions to speed-check query fragments, an affinity that helps explain why 63% strongly agreed the suggestions were useful.

Where the two groups converged was on the final decision aid. Only 32% endorsed the binary YES/NO recommendation. Indeed, user feedback revealed challenges with the lack of labels and comments in some ontologies, unclear CQs, domain-specific content, and flawed LLM-generated hierarchies, all of which required extra effort to interpret and evaluate. These findings highlight the need for adaptable AI interfaces that support both code-first and concept-first user strategies, while emphasizing the importance of providing transparent, context-rich justifications in addition to binary recommendations.

Quality of requirements. User comments and the varying wording of the CQs show that inconsistent, under-specified, or overly domain-specific requirements impede reliable evaluation. When the wording or granularity of the questions shifted from one task to the next, participants reported uncertainty about how to assess the related ontology. Domain specificity created an additional hurdle: evaluators who lacked subject-matter expertise could judge structural soundness but not semantic correctness, showing LLM suggestions seem to not be helpful in these cases.

Limitations and risks. While our findings indicate that LLMs hold considerable promise for user-based functional ontology evaluation, future work should involve a broader group of ontology engineers and a wider range of LLMs to validate these results, since our group included only five participants with expert qualification level. Additionally, some participants had worked on the ontologies several years earlier, so even though they reported remembering little, residual familiarity could still have influenced their answers. Finally, our study highlights the need to improve the quality of the requirements and to integrate the workflow into widely used tools, such as Protégé, in order to enhance overall usability. For dataset leakage, we did not find ontologies and their CQ annotations co-located online, so direct memorisation by the LLM is unlikely. Still, because pre-training data are opaque, we cannot rule out partial leakage. Running a full automated evaluation with *o1-preview* cost approximately \$851. By contrast, a single run with *GPT-4o* cost about \$40, and *o3-mini* just \$25, so we only ran *o1-preview* once. These differences highlight a trade-off between model strength and budget that must be considered when selecting engines or planning multiple runs.

4.5.4.6 The double-edged sword of LLM assistance

In this section, we have introduced OE-Assist, the first framework using LLMs for functional ontology evaluation via CQ verification through both automatic and semi-automatic methods. We also introduced OntoEval, a benchmark of 1,393 ontology–CQ pairs across 33 domains, annotated by ontology engineers. To know to what extent can LLMs evaluate ontologies using CQ verification, in automatic experiments, the *o1-preview* model achieved 0.66 macro-F1, outperforming other LLMs. Hence, we conclude that newer LLMs perform on par with human users. To evaluate practical impact of the semi-automated methodology, we ran a controlled study with 19 ontology engineers from eight institutions. When *o1-preview*'s suggestions were correct, average user accuracy improved by 13% and perceived task difficulty decreased. However, incorrect suggestions caused a 28% drop in accuracy; resulting in the same overall accuracy as the Unassisted mode since correct suggestions outnumbered incorrect ones. We conclude that the usefulness of LLM assistance heavily depends on the suggestion accuracy, and users are easily influenced by erroneous suggestions, while still perceiving the overall approach as useful. As LLMs improve, these results bode well for reasoning-powered ontology evaluation tools.

This section explored the dual role of LLMs in ontology evaluation, assessing their capabilities as both automated evaluators and as assistive tools in a human-in-the-loop workflow. The central finding is a clear demonstration of automation bias: while LLM suggestions can improve accuracy when correct, they significantly degrade human performance when wrong, highlighting that the reliability of the underlying model is a critical variable. This dependence on model consistency and the observed brittleness of results across different model versions raise a fundamental question of scientific and practical importance: how can we systematically measure and ensure the reproducibility of LLM-aided knowledge engineering? This is the core challenge we address in the following section, with the introduction of the CoLLM framework.

4.6 Defining reproducibility in the era of LLM-aided Knowledge Engineering

An unaddressed concern when using LLMs to generate and evaluate knowledge engineering tasks is the reliability of the underlying models. In this section, we will investigate a framework to assess the consistency of KE studies.

The motivation for this research is that scientific studies have long been the primary way to disseminate research efforts, with experiment reliability serving as the backbone of scientific inquiry (McNutt, 2014). An essential indicator of the trustworthiness of scientific research is the ability to repeat its evaluation and arrive at the same conclusions, assuring that the results are reliable. Thus, *trustworthiness in scientific research* refers to the credibility of findings and adherence to rigorous standards that ensure transparency in experimental setups and modeling (Fidler and Wilcox, 2018). For an experiment to be reliable, its method and results need to be *reproducible*, *replicable*, *repeatable* and *re-evaluable* (Werner et al., 2024). While reproducibility traditionally refers to obtaining the same computational results using identical input data, methods, and conditions, its definition varies across disciplines, often leading to ambiguity (National Academies of Sciences and Policy and Global Affairs, 2019). In this context, the wider term of *Consistency* is meant to define the characteristic of

a system, process, or its results to produce the same outcome when repeated under identical or comparable conditions.

The importance of consistency in scientific research extends beyond traditional experimental sciences and into the realm of contemporary computer science. In recent years, the growing reliance on LLMs (Kasneci et al., 2023) for tasks in the knowledge engineering field has further underscored the need for dependable and consistent results. In fact, LLMs, typically deployed as black-box systems, often lack transparency in their decision-making processes. This inherent opacity, along with their sensitivity to minor changes in prompts or parameters, can result in significant variability in outputs. While these models have shown considerable potential in advancing various computational tasks, their performance can fluctuate due to factors such as model architecture, training data, hyperparameters, and deployment environments. Furthermore, this non-determinism, or stochasticity resulting from different models results in a “built-in” consistency problem that can produce different answers to the same questions. In the field of knowledge engineering, LLMs have been increasingly used to automate and improve tasks (see Section 4.2). Therefore, the challenge of achieving consistency of results has become more urgent.

The necessity of benchmarking common semantic web tasks from the perspective of LLM-driven systems has thus evolved from a desirable to an essential requirement, making research in this area more critical than ever (Alharbi et al., 2024b; Alharbi et al., 2025; Allen and Groth, 2024; Plu et al., 2024; Rebboud et al., 2024a; Tsaneva, Herwanto, and Sabou, 2024). This requirement has become all the more compelling as new, specialized LLMs are introduced, or existing ones are deprecated over time, highlighting the need for studies to generalize existing research findings. Therefore, studying consistency in the context of semantic web tasks is essential for setting new standards for LLM-based research, ultimately reinforcing trust in the reliability of LLM results.

Recent trends in LLM and KG research suggest that many methods exploit LLMs to support various tasks (Khorashadizadeh et al., 2024). However, a reliance on the use of commercial LLMs such as OpenAI’s *GPT-4* (Achiam et al., 2024), which are accessible only through black-box APIs, can present several distinct challenges (Kadour et al., 2023):

1. *Deprecation* – What if the API provider deprecates access to a particular model?
2. *Licensing* – How might the variability of licensing restrictions impact the researchers’ or organizations’ ability to use these models?
3. *Updates* – If the model parameters are updated or a new version is launched, will existing methods still perform as intended?

Beyond the challenges of model updates and deprecations, LLM outputs are also influenced by factors such as prompt sensitivity, temperature settings, and latent prompt bias. Even minor rewordings of prompts can lead to substantial output variations, making consistency difficult to ensure in automated knowledge engineering tasks (Reynolds and McDonnell, 2021). Furthermore, proprietary APIs obscure model modifications, meaning researchers may unknowingly use an updated version without documentation of underlying changes. These challenges highlight the urgency of establishing clear consistency assessment criteria for LLM-driven research. A robust *consistency* framework in the semantic web domain is therefore essential to mitigate these challenges, given the field’s dependence on enriching human-readable web data with machine-readable annotations (Berners-Lee, Fischetti, and Dertouzos, 1999).

Inconsistencies in LLM-generated knowledge structures can significantly impact real-world applications such as automated ontology alignment (Osman, Pileggi, and Yahia, 2024), knowledge graph completion (Chen et al., 2020), and ontology-based knowledge representations (Anand and Kumar, 2022). For example, if an LLM provides different entity classifications or relation extractions across multiple queries, downstream systems relying on these outputs may exhibit unpredictable behavior. In high-stakes domains such as biomedical informatics (Nazi and Peng, 2024; Clusmann et al., 2023) inconsistent knowledge representations could lead to erroneous conclusions, reinforcing the necessity of rigorous consistency evaluations in LLM-driven knowledge engineering. In this section, we investigate this matter and propose the *Consistency of Large Language Models in Knowledge Engineering* (CoLLM) framework, which investigates the consistency of results of different studies by the Semantic Web research community that exploit LLMs.

An essential aspect of the evaluation of CoLLM is the empirical validation of several existing semantic web research studies that use LLMs as a core component. Thus, to establish a solid foundation for CoLLM, we have conducted an evaluation of 59 experiments taken from five separate studies from the past two years.

Five peer-reviewed studies were selected after database searches (ACM DL, IEEE Xplore, Scopus, Web of Science, ACL Anthology; 2020–2025) using combined keywords for LLMs, knowledge graphs and ontologies, and runnable code. Inclusion required: (1) employs an LLM within its pipeline; (2) has been peer-reviewed; (3) pertains to the field of knowledge engineering; and (4) has publicly available code to run the experiment. Of the five, only one was authored by the present thesis, and it is the preliminary Ontogenia study described in Section 4.3.1; the remaining four are independent external publications included to reduce selection bias. We also include these studies’ results for completeness of the analysis. The authors of all five works were contacted to run three consistency tests on their systems, with a four-week response window. More authors were contacted but did not respond within the established timeframe. Table B.1 summarizes each study’s original characteristics, including dataset, goal, benchmark, evaluation metrics, and the criteria defining a pass or fail. Furthermore, Table B.2 outlines the experimental settings that are used as part of the new reproducibility evaluation for each study (reported on in Section 5.4.3.6), including all relevant parameters. These studies collectively illustrate the applicability of the CoLLM framework to various challenges. Although the complete methodological details of each study are not replicated here, they are available in the original publications. The following sections provide details on each study, including task descriptions, datasets, evaluation criteria, and results.

Furthermore, the LLMs considered encompass the BLOOM family (Workshop et al., 2023), Meta AI’s LLaMA family (Touvron et al., 2023b), Google’s Flan-T5 (Chung et al., 2022), the OpenAI GPT family (Achiam et al., 2024), and the Mistral family of models (Jiang et al., 2023). But beyond technical details, one first must agree on the terminology employed to define consistent results.

4.6.1 A matter of terminology

A primary concern surrounding LLM-driven knowledge engineering studies, as sometimes noted by the researchers themselves, is reproducibility (Saeedizade and Blomqvist, 2024) and its very definition with respect to other related terms, as shown in Table 4.16. However, the way the term is referred to varies in meaning across works.

For example, Tan et al. (2024) define reproducibility as obtaining a consistent response from an LLM after repeating a question, but the term has also been defined as obtaining consistent computational results using the same methods, code, and steps (National Academies of Sciences and Policy and Global Affairs, 2019). Similarly, Werner et al. (2024) define reproducibility as a term that both defines: i) performing an experiment with the same software, the same parameters and the same data but by a different experimenter; and ii) with a broader meaning, the collective actions of repeatability, replicability, reproducibility, and re-evaluation.

The term “*Reproducibility*” has therefore been defined in various ways and with different acceptations, and has often been used as a synonym for the broader notion of “*Consistency*”: in fact, the process of obtaining the same results from repeating a scientific study is referred to differently depending on how the results are obtained (Benureau and Rougier, 2017; Werner et al., 2024). The re-execution of an experiment with the same code, parameters, and data by the same team is referred to as “*Repeating*” (Collberg and Proebsting, 2016; Werner et al., 2024), while the execution of an experiment by a different team using the same setups, parameters, and data is referred to as “*Reproducing*” the results (Donoho et al., 2009; Peng, 2011; Werner et al., 2024). “*Replicating*” (Rougier et al., 2017; Werner et al., 2024) refers to experimenting independently with different software or implementations on the same data, while “*Re-evaluating*” refers to performing the experiment independently using a different dataset (Werner et al., 2024). All of these activities are often, confusingly, encompassed by the term “*Reproducibility*” (Werner et al., 2024).

For what concerns the term “*Consistency*” and its use in recent KE work on LLMs, Machado et al. (2024) mentions consistency as something that, regardless of the underlying LLM or the specific configuration used, leads to an output that should be both structurally and semantically in line with a specified standard. In Wedholm (2024), consistency is defined as the degree to which an LLM produces similar and reliable outputs when processing the same prompt across different data formats. This definition encompasses two key aspects: (i) output similarity: the generated responses should be similar regardless of the format (if the same prompt is given under identical conditions, the outputs should not vary significantly from one format to another); (ii) result reasonableness: beyond just being similar, the underlying calculations that lead to the output must be reasonable and accurate. This means that the consistency measure also evaluates whether the responses are logically sound and faithful to the intended meaning of the prompt.

The definition of consistency as something that mirrors an expected goal or functional requirement is also outlined in Tufek et al. (2024), where consistency is defined as to whether a semantic artifact faithfully represents the information and relationships expected according to its governing rules or requirements. Even more widely, Uceda-Sosa et al. (2024) refer to Nguyen (2008) by defining consistency as a situation in which a knowledge base does not contain contradictions. Verifying consistency is crucial in scientific research (Djulgovic and Hozo, 2014; Goodman, Fanelli, and Ioannidis, 2016), as it not only reinforces confidence in the findings but also establishes a solid foundation for further advancements (Pineau et al., 2021; National Academies of Sciences and Policy and Global Affairs, 2019; Werner et al., 2024).

In this section, we follow the definitions outlined in Werner et al. (2024) and choose to use the term “*Consistency*”, which defines whether a system, process, or result exhibits a consistent behaviour (i.e., no contradictions) over multiple trials.

In the context of natural sciences, consistency is often classified into two distinct categories (Collins et al., 2017): “*Quantitative Consistency*”, which refers to the ability

to replicate the same scores/values through repeated experiments; and “*Qualitative Consistency*”, which seeks to achieve outcomes/findings consistent with those reported in prior studies. Given that LLMs are frequently accessed as black-box models via APIs and exhibit non-deterministic behaviour (Werner et al., 2024), we argue that Qualitative Consistency is particularly critical for KE tasks. The outputs of LLMs are influenced by numerous parameters, with significant variations depending on whether the models are open- or closed-source. Closed-source models are often deprecated, updated without transparency, or replaced with minimal notice. Furthermore, LLM outputs are inherently variable, influenced by random seeds, temperature, and other hyperparameters, which raises questions about the reliability of the findings reported in scientific publications (Kaddour et al., 2023).

Only one study to date has evaluated LLM consistency across diverse KE tasks. Frey et al. (2024b) designed six synthetic tasks to track variability over 2023 across GPT-3.5, GPT-4, and Claude on RDF Turtle and SPARQL within LLM-KG-Bench, showing that capabilities shift across versions and that LLMs lack consistency, but their tasks do not cover the breadth of KE and real KE experiments can be costly to run at scale.

In parallel, OAEI⁴⁸ provides year-over-year benchmarks for ontology matching, documenting version drift even on fixed datasets. Recent KG construction pipelines modularize the LLM as a reranker over fixed candidate sets, enabling isolated, repeatable measurement over time; for example, Zhang and Serban (2025) validate an LLM reranker with internal KG evidence and external web signals and report standard ranking metrics that support drift tracking. KonTest (Rajan, Soremekun, and Chattopadhyay, 2024) probes intra-model factual consistency via KG-grounded metamorphic testing, finding that model choice affects consistency outcomes.

Broader ML and NLP evidence underscores instability. Training with different random seeds can shift accuracy by over 1% on CIFAR-10 (Jordan, 2023), motivating multi-run reporting and toolkits such as Robustness Gym (Goel et al., 2021). API-served LLMs may change under stable names; Chen et al. Chen, Zaharia, and Zou, 2024 observed substantial behavior shifts between March and June 2023 on the same tasks. AIRer (Zeng et al., 2025) shows that minor model updates can break end-to-end reproducibility in data-analysis workflows even with fixed prompts and settings. A recent position piece proposes a six-tier evaluation ladder from repeatability to replaceability for deployment contexts (Tian et al., 2025).

These gaps call for a practical, study-level framework for KE.

4.6.2 The CoLLM Framework

The CoLLM framework investigates the consistency of LLM-based KE efforts by proposing three empirical tests that address the following challenges:

1. A lack of reproducibility in LLM-generated results due to factors such as stochasticity or non-determinism in the models themselves, variations due to sampling, differences in API settings, or even undocumented changes in models by their developers;
2. the rapid evolution of LLMs (Minaee et al., 2024) which can result in the release of new versions or entirely new models (and consequently the deprecation of previous versions) that may not necessarily support the findings of earlier studies;

⁴⁸<https://oaei.ontologymatching.org/>

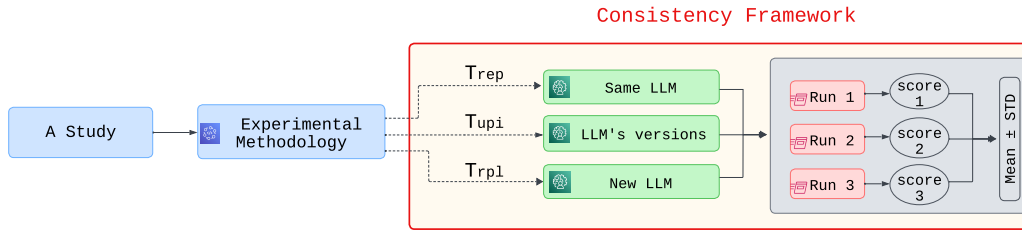


FIGURE 4.11: The CoLLM framework.

3. A model-specific bias or dependence, where many studies in KE rely on, or assume the use of, a specific LLM, leading to results that may be tightly coupled to the capabilities (or limitations) of that particular model. This phenomenon raises concerns about whether such methodologies are truly generalizable or just fine-tuned to maximise performance with a specific LLM;
4. The emerging nature of LLM-powered KE approaches that lack systematic frameworks for evaluating whether or not such approaches can yield consistent and robust results.

4.6.2.1 Preliminaries

We use terms such as “Study”, “Alternative/Distinct LLM” and “Updates/Versions of LLM” throughout this work. Here is a brief description of these concepts in our work:

Study: A knowledge engineering practice or method that uses an LLM as a component in its pipeline.

Alternative LLM: An LLM is considered an alternative model when it differs fundamentally from another in aspects such as architecture (e.g., dense vs. mixture-of-experts), training paradigm (e.g., supervised vs. RLHF), new types of training data, or core capabilities (e.g., text-only vs. multimodal). Identifying these differences can be challenging when dealing with closed-source LLMs. In such cases, we rely on the publisher’s identification of a new model without having clear information about the actual changes.

New Version of an LLM: It refers to an update of an existing LLM that retains its core architecture and training methodology while incorporating improvements in efficiency, scale, or output quality. Similar to identifying an alternative model, determining a specific update in closed-source LLMs is difficult. Instead, we rely on the producer’s classification of model versions. Examples of versions and distinct LLMs are given in Table 4.17.

4.6.2.2 CoLLM test formulation and metrics

CoLLM consists of three empirically grounded tests that constitute a complete evaluation, as depicted in Fig 4.11. The first test, the “LLM Repeatability Test” (T_{rep}) evaluates whether the results of previous studies are reproducible using the same LLM and settings. Subsequently, the “LLM Update Impact Test” (T_{upi}) investigates the extent to which updates to an LLM affect the results. Finally, the “LLM Replacement Test” (T_{rpl}) examines whether replacing the chosen LLM with another will affect the

results, thereby assessing if the findings are specific to a particular LLM or can be generalized across different ones.

TABLE 4.17: Examples of LLMs across distinct models and different versions. The first column represents the company that provided the LLM, the second column lists distinct LLMs, and the remaining columns provide examples of different versions of these models.

Provider	LLMs	Versions			
OpenAI	GPT-3.5	GPT-3.5-613	GPT-3.5-1106	GPT3.5-Turbo-0125	GPT-3.5-turbo-16k (0613)
	GPT-4	GPT-4-613	GPT-4-1106	GPT-4o	GPT-4 Turbo
	o1-preview	o1-preview	-	-	-
	ADA	text-embedding-ada-002	-	-	-
Meta AI	Llama with ~7B parameters	Llama-7B	Llama2-7B	Llama3.1-8B	Llama-3-8B
	Llama with ~13B parameters	Llama-13B	Llama2-13B	-	-
	Llama with ~70B parameters	Llama-3-70B	Llama-3.1-70B	Llama-3.3-70B	-
Mistral AI	Mistral with 7B parameters	Mistral 7B	-	-	-
	Mistral with 24B parameters	Mistral Large 24.02	Mistral Large 24.07	-	-

Given m studies, and a set of n results, $\mathcal{R} = \{r_1, \dots, r_n\}$, the CoLLM $\mathcal{T}_k$ test of Study $_k$ can be represented as:

$$r_i = E_{\text{setup}}([\text{LLM}_i, \text{Version}_i, \text{Prompt}_i]; \text{Study}_k), i \in [1, n], k \in [1, m]$$

where i represents an individual run; Version_i denotes the version(s) of the LLMs (LLM_i) employed in run i of a consistency test of Study $_k$; and Prompt_i is the set of prompts utilized in Study $_k$. The experimental setup E_{setup} is used in the study's evaluation that generates results r_i , which reflect the evaluation metrics used in the study. The results from n runs are stored in $\mathcal{R}$, and subsequently used to compute the mean μ and standard deviation σ . The mean is estimated by $\mu = \frac{1}{n} \sum_{i=1}^n r_i$ and the standard deviation is defined as $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_i - \mu)^2}$.

To facilitate comparison, let r_{orig} represent the reported score from the original scientific Study using the defined experimental setup E_{setup} . Performance Drift (PD) is the percentage quantifying the difference in a study's evaluation metrics (e.g., accuracy, F1 score) and reflects whether the update led to an improvement, decline, or consistency in performance. The PD formulation is defined as the ratio between the difference in score (between the original and evaluated score) and the original score, and is written as $PD_{\mathcal{T}} = \frac{\mu - r_{\text{orig}}}{r_{\text{orig}}}$, where a positive $PD_{\mathcal{T}}$ indicates an improvement in performance following the update, and a negative $PD_{\mathcal{T}}$ signals a decrease in performance. If $PD_{\mathcal{T}}$ is close to zero, it suggests minimal or no change in performance.

4.6.2.3 CoLLM Tests

The three CoLLM tests: *LLM Repeatability*, *LLM Update Impact*, and *LLM Replacement*, are used to evaluate the consistency of results across different experimental conditions.

LLM Repeatability Test ($\mathcal{T}_{\text{rep}}$): It investigates whether the results of prior scientific studies can be reliably reproduced when the experiments are re-run with identical settings. Experiments are conducted with the same LLM, version, parameters, prompts, and other relevant configurations, to assess the consistency of the outputs with respect to prior results. The emphasis here is on *reproducibility*, a foundational principle of scientific inquiry, ensuring that findings can be

verified and built upon in subsequent research. Accordingly, the choice of LLM, Version and Prompt will be the same as those used in the original Study with the tests executed n times to obtain $PD_{\mathcal{T}_{rep}}$ and $\mu_{rep} \pm \sigma_{rep}$ to support the $\mathcal{T}_{rep}$ analysis.

LLM Update Impact Test ($\mathcal{T}_{upi}$): It evaluates whether updates to the LLM yield significantly different results to those of earlier studies. By evaluating different versions of an LLM, it is possible to identify how changes in architecture, training data, or model parameters impact the outputs. In this test, the LLM should be the same as that used in $\mathcal{T}_{rep}$ (LLM_{Orig}); however, the version of that LLM, Version _{j} , should be different, where j represents the version number of the LLM used in E_{setup} for Study. Again, tests will be executed n times to obtain $PD_{\mathcal{T}_{upi}}$ and $\mu_{upi} \pm \sigma_{upi}$ to support the $\mathcal{T}_{upi}$ analysis.

LLM Replacement Test ($\mathcal{T}_{rpl}$): It investigates whether alternative LLMs can obtain similar performance to that observed in previous studies, or whether certain research findings are inherently dependent on a particular LLM. In this test, the only variable that differs is the LLM itself (i.e. LLM_{Alt} replaces LLM_{Orig}), which is chosen based on some rationale, such as being newer; however, the default hyperparameters of the alternative LLM may replace those used in the original work. The prompts and other configurations are held constant, remaining identical to those of the original study. This test facilitates the identification of alternative LLMs for a work that can replicate the results of prior studies, broadening the scope of tools and methodologies available for research. As before, tests will be executed n times to obtain $PD_{\mathcal{T}_{rpl}}$ and $\mu_{rpl} \pm \sigma_{rpl}$ to support the $\mathcal{T}_{rpl}$ analysis.

4.6.3 Evaluation setup

Table B.1 summarizes each study's original characteristics, including dataset, goal, benchmark, evaluation metrics, and the criteria defining a pass or fail. Furthermore, Table B.2 of the Appendix outlines the experimental settings that are used as part of the new reproducibility evaluation for each study (reported on in Section 5.4.3.6), including all relevant parameters. These studies collectively illustrate the applicability of the CoLLM framework to various challenges. Although the complete methodological details of each study are not replicated here, they are available in the original publications. The following sections provide details on each study, including task descriptions, datasets, evaluation criteria, and results.

4.6.3.1 Studies

Study 1 on Ontology Generation using Task Decomposition. As already described in Section 4.2.2, this study seeks to assess the capability of different LLMs for the task of generating ontologies using CQs and ontology stories (Saeedizade and Blomqvist, 2024), by exploring the effects of using various prompting techniques to instruct LLMs when generating ontologies. The primary research question posed by this study is: *To what extent can LLMs create an ontology that meets the requirements of an ontology story?* **Datasets:** To evaluate the performance, materials from a Semantic Web course were used with three ontology stories and 15 CQs in each. In the original case study, the students' performance was used as the baseline of the experiment when comparing performance. However, for purposes of this study, the scores of the generated models are compared with the highest score of this case study (which are

used as the baseline). **Evaluation:** The evaluation metric used in the study measures the proportion of the CQs that are correctly modeled in the resulting ontology by ignoring “minor issues”. For a competency question CQ_i and an ontology O , CQ_i is correctly modeled if O contain all necessary components to write a SPARQL query to answer CQ_i . A “minor issue” in this context is where only one property required to answer CQ_i is missing. **Results:** The experimental results showed that using a sub-task decomposed prompting (referred to as $CQ_{by}CQ$ by the study’s authors) with GPT-4 performs similarly to that of a novice ontology engineer taking the Semantic Web course.

Study 2 on Ontology Generation with Metacognitive Prompting. This case study proposed *Ontogenia* (§4.3.1) as a framework to assess the ability of using LLMs for generating OWL ontologies (Lippolis et al., 2024), starting from a gold standard of competency questions and a selection of ODPs (Gangemi and Presutti, 2009). The primary research question posed by this study is: *Can we automatically generate ontology drafts with LLMs from requirements and what is their assessment?* **Datasets:** We employed the Animal Wildlife ontology for use as a gold standard, and its associated CQs that are also available. Furthermore, we selected the relevant ODPs from the reference repository.⁴⁹ **Evaluation:** To evaluate *Ontogenia*, quantitative metrics taken from *OntoMetrics* (Lantow, 2016) and *OOPS!* (Poveda-Villalón, Gómez-Pérez, and Suárez-Figueroa, 2014) were employed together with a qualitative assessment by an ontology engineer. However, for purposes of this study, the qualitative assessment will not be taken into account, only the quantitative one. **Results:** Experimental results coupled with the qualitative analysis confirmed that MP and ODPs favored a richer formalization and the usage of a wider set of different axioms.

Study 3 on Ontology Learning. This study investigates the use of LLMs for Ontology Learning (OL) (Babaei Giglou, D’Souza, and Auer, 2023; Babaei Giglou, D’Souza, and Auer, 2024; Babaei Giglou et al., 2024). The *LLMs4OL* paradigm Babaei Giglou, D’Souza, and Auer, 2023 was proposed as a paradigm shift in OL in response to the emergence of LLMs, and the following research question was explored: *Can LLMs effectively apply their language pattern capturing capability to OL, which involves automatically extracting and structuring knowledge from natural language text?* This paradigm consists of three tasks based on the discovery of different ontology primitives (Maedche and Staab, 2001): *Task A - Term Typing*, aimed at discovering a generalized type for a lexical term; *Task B - Taxonomy Discovery*, where a taxonomic hierarchy between pairs of types is discovered; and finally, *Task C - Non-Taxonomic Relation Extraction*, the discovery of non-taxonomic semantic hierarchical relations between types. **Datasets:** The *LLMs4OL* paradigm tests were conducted using various ontological knowledge domains such as lexico-semantic knowledge in WordNet Miller, 1995, geographical knowledge in GeoNames *GeoNames geographical database 2023*, biomedical knowledge in UMLS Babaei Giglou, D’Souza, and Auer, 2024, and web content type representations in schema.org Patel-Schneider, 2014. However, for the reproducibility tests (reported in Section 5.4.3.6), only the biomedical knowledge sources in UMLS were targeted. **Evaluation:** The evaluation of *LLMs4OL* included the use of 11 state-of-the-art LLMs from the three main LLM architectures (encoder, decoder, and encoder-decoder) using a variant zero-shot prompting. The metrics used for evaluating the tasks included a MAP@1 metric (for *Task A*) and F1-scores (for *Task B* and *Task C*). **Results:** Although the experimental results for OL using zero-shot

⁴⁹<http://www.ontologydesignpatterns.org/>

prompts showed promise, they also suggested that task-specific fine-tuning may be required for the approach to be a practically viable solution. This was supported by later results that demonstrated that the fine-tuning of the Flan-T5 (Chung et al., 2022) LLM generated more credible outcomes.

Study 4 on Ontology Matching. This case study explored the *LLMs4OM* framework for the task of Ontology Matching (OM), i.e. the process of identifying correspondences between concepts (classes, properties, or individuals) in different ontologies (Euzenat, Shvaiko, et al., 2007). OM involves the use of concept mapping techniques to map ontology elements between source $C_{source} \in O_{source}$ and target $C_{target} \in O_{target}$ ontologies. Formally, for any $C_s \in C_{source}$, the task is to identify possible $C_t \in C_{target}$ such that $(C_s, C_t, S_{C_s \equiv C_t})$, where $S \in [0, 1]$ represents the likelihood of equivalence $C_s \equiv C_t$ Euzenat et al., 2011. The *LLMs4OM* framework (Giglou et al., 2024) uses two modules for retrieval and matching with enhanced zero-shot prompting across three types of ontology representations: concept; concept-parent; and concept-children. **Datasets:** Despite the large number of candidate datasets available from the *Ontology Alignment Evaluation Initiative* repository⁵⁰, datasets from six different tracks (with an overall of 20 tasks/datasets) were used to evaluate *LLMs4OM* in the original study Giglou et al., 2024. However, for the purpose of conducting the reproducibility tests (reported in Section 5.4.3.6), a subset of the originally selected datasets was used here. This subset corresponds to datasets taken from the following tracks: Anatomy (Mouse-Human task) Dragisic et al., 2017, Biodiversity and Ecology (biodiversity - Algaei-Zooobenthos tasks) Karam et al., 2020, Disease and Phenotype (phenotype - HP-MP task) Harrow et al., 2017, Biomedical Machine Learning (Bio-ML - OMIM-ORDO disease task) He et al., 2022, and finally, Material Sciences and Engineering (MSE - MI-EMMO task) Nas and Hushka, 2023. **Evaluation:** Ontology alignments were generated using the *LLMs4OM* framework with seven state-of-the-art LLMs across four retriever models and evaluated using recall for the retriever models, and precision, recall, and F1-scores for other evaluations. Note, however, that only F1-scores are generated here and reported in Section 5.4.3.6). **Results:** The extensive experimentation results confirmed that *LLMs4OM* with the right choice of LLM and retriever models can achieve a superior performance even surpassing the state-of-the-art OM models in 9 out of 20 datasets of the study.

Study 5 on Retrofitting Competency Questions. This case study proposed *RETROFIT-CQ* (Alharbi et al., 2024a; Alharbi et al., 2024c; Alharbi et al., 2024d) as an approach that *retrofits*, or re-generates missing or lost CQs by leveraging explicit knowledge from existing ontologies, represented by (s, p, o) statements. The approach utilises in-context prompts to guide parametric models such as LLMs in generating relevant CQs. The primary research question posed by this study was: *To what extent can 'usable' CQs be retrofitted from existing ontologies that are representative of the scope and tasks for which the ontology was designed?* **Datasets:** For evaluating *RETROFIT-CQ*, ontologies were selected from two main datasets: CORAL (Fernández-Izquierdo, Poveda-Villalón, and García-Castro, 2019) and the CQs Dataset (Jedrzej et al., 2020). These datasets provided diverse ontologies alongside their CQs to test the approach under various conditions. **Evaluation:** The evaluation of *RETROFIT-CQ* involved running numerous evaluations using several different settings:

1. the default configuration of LLMs (Alharbi et al., 2024a);

⁵⁰<https://oaei.ontologymatching.org/>

2. different creativity (temperature) parameters to observe output variability (Alharbi et al., 2024c; Alharbi et al., 2024d);
3. varying levels of prompt specificity (Alharbi et al., 2024c; Alharbi et al., 2024d).

These settings were applied to both closed-source LLMs (Alharbi et al., 2024a; Alharbi et al., 2024c) and open-source LLMs (Alharbi et al., 2024d). The generated CQs were compared against the original CQs associated with the ontologies (taken from the datasets) using Sentence-BERT (SBERT) Reimers and Gurevych, 2019, and both *precision* and *recall* metrics used to report performance. **Results:** Experimental results confirmed that *RETROFIT-CQ* effectively generates CQs from existing ontologies. The performance metrics indicated a strong alignment between the baseline and the *RETROFIT-CQ*-generated CQs, supporting the utility of this approach for generating representative CQs and thus enhancing ontology usability.

4.6.3.2 Experimental methodology

The CoLLM framework provides an approach to evaluate the consistency of LLM-driven methods through three different tests: $\mathcal{T}_{\text{rep}}$, $\mathcal{T}_{\text{upi}}$, and $\mathcal{T}_{\text{rpl}}$ (Figure 4.11), using the following methodological setup:

1. *Input Selection*: the input corresponds to a research study in which the use of one or more LLMs feature as a main component;
2. *Experimental Methodology*: the methodology used by that study should be clearly described, together with the source code to facilitate implementation/execution;
3. *Repetition*: a minimum of $n = 3$ iterations of E_{setup} are conducted. Using three iterations of the E_{setup} setup strikes a balance between noticing changes in the results and cost-efficiency, allowing enough variation in outcomes while minimizing computational and environmental impact. This number of iterations provides reliable results whilst avoiding unnecessary resource consumption for each test $\mathcal{T}$;
4. *LLM Selection*: the choice of LLMs should be aligned to the description of the related test (in Section 4.6.2.3), such that: for $\mathcal{T}_{\text{rep}}$, the same LLM (*family*⁵¹ and *version*) as that used in the original study should be chosen; for $\mathcal{T}_{\text{upi}}$, all alternate versions of the LLM to that used in the original study should ideally be chosen;⁵² finally, for $\mathcal{T}_{\text{rpl}}$, a different LLM to that used in the original study should be used.

Each study (summarised in Table B.1) varies in the choice of its original LLM, which affects the subsequent choice of LLM discussed in Section 5.4.3.6.

For example, given the original use of LLM in the *Retrofit-CQ* study (Alharbi et al., 2024a), the following LLMs were selected: $\mathcal{T}_{\text{rep}}$: GPT-4-0613; $\mathcal{T}_{\text{upi}}$: the currently available versions of the GPT-3.5-turbo family of LLMs (i.e. gpt-3.5-1106

⁵¹A *family* of LLMs are those that typically originate from the same source, but that may be released with different versions reflecting incremental changes over time.

⁵²It is possible that a previous version of some LLM may become deprecated and thus no longer available (for example, gpt-3.5-turbo-0613 and gpt-3.5-turbo-0301 as set out in <https://platform.openai.com/docs/deprecations>). In this case, it is advisable that a minimum requirement of at least two versions should be assumed to observe the changes in scores related to different versions of a specific LLM.

and gpt-3.5-0125) are used; and for $\mathcal{T}_{rpl}$: the new OpenAI model (O1 preview) is used, as it differs from the LLMs reported in that study (Alharbi et al., 2024d; Alharbi et al., 2024a).

4.6.4 Results

We conduct three consistency tests on five KE tasks (results are summarised in Table 4.18). First, we assess the Repeatability Test ($\mathcal{T}_{rep}$) by evaluating whether previously reported results are reproducible. Second, we perform the LLM Update Impact Test ($\mathcal{T}_{upi}$) to examine the effects of updates to LLMs on the KE tasks. Finally, we evaluate the LLM Replacement Test ($\mathcal{T}_{rpl}$) to determine whether LLMs used in KE tasks can be replaced with an alternative model or not. In these tests, we only focus on the qualitative reproducibility of the original work meaning the criteria for pass and fail is reproducibility of findings/outcomes.

TABLE 4.18: Summary of the CoLLM results across five studies. Each study considered several datasets, and the numbers show how many times the consistency tests passed or failed based on Tables[B.3 4.19 B.4 B.5 B.6]. Each pass or fail is the result of 3 runs. Scores in bold represent the results of all tests over 59 experiments.

Selected Studies	LLM Repeatability Test		LLM Update Impact Test		LLM Replacement Test		Study Result	
	Passes	Fails	Passes	Fails	Passes	Fails	Passes	Fails
Study 1	1	0	3	0	1	0	100%	0.0%
Study 2	0	1	0	2	1	1	20.0%	80.0%
Study 3	2	1	5	4	3	0	66.7%	33.3%
Study 4	5	0	6	2	5	0	88.9%	11.1%
Study 5	4	0	8	0	4	0	100%	0.0%
Test Results	85.7%	14.3%	73.3%	26.6%	93.3%	6.7%	81.4%	18.6%

TABLE 4.19: Results of Study 2: CoLLM tests using evaluation metrics (reported as $\mu \pm \sigma$ across three runs) and $PD_{\mathcal{T}}$ for ontology generation using the *Ontogenia method* ($\star$ indicates Ontogenia.). The results are given as percentages based on the highest score obtained across all tests. Results in green and red refer to the $PD_{\mathcal{T}}$ with respect to the original experiment, and results representing failed tests appear in italics.

Avg Score ($PD_{\mathcal{T}}$)	MP+Patterns $\star$	MP no Patterns	Patterns no MP	No MP no Patterns
<i>Reported Score (r_{Org})</i>				
LLM: GPT-4-1106				
Axioms	99.1	53.7	100	41.1
Class	80.7	53.8	65.3	100
Properties	100	68.7	50	30.7
<i>LLM Repeatability Test ($\mathcal{T}_{rep}$)</i>				
LLM: GPT-4-1106				
Axioms	65.05 $\pm$ 4.32 (-36.44%)	56.04 $\pm$ 8.16 (90%)	94.19 $\pm$ 22.40 (-87.3%)	54.9 $\pm$ 7.25 (29.1%)
Class	17.3 $\pm$ 3.02 (+17.6%)	61.98 $\pm$ 2.49 (16%)	57.03 $\pm$ 2.03 (11.7%)	47.91 $\pm$ 1.34 (-10%)
Properties	12.3 $\pm$ 1.34 (+53.75%)	100 $\pm$ 2.05 (39.09%)	76.88 $\pm$ 2.36 (-5.38%)	76.88 $\pm$ 1.25 (53.75%)
<i>LLM Update Impact Test ($\mathcal{T}_{upi}$)</i>				
LLM: GPT-4-1025				
Axioms	75.11 $\pm$ 15.9 (-26%)	70.51 $\pm$ 12.7 (27%)	78.36 $\pm$ 24.6 (-79%)	85.04 $\pm$ 1.6 (100%)
Class	38.02 $\pm$ 2. (-52%)	57.03 $\pm$ 4.5 (7%)	31.56 $\pm$ 2.5 (-85%)	49.43 $\pm$ 0.0 (-7%)
Properties	60 $\pm$ 0.4 (-0.4)	79.38 $\pm$ 5.0 (15%)	64.38 $\pm$ 1.8 (-86%)	75 $\pm$ 0.8 (50%)
LLM: GPT-4-turbo-2024-04-09				
Axioms	63.57 $\pm$ 11.5 (-37.8%)	49.95 $\pm$ 17.33 (-10%)	82.45 $\pm$ 37.23 (-20.11%)	54.7 $\pm$ 18.23 (29.11%)
Class	60.03 $\pm$ 2.13 (-23.8%)	58.28 $\pm$ 4.26 (9.5%)	35.47 $\pm$ 5.02 (-45.12%)	45.89 $\pm$ 0.58 (-14.19%)
Properties	91.9 $\pm$ 7.23 (-11.9)	71.89 $\pm$ 4.55 (0%)	65.81 $\pm$ 2.08 (-22.54%)	72.35 $\pm$ 0.58 (38.3%)
<i>LLM Replacement Test ($\mathcal{T}_{rpl}$)</i>				
LLM: o1-preview LLM				
Axioms	100 $\pm$ 64.5 (-3%)	39.29 $\pm$ 14.52 (-28%)	100 $\pm$ 81.9 (-3%)	38.16 $\pm$ 15.30 (-18%)
Class	95.06 $\pm$ 4.32 (19%)	39.16 $\pm$ 4.08 (-26%)	76.05 $\pm$ 8.52 (0.17%)	26.62 $\pm$ 1.41 (-50%)
Properties	75 $\pm$ 2.16 (-25%)	35 $\pm$ 0.47 (-49%)	50 $\pm$ 4.14 (-38%)	16.25 $\pm$ 2.10 (-67%)
LLM: Llama-3.1-405-Instruct-Turbo LLM				
Axioms	88.74 $\pm$ 22.45 (-13%)	55.81 $\pm$ 6.11 (4%)	71.64 $\pm$ 34.1 (-30%)	49.44 $\pm$ 22.56 (16%)
Class	73.38 $\pm$ 4.7 (-8%)	76.32 $\pm$ 1.53 (43%)	100 $\pm$ 14.26 (54%)	89.74 $\pm$ 16.17 (68%)
Properties	72.5 $\pm$ 1.25 (-27%)	87.50 $\pm$ 2.52 (27%)	93.75 $\pm$ 7.51 (15%)	35.63 $\pm$ 4.36 (-28%)

4.6.4.1 LLM Test results

Results of LLM Repeatability Test ($\mathcal{T}_{rep}$) An examination of the $\mathcal{T}_{rep}$ results in Tables B.3, 4.19, B.4, B.5, and B.6 together with the summary of results in Table 4.18, reveals that over 81% of the tests passed. This suggests that, in most cases (12 out of 14 experiments), reproducing the original work with the same settings yields consistent results. Studies 1, 4, and 5 demonstrate improvements in their scores, suggesting that they can be reproduced successfully.

Interestingly, in study 1, STD is still relatively high compared to other studies but significantly lower than what was reported in the original study. Despite the lowest score on this test, the findings remain consistent with those of the original study. In Research Study 5's *Video Game* task, a lower precision is observed, which could be attributed to the generation of more expressive questions through paraphrasing or new modelling perspectives compared to the benchmark. However, the results still support the same conclusion as those in the original study, by generating most of the benchmark competency questions (as suggested by the high recall). In contrast, for Research Study 2, the stochasticity or non-determinism behaviour of LLMs had high standard deviations and fluctuating performance drift meaning high variability in this study between the runs and compared to the reported scores caused a different setting to emerge as the optimal configuration. Similarly, the score of Task C in Research Study 3 led to a different conclusion, causing the replication to fail for those two cases.

In addressing the question of challenge (1), “*What is the extent of the stochasticity or non-determinism of the results of previous scientific studies?*”, the majority of the experiments were found to pass, suggesting that any stochasticity or non-determinism inherent in the LLMs used had little impact on the overall results and conclusions of the different studies evaluated. However, in some studies over some tasks, high *sigma* and PD values were observed, compared to those of the original studies. These could affect the conclusions by turning the study into a negative result or lead to better findings for those studies, and thus warrant further investigation.

Results of LLM Update Impact Test ($\mathcal{T}_{upi}$) By analyzing the $\mathcal{T}_{upi}$ results in Tables B.3, 4.19, B.4, B.5, and B.6, together with the summary of results in Table 4.18, we observe that 73.3% of the tests (22 out of 30) passed. This represents the lowest pass rate compared to other tests, indicating that in approximately one out of four cases, the results and findings differ from those of the original studies. These failures occurred across three different LLMs, suggesting that the discrepancies are not confined to any specific LLM.

Consistent with the previous section, PD was positive in studies 1 and 5 across all datasets, thus supporting the same conclusions as those found in the original studies. It should be noted that in two instances for study 5, negative PD scores were observed for recall (96.61% and 97.85%), compared to 99.1% in the original study. However, these differences remain small within the context of the study, thereby still validating the findings of the main study with another version of GPT-4.

This contrasts with study 2, which yielded different conclusions with two versions, failing the test for this case. Likewise, for studies 3 and 4, four scores were less than -5% and -4% respectively (a threshold related to each study that changes the conclusion of the original study, mentioned in Table 4.18), leading to conclusions that differ from those in the original studies.

To address the question posed in the challenge (2), “*To what extent do updates of LLMs affect the results of the studies?*”, later versions of LLMs can affect the results in approximately one out of four cases. Despite the test passing in 73.3% of the cases, there remains a significant risk, indicating that newer versions of an LLM should be re-evaluated before being adopted in research studies or industry applications.

Results of LLM Replacement Test ($\mathcal{T}_{rpl}$) Similar to the results of the two previous tests, our findings reveal that 93.3% of experiments passed the LLM Replacement test. The only study that yielded a different conclusion was study 2 with *Llama-3.1-405B-Instruct*; however, the test passed when using *OpenAI o1-preview*. This suggests that the KE studies investigated are not model-dependent and that they can yield similar outcomes when alternative LLMs are used.

Although the test succeeded in all but one case, the performance metrics differed from those in the original studies. The results for Study 1 were broadly stable, exhibiting very few differences. However, the differences between the ($\mathcal{T}_{rpl}$) tests and the original results were much greater for studies 2 and 5 when using *o1-preview*, suggesting that further investigation is needed when replacing the LLM for these two studies.

For example, in Study 5, *o1-preview* maintains a higher recall, generating most, if not all, of the competency questions from the ground truth. However, this is associated with a drop in the precision score due to the generation of a larger number of competency questions compared to the ground truth. The need for further investigation is discussed in Section 4.6.5. In contrast, the results for studies 3 and 4

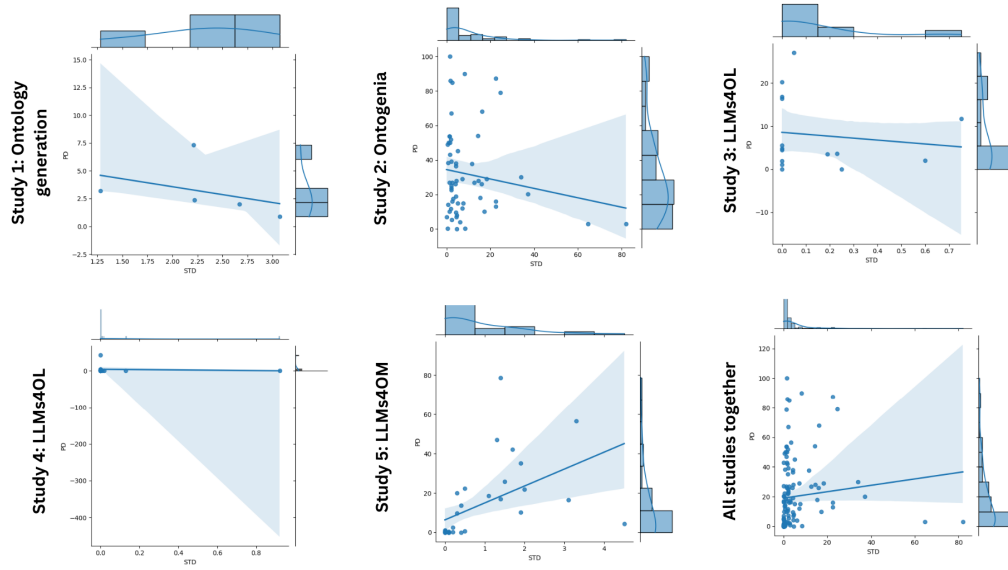


FIGURE 4.12: Bar chart for STD (x-axis), Performance Drift (PD) (y-axis), and Correlation of them. All Studies and all experiments combined as one.

yielded higher results for five of the experiments (exceeding +4% and +5% for these studies), with minimal σ . This suggests that these models consistently achieve a better performance, indicating high confidence and potential suitability for extending these studies in future.

In addressing the question posed in challenge (3), “*Can an alternative LLM achieve a similar performance to those used in previous studies, or are certain studies dependent on a particular LLM?*”, in 93% of cases, alternative LLMs were found to achieve a similar performance to that in the original studies, suggesting that these studies are not dependent on a specific LLM. However, the scores can vary significantly, suggesting that there could be substantial differences in the overall behaviour with some LLMs resulting in substantial improvements or decreases in scores.

4.6.4.2 Performance analysis

Finally, we discuss the performance drift and spread of results (represented by σ) of the experiments reported in Tables B.3 - B.6, and illustrated in Figure 4.12:

Performance Drift of Experiments. The results do not suggest that any trends exist for the PD with newer (regarding versions) or alternative LLMs; i.e. newer versions or alternative LLM models used within the tests do not lead definitively to lower or higher scores; on the contrary, the scores fluctuate significantly; meaning newer LLMs in KE tasks do not necessarily improve the scores. In some studies, PD was found to be considerably high (whether positive or negative). Although the original hypotheses or conclusions still held for the different studies with these new results, the impact of using different LLMs demonstrates the importance of further evaluation in these studies as new LLMs emerge.

Standard Deviations of Experiments. Similar to PD, there is no clear or obvious trend in the observed standard deviation (σ); this means newer LLMs do not

have lower non-determinism behaviour. However, it should be noted that the number of repeated runs for each test, n was low, and thus has an impact on the standard deviation. Ideally more runs should be performed; however this is challenging due to the time and costs associated with multiple queries with several LLMs. Ideally, this should be addressed in future research on KE where LLMs are used, to facilitate the generation of statistically significant results, as their use may account for any minor variance in results due to the non-determinism of the LLM models.

Correlation between Standard Deviation and Performance Drift. In some of the studies analyzed, the results exhibited a high performance drift (PD) and a high standard deviation (σ). However, as shown in Figure 4.12, there is no significant or predictable correlation between these results. This suggests that neither a high performance drift necessarily leads to a high standard deviation nor a high standard deviation implies a high performance drift; similarly, a low performance drift does not guarantee a low standard deviation, and a low standard deviation does not ensure a low performance drift.

4.6.5 Discussion

The consistency analysis resulting from the three proposed tests shows that (i) most prior KE results replicate, (ii) roughly one in four findings is brittle to LLM version bumps, and (iii) in nine out of ten cases it is possible to swap the original model for a modern one without invalidating the conclusions—though metric drift is substantial. In particular, for ($\mathcal{T}_{rep}$), 81.4% of the studies passed, for $\mathcal{T}_{upi}$ 73.3% of the studies passed, while for $\mathcal{T}_{rpl}$ 93.3% of experiments passed the LLM Replacement test. Furthermore, the magnitude and what the KE study analyzes—not just the direction—of drift matters. For example, in the case of Study 5, where competency questions were generated, a recall drift of 3 points is tolerable because in the study, after the LLM proposes its list of CQs, an expert reviews them: the drift is operationally harmless as the additional workload by the expert is negligible. However, in real-time ontology matching pipelines, even a -4 percentage points “pass” is operationally costly.

Overall, by applying the three tests on the studies in this work, we found some tests failed on some studies. We try to speculate the factors that may have driven these inconsistencies. An overview can be seen in Table B.7. Below, we focus on the motivations for Ontogenia.

Motivation for observed inconsistencies in the Ontogenia preliminary study.

The observed inconsistencies in the results can be attributed to several factors, including the lack of multiple test repetitions in the original work, which fails to account for stochastic variability inherent in LLMs, especially for what concerns ontology conceptualization, as evidenced by the high standard deviations in certain metrics. Additionally, the reliance on both qualitative and quantitative evaluations highlights the need for construct validity, namely more robust methods that can effectively capture the richness of ontology formalization, as the metrics employed in this work, due to the quantity of generated files, do not adequately reflect the qualitative depth required. Variations in results are also influenced by the extent of memorized ontological information within the models, which can bias the performance towards well-documented concepts. Therefore, for a more definitive evaluation of the claimed improvements with the *MP+Patterns* configuration, further research addressed these limitations by implementing new evaluation metrics that can be used

alongside or in lieu of the qualitative ones, and better examining the influence of memorization biases on performance (see the improved methodology in the second Ontogenia study (§4.3.2)). As we have seen from the results of the Ontogenia study, some qualitative evaluation metrics stemming from human observation could not be replaced by quantitative ones. This was the rationale for developing studies on ontology evaluation, which however carry some pitfalls (§4.5). As a result, the results of CoLLM need to always be considered with respect to the comprehensiveness of the metrics that have been employed.

4.6.6 Establishing consistency standards for LLM-aided knowledge engineering

In this section, we have presented the CoLLM framework, the first systematic approach to assessing consistency in LLM-aided knowledge engineering, addressing a critical gap as the field transitions from deterministic algorithms to stochastic generative methods. Through comprehensive evaluation of 59 experiments across five representative studies, our analysis reveals both encouraging resilience and concerning brittleness in current LLM-based approaches. The framework's three-tier assessment demonstrates that while the majority of LLM-aided knowledge engineering findings exhibit reasonable consistency—with 81.4% overall reproducibility across repeatability, update impact, and model replacement tests, significant variability persists. The finding that roughly one in four studies fails to maintain consistency under version updates establishes a concrete metric for the field's current reliability challenges. Conversely, the high success rate of model replacement tests (93.3%) suggests that methodological contributions are largely model-agnostic, providing confidence in the generalizability of research findings beyond specific commercial offerings.

These results show the dual nature of LLM consistency challenges: while stochastic variability appears manageable for most knowledge engineering tasks, the rapid evolution of model versions introduces systematic risks that demand proactive management. The substantial performance drift observed across studies underscores that consistency assessment cannot be treated as a one-time validation but requires continuous monitoring as models evolve. The CoLLM framework thus establishes both measurement capabilities and methodological imperatives for the field. It demonstrates that systematic consistency evaluation is not only feasible but essential for maintaining scientific rigor in an era of rapid model development. The framework's application reveals, in addition to those discussed in Sections 4.3-4.4.1, specific failure patterns that practitioners can anticipate and mitigate, which concern for instance the choice of LLMs and the design of experiments.

Moving forward, the consistency standards established through CoLLM point toward a mature paradigm for LLM-aided knowledge engineering, one that embraces the powerful capabilities of generative models while maintaining the reproducibility and reliability standards essential for scientific progress. As the field continues to evolve, these consistency assessment practices will prove crucial for building trust in AI-assisted knowledge engineering and ensuring that the transformative potential of LLMs translates into dependable, reproducible scientific advancement. Still, these are not the only ethical concerns. In the following section, we comprehensively address ethical limitations of the studies in this chapter, which remains a critical concern on top of performance.

4.7 Ethical considerations

In the previous sections, we have seen that employing LLMs for KE raises a set of ethical and methodological concerns that cut across the studies in this chapter and, more broadly, the emerging practice of LLM-assisted ontology work. While the reported results indicate that current closed models can produce draft ontologies that meet many requirements and even outperform novice modellers under certain prompting strategies, the same studies consistently surface variability, non-trivial modelling errors with mistakes concerning taxonomy and axioms, and the need for multi-dimensional, expert-led evaluation. These studies show that these systems must be positioned as assistive, not autonomous, agents in KE workflows. For instance, in the Ontogenia methodology (§4.3), the best-performing model (*o1-preview*) produced outputs of sufficient quality to serve as starting points for ontology engineers, yet we also observed recurrent issues such as erroneous domain/range restrictions, faulty inverse axioms, and superfluous elements, showing impossibility to rely on a single evaluation metric and the need to include ontology engineers in the evaluation to get insight on error patterns.

Automation bias and human-in-the-loop are topics of the utmost urgency in this context. When examining LLMs as assistant evaluators (§4.5), LLM suggestions improved mean accuracy and reduced perceived task difficulty when correct, but incorrect suggestions decreased accuracy substantially, yielding no net gain over unassisted work. This exposes a central risk for hybrid KE tooling: efficiency benefits come with a real possibility of misguided reliance when model guidance is wrong, which motivates transparency about suggestion confidence and careful interface design to sustain human judgment.

Even when overall performance is promising, therefore, the studies report quality variability and recurring pitfalls. In safety-critical or compliance-heavy domains, such errors can silently propagate into downstream decision logic if draft ontologies are adopted without expert curation. We therefore advocate expert qualitative analysis layered on top of structural checks and functional testing via CQ verification; this recommendation carries an ethical imperative to avoid overstating model competence and to maintain accountability with clear human sign-off before deployment.

A second ethical axis involves model openness, transparency, and reproducibility. The studies in Section 4.2.2.4 explicitly compare closed and open models (e.g., OpenAI *o1-preview* versus *Llama-3.1-405B-Instruct* and *GPT-4* variants), and document that conclusions depend on model choice and version. This creates fairness and replicability concerns: closed models can be stronger on some tasks yet remain non-auditable and subject to version drift; open models are inspectable and locally deployable but may underperform without careful configuration. Ethical reporting thus requires declaring the exact engines and versions used, together with their differing effects on outcomes.

In a related sense, there is an economic and access dimension that must be kept into account when using LLMs. Running the studies in this section raised equity and sustainability questions about who can afford to reproduce or extend such experiments at scale and whether cost pressures might bias evaluations toward weaker but cheaper engines.

Data governance and leakage are also important. In Section 4.2.2.4 and 4.4.1, we did not find the ontologies and their CQ annotations co-located online and thus consider direct memorisation unlikely; still, because the pre-training corpora of closed models are opaque, they cannot exclude partial leakage. Moreover, the experiments

involved routing material (ontologies, stories, queries) to third-party endpoints (e.g., via an Azure API for *o1-preview*), which heightens obligations around confidentiality when KE artefacts encode sensitive organisational knowledge or proprietary vocabularies. Practitioners should therefore prefer local execution with open models when requirements are sensitive; otherwise, they should apply redaction, contractual safeguards, and strict logging minimisation for cloud inference. Across the studies considered, the risk of data leakage depends primarily on whether the artefacts evaluated by the LLMs (ontologies, competency questions, SPARQL-OWL queries, and gold solutions) are publicly available and co-located in training corpora. In particular, in the preliminary Ontogenia study (§4.3.1), the dataset is derived from a recognised benchmark of competency questions, with a specific use case from the African Wildlife Ontology comprising fourteen CQs; thus the material comes from an established benchmark which, despite gold standard ontologies being published separately from the CQs, implies potential—but unverified—public exposure in pre-training data.

In the ontology generation study of Section 4.3.2, the selected ontologies from the “SemanticWebCourse” had their stories and CQs published online since 2008–2009, but not the OWL solutions, and for all cases the CQs/stories are published separately from the ontology models; this reduces the likelihood that models could have memorised paired inputs and gold answers. To avoid LLMs to enter the created dataset as output as future training corpus, the data has been published online with a password.

In the ontology evaluation study of Section 4.5, we explicitly searched for co-location of ontologies with their CQ annotations and report that we did not find ontologies and their CQ annotations co-located online, which makes direct memorisation unlikely, while acknowledging that opaque pre-training data prevent ruling out partial leakage altogether. In addition, in the study, we have drawn on multiple sources, with requirements and models aggregated from prior work, again reducing the chance that any single bundle of question and answers was ingested as-is by an LLM.

In the CoLLM study of Section 4.6.2.3, for Study 1 (ontology generation from stories/CQs), the authors note that the stories and competency questions have been publicly available since 2009, but the ontology solutions were not available, limiting straightforward next-token learning of the correct outputs. For Study 2, which is Ontogenia, the dataset has been public since 2020, yet gold outputs are stored separately from CQs and gold ontologies; observed variability in Jaccard overlap (0%–90.91%) and degraded precision for *o1-preview* relative to predecessors argue against trivial memorisation of prior runs. For Study 3 (ontology learning with UMLS), the data are *not* publicly available, so direct training-time access is unlikely and the leakage risk is minimal. For Study 4, datasets have long been public, so some exposure in pre-training is possible; nevertheless, reproduced scores do not exceed the original baselines, suggesting any leakage did not inflate performance. For Study 5, ontologies and their CQs have been public since 2019, but LLM-generated questions and outputs are stored in files separate from the ontologies; together with model-behaviour changes across versions, this setup again weakens the case for simple memorisation effects.

This section also shows the importance of requirements: long, memory-heavy prompts can distract models and degrade output quality, and that token/output limits force decomposition strategies and iterative merging—each with its own failure modes (e.g., overlap, drift). Because functional adequacy is evaluated against CQs and stories of varying complexity and clarity, poor or ambiguous requirements

can yield misleadingly positive or negative assessments. Ethically robust pipelines should control context length, make requirement quality explicit, and report decomposition strategies alongside results.

In sum, the ethical posture that emerges from this chapter about the integration of LLMs in ontology engineering is one of guarded integration. Closed models may currently offer stronger baseline performance for KE tasks, but at the cost of opacity, expense, and potential data-governance risks; open models enhance transparency and local control, but often require additional engineering to reach comparable quality. Across both, the studies converge on three safeguards: maintain human oversight to counter automation bias; adopt multi-dimensional, requirement-aware evaluation (structural, functional, and expert); and document models, versions, costs, and data flows to support reproducibility and responsible reuse.

4.8 Reframing ontology engineering in the LLM era

This chapter has traced the methodological evolution of pattern extraction and ontology engineering from its traditional roots to the contemporary landscape of LLM-augmented knowledge engineering. The main research question set out to ask was: *How has automation in knowledge engineering evolved with LLMs, and how can LLMs be systematically integrated into a reproducible, human-in-the-loop pipeline for KE automation of metaphor processing?*

In this chapter, we saw that automation has evolved from manual methods to LLM-assisted workflows where models act as reactors that on premise draft axioms, instantiate patterns, and evaluate outputs, with humans retaining oversight.

By reviewing OEMs, where once the process demanded extensive manual effort at every stage, we now witness a hybrid paradigm where LLMs serve as cognitive amplifiers, capable of generating initial drafts, suggesting patterns, and evaluating outputs, yet still requiring human expertise for validation and refinement.

The studies presented in the chapter demonstrate both the promise and the perils of this transformation. The successful generation of ontologies that meet 85-90% of competency questions using techniques like Ontogenia confirms that LLMs can indeed produce outputs comparable to novice ontology engineers. The domain-agnostic nature of these results further validates the generalizability of these approaches. Yet the persistent challenges—superfluous elements, erroneous domain/range specifications, and susceptibility to incorrect suggestions—underscore that we have achieved augmentation, not replacement, of human expertise.

Our review and contribution across multiple LLM-based ontology engineering approaches reveals three systematic failure patterns that practitioners must anticipate and mitigate. First, superfluous element generation emerged as a persistent issue, with Llama-generated ontologies producing approximately 40% superfluous classes and properties, manifesting as redundant concepts. Second, domain/range specification errors represented the most common critical flaw according to OOPS!. Third, automation bias in human-AI collaboration demonstrated measurable impacts: while correct LLM suggestions improved user accuracy by 13%, incorrect suggestions caused a 28% accuracy drop, with statistically significant effects that ultimately yielded no net performance gain over unassisted work.

The CoLLM framework's findings on reproducibility (81.4% repeatability, 73.3% update robustness, 93.3% model replaceability) provide crucial empirical grounding for the field's maturation. These results suggest that while LLM-based knowledge engineering exhibits reasonable consistency, the one-in-four brittleness to version

updates demands careful versioning practices and continuous re-evaluation as models evolve.

Three methodological innovations stand out as particularly significant contributions. First, the shift from monolithic to modular ontology generation, exemplified by the refinement of XD principles for the LLM era, enables more tractable and testable knowledge structures. Second, the emergence of multi-dimensional evaluation frameworks—combining structural metrics, functional testing via CQ verification, and expert qualitative assessment—provides the necessary rigor for assessing LLM-generated artifacts. Third, the development of hybrid human-AI workflows, as demonstrated in the OE-Assist studies, charts a pragmatic path where LLM suggestions enhance rather than supplant expert judgment. Perhaps most importantly, the consistent finding that context engineering matters as much as model selection suggests that the next frontier lies not in larger models but in more sophisticated workflow designs.

The ethical considerations, however, remind us that the transformation of knowledge engineering is not merely technical but fundamentally sociotechnical. The cost disparities between models, the opacity of closed systems, and the variability in outputs all pose challenges that extend beyond algorithmic optimization to questions of access, accountability, and trust.

The answer to our initial research question is thus neither automation only nor continued manual effort, but rather a careful combination of human expertise and machine capability, guided by multidimensional evaluation. This is the new methodological reality of pattern extraction and knowledge engineering in the era of large language models. This hybrid approach is particularly crucial for metaphor research, which faces fundamental scale and coverage problems—diverse conceptual mappings, sparse data across domains, and long-tail rare cases that traditional methods cannot address efficiently.

Therefore, practitioners implementing LLMs for ontology engineering should adopt a three-tier mitigation strategy: (1) Implement systematic post-processing to identify and merge semantically equivalent elements, using tools like OOPS! for automated structural validation followed by expert review of flagged domain/range inconsistencies; (2) Design human-in-the-loop workflows that present LLM suggestions with guardrails, such as explicit confidence indicators, and require active validation rather than passive acceptance, particularly for critical structural decisions. These workflows are needed to form comprehensive evaluation setups; and (3) Establish cost-performance optimization by starting with mid-tier models for initial drafts, then selectively upgrading to more expensive models only for complex requirements where the 13% accuracy difference justifies the cost increase. This evidence-based approach acknowledges that LLMs serve best as drafting tools, or amplifiers inside a methodologically constrained workflow, requiring systematic validation rather than autonomous ontology generators.

With these methodological foundations established, the next chapter operationalizes this approach for metaphor: domain-scoped datasets exposing graded metaphoricity, a Blending Ontology encoding analogical structure, and a LAG-based pipeline integrating analysis and reasoning.

System	Input	Goal / Output
OntoChat (Zhang et al., 2024; Zhao et al., 2024)	Stakeholder user stories expressed in natural language.	Generate and iteratively refine CQs in a conversational loop to support participatory ontology engineering.
NeOn-GPT (Fathallah et al., 2024)	Structured requirements plus ontology metadata (e.g., classes, properties, labels, naming).	Produce requirements-grounded CQs using NeOn-inspired prompt templates, then filter and test them; map validated requirements and CQs to a draft ontology module, yielding ontology artefacts (axioms and vocabulary) aligned with the specified requirements.
AgOCQs (Antia and Keet, 2023)	Domain-specific corpora such as scientific articles.	Generate well-formed, answerable CQs that reflect intended ontology semantics, using lexical, syntactic, and semantic filters.
Pan et al. 2024 (RAG for CQs) (Pan et al., 2024)	Retrieved passages from a corpus of scientific papers (chunked and embedded).	Generate context-aware CQs conditioned on retrieved evidence; compute a semantic similarity score against reference questions.
Rebboud et al. 2024 (Rebboud et al., 2024b)	An existing ontology (classes, relations, labels).	Generate CQs for the given ontology and evaluate CQ quality via similarity to ground-truth questions (SentenceBERT).
RETROFIT-CQs (Alharbi et al., 2024a)	Ontology triples extracted from the target knowledge base.	Generate candidate CQs directly from triples.
Di Nuzzo et al. 2024 (Di Nuzzo et al., 2024)	Domain standards (e.g., BREEAM) together with entity-and community-level summaries retrieved from a knowledge graph.	Automatically generate CQs that operationalize the standards, using KG summaries to condition generation.

TABLE 4.2: CQ generation systems: overview.

TABLE 4.16: Summary of key terms related to reproducibility and consistency, with definitions and supporting sources.

Term	Definition	Synonyms/Notes	Sources
Reproducibility	Obtaining consistent computational results using the same methods, code, and steps, or obtaining a consistent response from an LLM after repeating a question.	Often used interchangeably with consistency.	Tan et al., 2024; National Academies of Sciences and Policy and Global Affairs, 2019; Werner et al., 2024; Benureau and Rougier, 2017
Repeating	Re-execution of an experiment by the same team with identical software, parameters, and data.	Used as a synonym of reproducibility.	Collberg and Proebsting, 2016; Werner et al., 2024
Reproducing	Execution of an experiment by a different team using the same setups, parameters, and data.	Used as a synonym of reproducibility.	Donoho et al., 2009; Peng, 2011; Werner et al., 2024
Replicating	Independently performing an experiment with different software or implementations on the same dataset.	Used as a synonym of reproducibility.	Rougier et al., 2017; Werner et al., 2024
Reevaluating	Performing an experiment independently using a different dataset.	Used as a synonym of reproducibility.	Werner et al., 2024
Consistency	The degree to which a system, process, or result exhibits the consistent behavior over multiple trials.	Sometimes used interchangeably with reproducibility; can be split into quantitative and qualitative aspects.	Machado et al., 2024; Wedholm, 2024; Tufek et al., 2024; Uceda-Sosa et al., 2024; Nguyen, 2008
Quantitative Consistency	The ability to replicate the same numerical scores or values through repeated experiments.	Focuses on numerical reproducibility.	Collins et al., 2017
Qualitative Consistency	Achieving outcomes that are logically sound and align with findings reported in prior studies.	Particularly critical for KE tasks given LLMs' non-deterministic behavior.	Collins et al., 2017

Chapter 5

Operationalizing metaphor: from phenomenon to method

Metaphor, as explored in previous chapters, represents one of the most fundamental yet computationally elusive aspects of human cognition. While theoretical frameworks, especially when combined, provide rich accounts of how metaphors structure thought and language, translating these theories into computational methods remains a challenge.

The operationalization of metaphor, intended as building a bridge from concepts to measurement¹, requires addressing three interconnected challenges that were highlighted in the previous chapters. First, we lack the data necessary for systematic computational analysis, particularly in specialized domains where metaphors carry high stakes (such as medical communication). Second, existing ontological representations fail to capture the dynamic, compositional nature of metaphorical meaning-making, which goes beyond the mere recognition of the two domains involved in a metaphor with their mapping and their justification, and the sensitivity to contextual, plastic and pragmatic cues. Third, there is a plausibility challenge: how do we discover and generate metaphors that are not just a combination of concepts, but rather plausible ones. A neurosymbolic approach can provide flexible but stable structures, such as knowledge patterns (§2.1.3), to solve this challenge and allow the applicability of metaphors in real-world systems.

Conversely, we have seen now is the time to understand how symbolic systems, which operate over a finite vocabulary, can collaborate with LLMs, which foreground the interpretive and adaptive aspects of the world, evident in metaphor and, as shown in the previous chapter, in ontology generation and evaluation. This chapter presents an integrated framework that addresses these challenges through complementary components, each building upon the previous to create a complete operationalization pipeline.

In particular, the implementation presented here instantiates the earlier commitments in working resources and methods: datasets that model uncertainty and domain specificity, an ontology that encodes analogical mapping over frames and allows the generative power of metaphor, and a neurosymbolic pipeline that executes and evaluates the process across detection, understanding, reasoning across different modalities and domains.

The RQ we aim to address is: *How can metaphor studies be operationalized?* We also address three sub-questions.

RQ1 Data. What characteristics of metaphor usage emerge in scientific/medical discourse, and how do current LLMs handle graded metaphoricality under human uncertainty?

RQ2 Representation. Can a formal ontology capture the non-conservative nature of metaphor while aligning with perspectivisation in context?

RQ3 Operational analysis. What pipeline can reliably recover the blending property across metaphorical text and images?

We begin by answering the first RQ in Section 5.1 by establishing the necessary empirical groundwork through two novel datasets: the Medical Metaphor Corpus (MMC) and the Balanced Conceptual Metaphor Testing Dataset (BCMTD). These resources address the critical absence of fine-grained, domain-specific metaphor annotations, providing both the breadth needed for robust evaluation and the specialized

¹Throughout this work, we use the definition of F. Moretti (2013), which means operationalizing as turning theoretical notions into explicit procedures so they can be counted, compared, and tested. However, operational thinking had its original roots in physics, then continued to be used in social sciences and, with Moretti, into the humanities at large (Goertz, 2006).

focus required for high-stakes applications. By capturing metaphoricity as a spectrum rather than a binary property, these datasets reveal the nature of metaphoricity perception that computational systems must navigate.

In Section 5.2.1 we address the second RQ through the Blending Ontology, a formal framework that computationally instantiates Conceptual Blending Theory. This ontology provides the structure necessary to represent metaphors as dynamic conceptual integrations, introducing formal axiomatizations for concepts like blendable frames, blending spaces, and the new blended entities that result from metaphorical processes. In this section, our attempts to computationally identify the blending property, that is, the core conceptual bridge that enables metaphorical mappings, reveal fundamental limitations in current approaches. Through experiments with frame-based, embedding-based, and LLM-based methods, we demonstrate that neither structured knowledge bases nor neural approaches alone can reliably extract this fundamental component. These failures motivate the need for a hybrid approach that combines symbolic reasoning with neural flexibility.

Building on lessons learned from our previous attempts, we answer the third RQ by introducing in Section 5.4 the LAG framework, which integrates LLMs with structured knowledge representations to achieve robust metaphor processing. By combining the pattern recognition capabilities of neural models with the interpretability of graph-based representations, LAG successfully operationalizes metaphor detection and understanding across textual, visual, and multimodal contexts, significantly outperforming existing baselines.

The empirical evidence for this argument, established through our systematic evaluation of individual approaches, directly motivates the LAG framework’s architectural design.

5.1 The need for domain-specific datasets

As shown in Section 3.1.1.1, one of the main problems pointed out by current metaphor studies is that there is no theoretical standardization in metaphor studies, but at the same time, it is not possible to move forward in metaphor processing without dedicated, domain-specific fine-grained datasets, which are lacking. Furthermore, the previous chapter has shown that work on the collaboration between KE and LLMs needs task-specific datasets. Thus, we have a need for scientific metaphor datasets, whose development we propose in this chapter.

5.1.1 The Medical Metaphor Corpus (MMC)

MMC is a 792-item dataset that aggregates medical metaphors from nine sources across heterogeneous channels from scholarly articles to social media, and enriches each sentence with crowd-validated ratings of perceived metaphoricity.

5.1.1.1 Motivation

Metaphor is pervasive even in the most technical genres; yet, precisely these high-stakes domains expose severe blind spots in current language technologies. In previous chapters, we have shown (§ 3.1.1.3) that despite advances in LLMs, figurative language understanding remains brittle and these systems tend to struggle with higher-order relations such as metaphors, and these problems tend to be most evident in domain-specific metaphors rather than in generic ones.

To the best of our knowledge, no publicly available resource offers fine-grained annotations of domain-specific metaphors in scientific writing. Likewise, downstream applications such as clinical decision-support and patient-centric text generation lack training data that distinguishes conventional metaphors from perceived novel ones. Therefore, despite the growing interest in figurative language, the community still lacks a curated resource that brings together medical metaphors across the full spectrum of scientific discourse and lay communication. Larger, balanced samples of metaphors are particularly called for, since current studies use small and non-balanced examples, and among the various domains, no corpora of medical metaphors exist (Liu, 2025). This lack of resource hinders quantitative and multi-modal analyses of figurative language. Furthermore, many studies converge on the fact metaphoricality is a range rather than a binary label; however, this continuum has not been usually annotated in metaphor datasets (Julich-Warpakowski and Jensen, 2023; Bisang et al., 2006; Dunn, 2010; Gibbs, 2015). Existing evaluation frameworks currently treat all human annotations equally, despite varying levels of annotator consensus. A model that fails to detect metaphors with high human agreement represents a more serious limitation than one that struggles only with cases where human annotators themselves show substantial disagreement.

MMC is a step towards closing this gap. By gathering 792 manually verified items from 9 complementary channels (peer-reviewed articles, specialist handbooks, news media, patient-oriented social platforms, and open crowdsourcing) it offers an unprecedented, discourse-aware panorama of how illness and care are framed metaphorically. By providing the first discourse-aware, domain-balanced resource of this kind, we enable systematic testing of LLMs' domain-specific metaphor competence, support contrastive studies between expert and lay framing, and lay empirical foundations for applications ranging from claim mining to the generation of patient-friendly explanations. In this context, and thanks to the annotation scheme, we also propose the use of confidence-weighted evaluation metrics that prioritize items with stronger human consensus while de-emphasizing controversial cases.

MMC is indeed suited for different use cases, including: (i) contrastive analyses of metaphor usage between clinicians, patients, and the public; (ii) benchmarking figurative-language models beyond generic conceptual metaphor detection and understanding; and (iii) downstream tasks such as claim mining, stance detection, and the generation of patient-friendly metaphors. In short, MMC provides the empirical footing necessary to advance both cognitive-linguistic theory and practical NLP applications in the medical and health domains, fields in which metaphors have been theoretically long studied, but never empirically and on a large-scale.

5.1.1.2 Data collection

The data collection followed a systematic approach to identify annotated scientific metaphors in existing literature. We conducted searches using keywords: "scientific metaphor", "medical metaphor", "biological metaphor", "conceptual metaphor AND science", across major scholarly and linguistic databases (Linguistics and Language Behavior Abstracts, MLA International Bibliography) and computational linguistics venues (ACL Anthology). Sources were included if they: (1) contained explicit sentence-level metaphor annotations in medical or biological domains; (2) provided source-target mappings following CMT; and (3) offered sufficient context for metaphoricality assessment. The literature search returned nine primary sources spanning different discourse types (academic literature, news media, social platforms, patient narratives, crowdsourced data). Where sources were not tabular, we

digitized them through manual transcription or, in most of the cases, extracted sentences verbatim, preserving original wording and punctuation.

TABLE 5.1: Primary sources for MMC divided by channel (Chan) and number of metaphors (N). For Channels, *Lit* concerns academic literature, *News* the news domain, *SoMe* social media, *Crowd* stands for crowdsourced.

Source	Chan	N
Medical metaphors (Tongereren, 1997)	Lit	455
UK News (Camus, 2009)	News	19
UK News (Kaikarytė, 2020)	News	145
Patient forum (Semino, Demjén, and Demmen, 2018)	SoMe	27
Cancer stories (Feralda, Duwila, and Setyowati, 2022)	SoMe	35
Medical metaphors (Cheded, Liu, and Hopkinson, 2022)	News	35
Cancer narratives (Gibbs Jr and Franks, 2002)	SoMe	50
Diabetes on Twitter (Sinnenberg et al., 2018)	SoMe	40
Metamia (Metamia, n.d.)	Crowd	16
Total		792

An example is the MEDICINE IS WAR conceptual metaphor in scientific and clinical discourse: “Immunotherapies hijack our natural defences to combat cancer” (Kaikarytė, 2020). This example supports our working definition of scientific metaphor: a metaphor whose target concerns scientific entities, processes, or practices and whose use occurs in scholarly, clinical, or science–communication contexts.

Literature. In *Metaphors in Medical Texts* (Tongereren, 1997), the author analyzes how conceptual metaphors are used in medicine by analyzing scientific articles. In the text, the author does not provide explicit annotations for the subset that is classified as isolated metaphors, that is, metaphors that have become highly conventionalized within the field. However, we include these instances in MMC and retrospectively annotate their source and target domains basing on the descriptive text to maintain consistency across the corpus.

News outlets. Camus (2009) analyzes 19 cancer conceptual metaphors found in *The Guardian*. Kaikarytė (2020) analyzes conceptual metaphors in popular medical discourse: 145 from popular UK news outlets such as *The BBC*, *The Guardian*, or *The Daily Mail*. The scope of these works is to analyze how diseases are talked about in popular discourses from the point of view of CMT. Cheded, Liu, and Hopkinson (2022) analyze 35 medical metaphors for understanding the consumption of preventative healthcare in a news setting.

Social media. Many works focus on social media discourse of illnesses. In fact, people anonymously can share more freely what they think, and it is a different perspective than one of both “institutional” outlets such as news or scientific literature. In this way, it is possible to get a glimpse into what the patient really experiences.

While proposing an integrated approach to metaphor and framing, Semino et al. (2017) select for presentation 27 metaphors from a UK-based online forum for people with cancer and identify 35 metaphors apt for discussion about the use of conceptual metaphor in cancer patient stories.

Fereralda, Duwila, and Setyowati (2022) present five metaphors in popular discourse online derived from the FORCE forum. Finally, Sinnenberg et al. (2018) collect 40 metaphors of diabetes from Twitter.

Interviews. Gibbs Jr and Franks (2002) collect 50 conceptual metaphors from interviews with 6 middle-class women in recovery from cancer.

Crowdsourced data. Metaphors can also be collected from crowdsourced data. In particular, Metamia² is a website where users can freely submit metaphors and analogies found online. They can specify the source and the target of the trope, along with author and link of the source. The website is not structured by themes but rather has a keyword-based search option. To collect medical metaphors, the following keywords: “cell”, “disease”, “illness”, “cancer”, “biology” were searched from the website to filter from inputs by users. Furthermore, these results were manually filtered by an expert according to their actual presence of a metaphor and the presence of a good example (namely, not a paraphrase). As a result of this process, we obtain 16 annotated metaphors.

5.1.1.3 Data processing

Each source underwent standardization: sentences were extracted verbatim and tagged with provenance information. Pre-existing annotations (source/target domains, metaphor types) were preserved where available to maintain scholarly continuity.

Van Rijn-van Togerren (Tongerren, 1997) clusters *isolated metaphors*, which instantiate only a single image from a donor field. For these cases, we still assign source and target frames, and we type the instance as the isolated type, preserving the original catachretic characterization in the source.

Furthermore, we enriched bibliographic provenance with persistent identifiers. Where a DOI was available we recorded it; otherwise we stored title and author along with a resolvable link to the article.

Annotation procedure. Given the heterogeneous nature of our source material and the need to capture information beyond basic source-target mappings, we sought to measure the perceived metaphoricality of each expression. This approach addresses two key insights from the literature: metaphoricality exists on a continuum rather than as a binary property, and many medical metaphors are highly conventionalized, potentially affecting their perceived figurativeness (Julich-Warpakowski and Jensen, 2023; Bisang et al., 2006; Dunn, 2010; Gibbs, 2015).

Thus, we expanded the data already available from the original sources with two annotations:

²<http://www.metamia.com/>

1. **Binary metaphoricity** (M/L).
2. **Perceived metaphoricity scale** (0 = literal ... 7 = highly metaphorical).

All the metaphors were annotated through a Qualtrics survey upon specific instructions by twenty-seven advanced students of the *Informatica Umanistica* course at the University of Bologna participated (with C1 level English). To these, 15 online linguists recruited through the *Linguistlist* newsletter were added, with English as a primary language, making up a total of 42 annotators who annotated about 80 sentences each. Prior to annotation, all participants received instructions on defining metaphor and metaphoricity along with an example; the complete instructions can be found on the dedicated GitHub repository³.

Qualtrics presented items in complete random order per annotator; no time limits were imposed. Annotators saw only the sentence text; channel, provenance, and any source/target labels were hidden to avoid priming.

To ensure reliability, each sentence received independent annotations from a minimum of two annotators, with systematic overlap designed to calculate inter-annotator agreement. Each sentence received independent judgements from a median of 3 raters (mean 3.28 ± 1.18).

For each sentence, two questions were asked: (i) “Does this sentence contain a metaphor?”; (ii) “On a scale from 0 (literal) to 7 (very metaphorical), how metaphorical do you perceive this sentence to be?. If you put *No* to the previous question, write 0.”

Quality control. The responses were filtered according to consistency of the answer with the yes/no responses: if the metaphor was judged literal, the metaphoricity was explicitly said to be 0. Empty or invalid submissions were removed.

Fleiss’ kappa was 0.28, with the average percent agreement of 60%. For the agreement on the Likert scale, average Pearson r is 0.4 with the Spearman ρ being 0.4.

The relatively low inter-annotator agreement for binary rating reflects the inherent gradient nature of metaphoricity rather than annotation failure. This result aligns with established findings in metaphor research (Shutova, 2015) due to the subjective nature of figurative language perception. Our Likert scale ratings capture indeed this gradient nature more effectively than binary judgments, suggesting that metaphoricity is better understood as a spectrum of literality rather than discrete categories.

5.1.1.4 Frame linking and graph construction

We converted the MMC into an RDF graph aligned with Framester. In this release, for coherence with the metaphorical data already present in Framester, we model only the core elements: the conceptual metaphor, its source frame, its target frame, the linked textual examples, the language, and the provenance for every metaphor.

Because Framester has limited coverage of medical knowledge, we performed frame discovery and linking with an LLM-based linker and then conducted manual verification. We issued a single prompt to the Claude Sonnet API⁴ to propose links between MMC items and existing MetaNet frames, explicitly instructing it to refrain from forcing links when no match existed and to suggest new frames in those cases.

³github.com/dersuchendee/medical-metaphors-corpus

⁴We used *claude-sonnet-4-20250514* via the API. Temperature was 0, max_tokens were set as 1000. The exact system and user prompts are available on Github, along with the returned candidates.

We then verified each suggestion against the knowledge base and corrected them where needed. The model was conservative and tended to produce links with existing but not exactly coherent frames even when a suitable frame was not available, despite our instruction to propose new frames when an exact match did not exist. Approximately two-thirds of the initial LLM suggestions required manual adjustment, either through reassignment to more appropriate existing frames or through creation of new domain-specific frames.

We decided on a new frame only when two conditions held: first, no existing MetaNet frame captured the biomedical concept with sufficient specificity; second, the closest candidate would have required conflating distinct roles or mechanisms in our data. Each new frame was then aligned by asserting `fs:subsumedUnder` to an existing MetaNet superframe or to a FrameNet frame already bridged in Framester. We followed a naming and IRI policy consistent with MetaNet and Framester: singular noun phrases without leading articles as English labels (`rdfs:label`), lowercase hyphenated slugs in IRIs, and domain-specific frames placed under established biomedical or general superframes such as `Body_parts`, `Medical_conditions`, `Boundary`, `Competition`, or `Process_consisting_of_stages`.

As a result, 102 frames ended up with no suitable frame, leading to the creation of those frames. For example, `Cell` and its specializations `Mammalian_cell`, `Red_blood_cells`, and `Ependymal_cells` are subsumed under `Body_parts` to preserve consistency with existing MetaNet physiology coverage. `Cell_membrane` is subsumed under `Boundary` to reflect its enclosing function in the source domain of many MMC metaphors.

We transformed the tabular data into RDF with pyRML (Nuzzolese, 2025); the final graph is released in Turtle and N-Triples serializations. We adopt <https://w3id.org/framester/> as the canonical base for new resources.

We then uploaded the graph to Framester. In total, the construction produced more than 1,300 triples from 105 unique metaphors, and the addition of 102 new frames.

This graph is now queryable within the broader MetaNet ecosystem in Framester, allowing users to retrieve information on scientific metaphors, and to connect MMC metaphors to related resources such as the ones linked via the Amnestic Forgery work (Gangemi, Alam, and Presutti, 2018).

5.1.1.5 Technical validation

We conducted a preliminary analysis to address three essential inquiries: (1) the relationship between binary metaphoricity judgements and graded ratings in scientific discourse, including patterns of disagreement; (2) whether MMC supports information retrieval in Framester via SPARQL; and (3) a baseline zero-shot detection benchmark using LLMs. This last analysis establishes performance benchmarks for future method development while demonstrating the challenging nature of scientific metaphor detection. More sophisticated evaluation protocols (few-shot learning, fine-tuning, comparison with specialized metaphor detection models) represent future work that our dataset enables.

Annotation analysis The MMC contains 792 sentences drawn from scientific writing in which metaphorical language is either suspected or confirmed. Within these sentences, we identified 105 distinct metaphor types, spanning 57 unique target domains and 91 unique source domains. We derived a gold-standard label via majority vote together with the mean metaphoricity score for that sentence.

For the binary judgments we quantified disagreement with the following procedure:

We first compute the proportion of “yes” votes, denoted p_{yes} , and take the remaining fraction $1 - p_{\text{yes}}$ as the “no” votes. The disagreement score is then defined by

$$d = 1 - |p_{\text{yes}} - (1 - p_{\text{yes}})| = 1 - 2|p_{\text{yes}} - \frac{1}{2}|.$$

This index ranges from 0 when all annotators agree, to 1 when the panel splits exactly fifty–fifty.

For metaphoricity-rating questions (0–7 scale) we used the standard deviation σ of the ratings as the disagreement index: higher σ indicates greater annotator divergence about metaphorical intensity.

From the analysis of the data, upon close reading, we obtain the following findings:

- **Binary metaphoricity vs. Range.** Sentences judged metaphorical receive substantially higher metaphoricity ratings than non-metaphorical ones ($\mu_{\text{META}} = 3.41$ vs. $\mu_{\text{NON}} = 0.16$, $\Delta = 3.25$ points), a pattern that holds for 95% of question pairs.
- **Boundary cases.** The highest binary disagreement (perfect 50/50 splits) seems to arise in three main situations: (i) scientific terminology with a possible metaphorical reading (e.g. “*drug transport*”), (ii) highly lexicalised conventional metaphors, and (iii) domain-specific phrases whose interpretation depends on the context in which they are set.

We list below examples of highest rated and lowest rated metaphors:

Highest-rated (a) *It is inside the lungs that the virus turns nasty. It invades the millions of tiny air sacs in the lungs, causing them to become inflamed.*

(b) *(about cell biology) Three-step theory of invasion.*

Lowest-rated (a) *Two of its main activities—of the plasma membrane—are selective transport of molecules into and out of the cell.*

(b) *(Of a person who has cancer) I have learned to let the little things go.*

SPARQL queries To illustrate how the MMC can be explored and connected to MetaNet in Framester, we provide queries that retrieve metaphors involving a given frame and example sentences linked to those metaphors. Below, we show a query on metaphors that involve *Cancer* as source or target.

```
PREFIX mns: <https://w3id.org/framester/metanet/schema/>
```

```
SELECT ?m ?role ?otherFrame
WHERE {
  VALUES ?Cancer {
    <https://w3id.org/framester/metanet/frames/Cancer>
  }
  GRAPH ?g {
    { ?m mns:hasSourceFrame ?Cancer ; mns:hasTargetFrame ?otherFrame .
      BIND("Cancer_is_source" AS ?role) }
    UNION
```

```

    { ?m mns:hasTargetFrame ?Cancer ; mns:hasSourceFrame ?otherFrame .
      BIND("Cancer_is_target" AS ?role) }
  }
}
ORDER BY ?role ?m
LIMIT 500

```

The query returns all MetaNet metaphors where CANCER is either a source or a target frame. Running this query yields 22 distinct metaphors and shows when CANCER is mainly used as a target frame.

Another possibility is to see what are the most common source frames for a given target frame, for instance CANCER in our case.

```

PREFIX mns: <https://w3id.org/framester/metanet/schema/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT ?sourceFrame (SAMPLE(?sLbl) AS ?sourceLabel) (COUNT(DISTINCT ?m) AS ?n)
WHERE {
  VALUES ?Cancer {
    <https://w3id.org/framester/metanet/frames/Cancer>
  }
  GRAPH ?g {
    ?m a mns:Metaphor ;
      mns:hasTargetFrame ?Cancer ;
      mns:hasSourceFrame ?sourceFrame .
    OPTIONAL { ?sourceFrame rdfs:label|skos:prefLabel ?s1 }
    BIND(COALESCE(?s1, STR(?sourceFrame)) AS ?sLbl)
  }
}
GROUP BY ?sourceFrame
ORDER BY DESC(?n)
LIMIT 50

```

The query identifies 17 source frames; prominent examples are EDUCATION, JOURNEY, DIRT, WAR, and VISION. The graph structure enables queries that would be cumbersome in tabular formats: traversing from sentences to metaphors to frames, and linking to MetaNet’s broader conceptual network. The integration with Framester’s existing 2M triples enables cross-domain metaphor discovery. We report sample SPARQL queries, outputs, and related documentation on the Github repository.

LLM-based metaphor detection **Evaluation metrics.** To try to answer to the third question, we begin with the established standard from human annotations. As inter-annotator agreement is moderate, we can refer to it as a silver standard. For each metaphor detection item q_i , multiple human annotators provided binary judgments $R_i = \{r_1, r_2, \dots, r_n\}$ where $r_j \in \{\text{yes}, \text{no}\}$. We compute vote counts as $\text{yes\_count}_i = \sum_{j=1}^n \mathbf{1}(r_j = \text{yes})$ and $\text{no\_count}_i = \sum_{j=1}^n \mathbf{1}(r_j = \text{no})$, where $\mathbf{1}(\cdot)$ is the indicator function. The majority label is determined as $\text{majority}_i = \text{yes}$ if $\text{yes\_count}_i >$

no_count_i , $\text{majority}_i = \text{no}$ if $\text{no\_count}_i > \text{yes\_count}_i$, and $\text{majority}_i = \text{tie}$ otherwise. Additionally, we calculate the confidence of each annotation as $\text{confidence}_i = \frac{\max(\text{yes\_count}_i, \text{no\_count}_i)}{\text{yes\_count}_i + \text{no\_count}_i}$, representing the proportion of annotators who agreed with the majority decision.

To account for varying levels of human agreement, we implement a confidence-based weighting scheme that assigns higher importance to items with stronger annotator consensus. The weight for each item is calculated as $w_i = 2 \cdot (\text{confidence}_i - 0.5)$ when $\text{confidence}_i > 0.5$, and $w_i = 0$ when $\text{confidence}_i = 0.5$ (ties). This linear mapping transforms confidence scores from the range $[0.5, 1.0]$ to weights in $[0.0, 1.0]$, ensuring that items with perfect consensus receive full weight while barely-majority cases receive minimal weight. Items where annotators were evenly split (ties) are effectively excluded from weighted calculations by receiving zero weight.

LLM predictions are evaluated against the human silver standard using both traditional and confidence-weighted metrics. Let $S = \{(y_i, \hat{y}_i, w_i) : \text{majority}_i \neq \text{tie}\}$ represent the set of non-tie predictions, where y_i is the silver standard label, $\hat{y}_i$ is the model prediction, and w_i is the confidence weight. Standard accuracy is computed as $\text{Accuracy} = \frac{1}{|S|} \sum_{i \in S} \mathbf{1}(y_i = \hat{y}_i)$, treating all items equally. The confidence-weighted accuracy is calculated as $\text{Weighted Accuracy} = \frac{\sum_{i \in S} w_i \cdot \mathbf{1}(y_i = \hat{y}_i)}{\sum_{i \in S} w_i}$, giving higher importance to items with stronger human consensus. Similarly, precision and recall metrics are computed both in standard form and with confidence weighting, where for class c , weighted precision is $\frac{\sum_{i \in S} w_i \cdot \mathbf{1}(y_i = c \wedge \hat{y}_i = c)}{\sum_{i \in S} w_i \cdot \mathbf{1}(\hat{y}_i = c)}$ and weighted recall is $\frac{\sum_{i \in S} w_i \cdot \mathbf{1}(y_i = c \wedge \hat{y}_i = c)}{\sum_{i \in S} w_i \cdot \mathbf{1}(y_i = c)}$.

Items where human annotators reached no consensus (ties) receive special treatment in our evaluation framework. During silver standard construction, tie items are identified and labeled but not assigned a definitive binary (yes/no) classification. In the weighting phase, these items receive zero weight ($w_i = 0$), effectively removing them from confidence-weighted calculations while preserving them in the dataset for transparency. During experimental model evaluation, tie items are completely excluded from all metric calculations, ensuring that models are only assessed on cases where human consensus exists. This exclusion strategy ensures that the evaluation focuses on cases with clear ground truth while avoiding penalizing models for predictions on inherently ambiguous examples where even human experts disagree.

Models and parameters setup. We used four LLMs exclusively through their APIs: *GPT-4*, *o1-preview*, *o3-mini*, *DeepSeek*, *Claude Opus 4*. All experiments used default inference settings, with the sampling temperature fixed to 0 to obtain deterministic outputs. The sole exception is *o1-preview*, whose API mandates a default temperature of 1. For the technical validation, we evaluate with and without the confidence-based weighting scheme. We include a simple zero-shot detection baseline solely to show the dataset’s reuse.

Results.

Table 5.2 presents standard evaluation metrics, while Table 5.3 shows our confidence-weighted results.

5.1.1.6 Discussion

In this section, we discuss the results of the experimental setup and the potential of the dataset in computational metaphor research.

TABLE 5.2: LLM Performance on scientific metaphor detection without weights. Precision, Recall and F1 are macro-averaged.

Model	Acc	Prec	F1	Rec
o1-preview	0.716	0.714	0.714	0.714
Claude-Opus 4	0.711	0.746	0.707	0.725
o3-mini	0.706	0.785	0.695	0.727
DeepSeek	0.683	0.745	0.673	0.702
GPT-4	0.655	0.785	0.695	0.727

TABLE 5.3: LLM Performance on scientific metaphor detection with weights.

Model	wAcc	wPrec	wF1	wRec
o1-preview	0.758	0.716	0.716	0.714
Claude-Opus 4	0.755	0.756	0.705	0.721
o3-mini	0.752	0.799	0.690	0.706
DeepSeek	0.725	0.757	0.668	0.683
GPT-4	0.690	0.776	0.626	0.655

The relatively low inter-annotator agreement for binary rating reflects the inherent gradient nature of metaphoricality rather than annotation failure. This aligns with established findings in metaphor research: Shutova (2015), for instance, notes that moderate agreement is typical in metaphor annotation tasks due to the subjective nature of figurative language perception. Our Likert scale ratings (Pearson $r = 0.441$) capture indeed this gradient nature more effectively than binary judgments, suggesting that metaphoricality is better understood as a spectrum of literality rather than discrete categories.

Our exploratory analysis reveals a strong positive correlation between binary metaphoricality judgments and high metaphoricality ratings in scientific discourse. The substantial difference in literality ratings between metaphorical ($\mu = 3.41$) and non-metaphorical expressions ($\mu = 0.16$) suggests that annotators do perceive metaphors as having a low literality level.

The disagreement patterns we identified also provide insights into the inherent challenges of metaphor annotation. Annotator judgments, in some cases, reveal genuine boundary cases involving scientific terminology with potential metaphorical readings, highly conventionalized metaphors, and domain-specific expressions where scientific expertise influences perception. These instances represent the most difficult cases for both human annotators and automated detection systems. Furthermore, such hard cases with low agreement tend to often represent the most theoretically interesting boundary phenomena rather than annotation failures. The metaphoricality ranges in our dataset naturally enable confidence-weighted evaluation methodologies that account for varying levels of human consensus. By leveraging the degree of annotator agreement on each item, we can develop evaluation metrics that prioritize clear-cut cases while appropriately handling inherently ambiguous instances where human judgment varies.

To make these borderlines empirical and reproducible, we qualitatively audited the full set of tie items and coded them into a small taxonomy of ambiguity/error types in Table 5.4. The goal is twofold: (i) to see where and why human judgments diverge, and (ii) to provide concrete adjudication guidelines so future annotation

and model evaluation can treat these cases consistently.

TABLE 5.4: Ambiguity / error typology observed in *tie* items.

Type	Why ambiguous	Examples
Technolect polysemy (literal-looking metaphor)	Terms with metaphorical surface forms are conventionalized in biomedicine; raters split on whether any cross-domain tension remains.	<i>“at the end of the dissemination phase...”</i> ; <i>“to suppress colonizing organisms”</i> ; <i>“At a metastatic site these cells...”</i> .
Lexicalized conventional metaphors	Figurative origin is bleached; some readers still perceive a mapping, others treat it as fixed jargon/idiom.	<i>“outspoken about his battle with OCD”</i> ; <i>“So how does the virus attack?”</i> ; <i>“destroying the engine at the heart of the disease”</i> .
Context / lens dependence	Metaphoricity depends on reading (emotional/supportive vs structural/physical; cognitive vs physical).	<i>“friends and family...support me”</i> ; <i>“I couldn’t absorb what was going on”</i> ; <i>“feels like two divergent realities”</i> ..
Co-occurring metaphors (competing labels)	Multiple plausible metaphorical cues in one span; single-label gold forces an arbitrary choice.	<i>“blocking...steps for invasion...prevent tumor...”</i> ; <i>“destroying the engine at the heart of the disease”</i> ; <i>“dissemination phase...seeds of metastases”</i> .
Property-level ambiguity (blending)	Raters agree a mapping exists but disagree on <i>which property</i> is projected (cause, centrality, aggressiveness, etc.).	<i>“the virus turns nasty...invades...”</i> — AGGRESSION VS VIRULENCE; <i>“two new trouble spots”</i> — LOCALIZATION VS SEVERITY; <i>“engine at the heart of the disease”</i> — CAUSAL CORE VS POWER SOURCE.
Affective idioms misread as metaphor	Formulaic stance/affect expressions flagged as metaphor despite no cross-domain mapping being required.	<i>“a few brutal panic attacks”</i> ; <i>“stop worrying ... give thanks to God...”</i> ; <i>“friends and family...support me”</i> ..

Our confidence-weighted evaluation framework reveals that the consistent 3-4.6% improvement across all models when weighted by human consensus indeed demonstrates that current LLMs perform systematically better on cases where humans strongly agree, while struggling disproportionately with ambiguous instances.

This pattern has important implications for practical applications: *o3-mini*’s largest weighting benefit (+4.6%) suggests it could serve reliably in high-confidence scenarios while requiring additional safeguards for borderline cases. *o1-preview*’s balanced performance across both weighted and standard metrics indicates more robust handling of metaphor ambiguity, making it suitable for applications requiring consistent performance across diverse linguistic contexts. We attribute these models’ success to the fact that they are tuned for deliberate reasoning in few-token budgets; their internal chain-of-thought appears particularly effective for short, domain-specific classification zero-shot prompts like ours.

The conservative precision-recall profiles observed across all models (high precision but low recall for metaphor detection) reflect a systematic bias toward literal interpretation. This case suggests that LLMs adopt a cautious decision boundary, labelling a sentence as metaphorical only when strongly lexical cues (e.g. *“battle,” “storm,”* or explicit anthropomorphism) are present. While this reduces false positives, it may limit utility in applications requiring comprehensive metaphor identification, such as literary analysis or patient communication assessment. Therefore, the MMC dataset surfaces cases that even frontier LLMs find non-trivial, making it a valuable stress-test for future metaphor-aware language technology.

5.1.1.7 Data availability

The MMC is publicly available on GitHub at <https://github.com/dersuchendee/medical-metaphors-corpus>. The contents of the repository are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

A permanent archive of the MMC, the non-annotated and the annotated version, along with the graph in ttl serialization is also deposited on Zenodo with the DOI [10.5281/zenodo.17313744](https://doi.org/10.5281/zenodo.17313744).

The MMC graph can be queried via the Framester YASGUI interface at https://etna.istc.cnr.it/framester_web/yasgui.html and the programmatic SPARQL endpoint at <https://etna.istc.cnr.it/framester2/sparql>.

5.1.1.8 Code availability

All code used for data processing, annotation aggregation, and analysis is available in the same GitHub repository at <https://github.com/dersuchendee/medical-metaphors-corpus>.

Dependencies and execution instructions are documented in the repository's README file.

5.1.2 The Balanced Conceptual Metaphor Testing Dataset (BCMTD)

BCMTD is a small subset of the MMC designed to serve as an evaluation set designed to test a system's capabilities for the processing of generic conceptual metaphors and specialised, domain-bound metaphors, as opposed to literal sentences.

Partly derived from MMC, the use of the BCMTD should be complementary to its bigger counterpart: BCMTD stresses balance and control for cases in which a person wants to both assess metaphoricity of general and domain-specific metaphors, whereas MMC maximises ecological coverage.

5.1.2.1 Motivation

Large open-domain benchmarks such as VUA (Steen et al., 2010) or TroFi (Birke and Sarkar, 2006a) capture breadth, but they are heavily skewed toward everyday language and contain only a handful of examples drawn from specific domains such as academic language. Conversely, specialised examples of scientific metaphor (e.g. the COVID datasets in Semino, Demjén, and Demmen (2018)) are not publicly available, not annotated, and lack a built-in literal control set, so a high F_1 can mask trivial domain cues. BCMTD fills this gap by presenting three perfectly balanced strata (Table 5.5) that force a model to discriminate along two orthogonal axes: *figurative vs. literal* and *general-purpose vs. scientific domain*.

5.1.2.2 Data Collection

In this section, we describe the data sources collected to build the dataset.

- **General conceptual metaphors**, sampled from Metanet-derived annotations already harmonised in Framester (Gangemi et al., 2016). To avoid overlap with the set published by Wachowiak and Gromann (WG) (Wachowiak and Gromann, 2023), all candidates present in WG were removed before sampling.

- **Scientific conceptual metaphors** come from two medically oriented sources contained in the MMC: (i) the appendix of Van Rijn & Van Tongeren’s typology of scientific metaphors (Tongeren, 1997), and (ii) Semino et al.’s medical metaphors collection (Semino, Demjén, and Demmen, 2018). These two sources were specifically selected from the MMC due to their uniformity and due to the online absence of the data used by Tongeren (1997), preventing data leakage issues.
- **Literal controls** were extracted from VUA after applying the mis-label corrections of Wachowiak & Gromann (Wachowiak and Gromann, 2023).

5.1.2.3 Annotation model

Each sentence carries the annotation of a source and target domain, as provided by the sources, and information linking back to the original publication. It also carries a label of the sentence, showing whether it is considered literal or metaphoric. In this work, metaphoricity is reduced to a binary label for ease of testing, since the aim of this dataset is to provide a quick preliminary evaluation set.

5.1.2.4 Dataset statistics

The dataset is composed of 49 conceptual metaphors, 49 scientific metaphors and 49 metaphors from VUA, for a total of 147 samples. Table 5.5 provides statistics and examples from each dataset layer.

TABLE 5.5: Summary of BCMTD dataset statistics by type, including instances, percentage of metaphorical sentences, samples, examples, and metaphors.

Type	# Instances	% Met.	# Samples	Example	Metaphor
Conceptual	49	100	49	<i>“The money is all dried up.”</i>	MONEY AS LIQUID
Scientific	49	100	49	<i>“Tissues of cells independent of normal growth-controlling forces.”</i>	CELLS AS SOCIETY MEMBERS
VUA (literal)	49	0	49	<i>“Then I normally watch it with you.”</i>	None

5.1.2.5 Data availability

The dataset is available at the following Github link: <https://github.com/dersuchendee/metaphorical-logic-augmented-generation/blob/main/data/datasets/bctmd.csv> and the Zenodo DOI: <https://doi.org/10.5281/zenodo.15355851> as a part of the study presented in Section 4.2.3.1.

To sum up, BCMTD complements larger resources by offering a balanced, diagnostic test of whether models that master generic conceptual metaphors can transfer to specialised scientific metaphors while avoiding spurious flags on literal scientific prose. Its three matched strata make per-stratum evaluation straightforward. We recommend using BCMTD strictly as a held-out challenge set with macro-level reporting and sanity baselines; success here indicates robustness to domain shift and better precision on literal scientific text, but it should be interpreted alongside

broader, ecologically diverse benchmarks, such as the MMC, which treats metaphoricity as a range rather than a binary label. The dataset will be experimentally evaluated as part of the study described in Section 4.2.3.1.

5.2 Representing, recognising, and producing metaphor

In this section, we discuss how to operationalize metaphor according to the complementary facets of representation and analysis. These aspects work together: pattern-based representation provides the stable yet flexible dynamic framework needed for integration with modern LLM technologies, while analysis addresses the deeper question of what it actually means for a computational system to understand a metaphor.

Regarding the plausibility issue mentioned at the beginning of this chapter, metaphor understanding serves as a good testing task. The success of such systems will ultimately be measured not by their ability to identify metaphors, but by their capacity to understand, interpret, and create them in contextually appropriate and meaningful ways.

5.2.1 Representing: the Blending Ontology

5.2.1.1 Motivation

The Blending Ontology has been developed to computationally address cases of computational blending for metaphorical phenomena. The motivation for this work emerges from critical gaps identified in computational approaches to metaphor processing, particularly the limitations of current knowledge representation frameworks in capturing the emergent nature of metaphorical meaning construction.

As demonstrated in Section 3.1.1.3, existing computational approaches to metaphor remain fragmentary and theoretically incomplete. In fact, current metaphor processing systems assume static mapping, thus fail to capture the bidirectional influences proposed in CBT. Furthermore, while emergent properties arise in blend spaces, the representation of this structure is not properly addressed in current works (§ 3.0.6). As we also saw in § 2.2.2.5, current systems treat metaphors as isolated linguistic phenomena. However, metaphors are born and evolve within discourse contexts. The ability to perform reblending when new questions or contexts arise, just how humans process extended metaphors, remains computationally unaddressed.

To address these current gaps, we describe in the next section the Blending Ontology.

5.2.1.2 Ontology structure

The Blending Ontology⁵ is engineered through a top-down, pattern-based strategy that extends the Description and Situations pattern (§2.1.3) via mDnS to formalise conceptual blending. It is based upon mDnS (mereological Descriptions and Situations)⁶ and is aligned to the Cognitive Perspectivisation ontology as defined in Gangemi and Presutti (2022)⁷ and the MetaNet schema from the Amnestic Forgery ontology (Gangemi, Alam, and Presutti, 2018).

⁵Available at <https://github.com/dersuchendee/BlendingOntology> and <http://www.ontologydesignpatterns.org/ont/mdns/blending.owl>

⁶<http://www.ontologydesignpatterns.org/ont/mdns/mdns.owl>

⁷<http://www.ontologydesignpatterns.org/ont/persp/perspectivisation.owl>

In particular, mDnS consists of a novel axiomatization for the D&S ontology pattern⁸, fully compatible with classic D&S and constructive D&S (cDnS) and is based on intensional/extensional mereology and optimizing expressivity with respect to low complexity for a hybrid reasoner. By introducing the concepts of descriptions and situations, the Description and Situations ODP (Gangemi and Mika, 2003b) provides a formalization of frame semantics where a situation is a set of facts as described by an observer that is interpreted by a description, a schema that defines the entities observed in a situation.

In mDnS, each description d corresponds to a class of situations $s_1, \dots, s_n$ as usual, such that each situation s_i is classified by d , and d classifies (is satisfied by) some s_i . We use the notation $s_i \in \text{nClassifiedBy } d$ to indicate that situation s_i is classified by description d , and $d \in \text{nClassifies}$ to indicate that d classifies some situation s_i . The components at the intensional and extensional levels in the mDnS vocabulary are represented as follows:

- Intensional component (*iComponent*): used for the `usesConcept` relationship.
- Extensional component (*eComponent*): used for the `isSettingFor` relationship.

These roles are foundational in the classic DnS (Descriptions and Situations) ontology framework, where:

$$\begin{aligned} iComponent &\longleftrightarrow \text{usesConcept}, \\ eComponent &\longleftrightarrow \text{isSettingFor}. \end{aligned}$$

The Blending Ontology can generalise the notion of metaphors as frame mappings, and uses mDnS for ground axiomatisations. Below, we list the main components of the ontology.

Classes such as `Blend`, `Blendable`, `Blended`, and `Blending` represent different stages and elements of the conceptual blending process. More specifically, the basic Descriptions include:

1. `Blending`, a generic Description enabling `Blendable` frames to be mapped by subsuming some of their components;
2. `Blendable`, two or more Descriptions with components that are subsumed by components in the `Blending` frame;
3. `Blended`, the Description that inherits components from the `Blendable` frames, firstly through the `Blending` sharable space, and eventually inheriting other components from the `Blendable` frames, making emerge a novel set of blended situations;
4. `Blend`, the meta-description with components `Blending`, `Blendable`, and `Blended`.

Object properties, namely `blendableComponent`, `blendedComponent`, and `blendingComponent`, create links between classes that mirror underlying cognitive processes. They facilitate the mapping of components from one conceptual domain to another, enabling the emergence of blended concepts.

Specialized properties such as `enablesBlending` and `inheritsRoleFrom` define the interactions among conceptual elements within the ontology, ensuring that the blending process remains both logically consistent and semantically coherent. When

⁸<http://www.ontologydesignpatterns.org/cp/owl/descriptionandsituation.owl>

a conceptual metaphor is derived from two input frames, the roles within these frames are first generalized and then inherited by a broader Blending construct via the `inheritsRoleFrom` property. This process, in turn, triggers the blending mechanism through the property `enablesBlending`, allowing the resulting Blended space to assimilate components from the original Blendable and Blending frames. Consequently, the Blended space emerges as a new, independent domain representing the metaphor.

All components of Blend are Descriptions (Axiom 1), whose component requirements are specified in Axiom 2:

$$(\text{Blend} \sqcup \text{Blendable} \sqcup \text{Blending} \sqcup \text{Blended}) \sqsubseteq \text{Description} \quad (5.1)$$

$$\text{Blend} \equiv \geq 2 \text{eComponent.Blendable} \sqcap = 1 \text{eComponent.Blended} \sqcap \exists \text{eComponent.Blending} \quad (5.2)$$

Every Blend Description necessarily generates a specific subdomain context in which all the situations it classifies are situated. This dependency involves quantification over contexts and the use of predicates that capture meta-level relationships between descriptions and parts of the domain. Therefore, for Axiom 3, we choose a First Order Logic formulation, though the same structural dependency can be enforced in practice using SHACL rules:

$$\forall x(\text{Blend}(x) \rightarrow \exists C(\text{sub}\Delta(C) \wedge \text{ContextDep}(x, C) \wedge \forall s((\text{Situation}(s) \wedge \text{ClassifiedBy}(s, x)) \rightarrow \text{InContext}(s, C)))) \quad (5.3)$$

Blendables and Blended Descriptions are subsumed by the Blending description (Axiom 4):

$$(\text{Blendable} \sqcup \text{Blended}) \sqsubseteq \text{Blending} \quad (5.4)$$

Subsumption is however debatable when we consider the extension provided by Blended frames and the consequential possible existence of new blended situations emerging from the application of a Blend. We imagine three possible semantics:

$$\begin{aligned} \text{a) } & \text{Blended} \sqsubseteq (\text{FirstBlendable} \sqcap \text{SecondBlendable}) \\ \text{b) } & \text{Blended} \sqsubseteq \text{DominantBlendable} \sqcap \text{Blended} \sqcap \text{NonDominantBlendable} = \emptyset \\ \text{c) } & \text{Blended} \sqcap \text{Blendable} = \emptyset \end{aligned} \quad (5.5)$$

In the first case (5a), we would be incoherent if the two Blendable situation types are disjoint (incompatible frames). Even if we introduce a new Blended situation s , it would result as an instance of both Blendables, so generating an inconsistency.

In the second case (5b), we introduce the principle of Dominance for metaphoric blending (Lakoff and Johnson, 1980) (See Section 2.2.2.1), and e.g., a CRIME AS INFECTION situation (a Blended situation t) would still be a CRIME, but not an INFECTION. We cannot assume that t is the same individual as a situation s in the DominantBlendable class, but for sure t is not in the NonDominantBlendable class. However, also in this case t might contain features from the NonDominantBlendable class that are inconsistent with those of the DominantBlendable one.

In the third case (5c), besides Dominance, we introduce a Composition principle, by which a new Blended situation cannot be an individual in any Blendable class: $t \neq s$. In this case, we take a radical decision: when a Blend conceptualisation is applied to generate a Blended description d , while maintaining the schematic features of a Blending description (Axiom (4)), d enables the creation of a non-conservative extension of the domain of interpretation Δ . In other words, any situation s that is classifiable by a Blended description must be assumed as a novel entity in Δ .

Since all Descriptions have intensional components (a.k.a. *semantic roles*), Axiom 5 formalizes the intensional composition of a Blended:

$$\text{Blended} \sqsubseteq \exists \text{Component}. (\text{Concept} \sqcap \exists \text{ComponentOf}. \text{DominantBlendable}) \sqcap \sqcap (\text{Concept} \sqcap \exists \text{ComponentOf}. \text{NonDominantBlendable}) \quad (5.6)$$

5.2.1.3 The nature of the blended

The discussion of Axioms 5c and 6 align to Gangemi, Alam, and Presutti (2018)'s formalization of Amnestic Forgery, where a question was raised: once a metaphor generates a blended frame, should we assume that the role value from one domain simply becomes identical to the role value from the other (an economic, lean commitment) or should we allow for the emergence of an entirely new hybrid entity (a more complex, multiplicative design)? In the Blending Ontology, and consistent with literature in the field (§ 2.2.2.3), the nature of the Blended component tends to the second option.

In fact, in a metaphor, the Blendable elements do not share an overlap because their relationship is established through selective intensional projection. Rather than merely juxtaposing frames, this process involves an epistemological superposition in which the dominant Blendable overlays the non-dominant one, creating a new, transformed entity. In this case, the treatment of identical entities is different from that in qua-entities, which isolate forms of aspectual slices of core entities, yet each instance remains identifiable and maintains its original characteristics (Guizzardi and Almeida, 2020). In fact, the Blended is not merely a copy or a type of the dominant Blendable, but rather represents its conceptual reconfiguration where the original identity is fundamentally altered.

This idea finds confirmation, among others, in several studies in medical metaphors, already described in Section 2.3.3, whose conceptualization, as opposed to their literal counterparts, has been proven to influence decision making as well as thought itself. In our view, just like describing the flu metaphorically (e.g., as a war) increased individuals' willingness to get vaccinated with respect to its literal wording (Scherer, Scherer, and Fagerlin, 2015), the blended entity is conceptually different from its target frame alone (the flu): it is enriched with novel attributes and loses some aspects of its initial form.

The interference effect of the source, non-Dominant frame over the other results in a "warping" of the original properties of the Dominant frame, leading to a shift in meaning that cannot be reversed without a fundamental change (WAR AS A FLU works entirely differently). In this context, perspectivisation reveals the interpretative standpoints that shape the blended entity.

5.2.1.4 The metaphorical mechanism as perspectivisation.

Perspectivisation occurs when at least two distinct views can be selected from a ground situation, highlighting the inherent contrast between different interpretations. This contrastive approach is essential for understanding metaphor, as it often juxtaposes the expectation with the fact, the literal with the metaphorical, or even multiple metaphorical readings alongside the literal. In essence, metaphor is always an act of perspectivisation, whether intentional or not, since it involves integrating an anomalous viewpoint from reality into a unified conceptualization. In this sense, the Blending Ontology links to the Cognitive Perspectivisation Ontology (Gangemi and Presutti, 2022), which models the Perspectivisation frame—a scenario where facts are narrated through a Lens that shapes the Conceptualiser's view based on

their Background. In the next section, we demonstrate this integration through concrete examples that evaluate the Blending Ontology.

5.2.1.5 Evaluation

In this section, we present a two-fold assessment of the Blending Ontology: first, a formal validation with automated reasoning, and second, case-study demonstrations of its use in metaphor analysis.

Formal evaluation. We verified the Blending Ontology with the HermiT 1.4.3.456 reasoner in Protégé 5.6.4. On the test data that instantiates the CRIME AS INFECTION blend, HermiT pre-computed all inferences (class and property hierarchies, assertions, same-individuals) in 122 ms, showing that the model is tractable for interactive work.

The run returned zero unsatisfiable classes and no range/domain violations, confirming global consistency. In addition, HermiT (i) classified the individual ‘crimeAsInfectionBlend_1’ as a ‘Situation’, (ii) materialised the component structure, and (iii) activated the property-chain axiom ‘iComponent subsumes iComponentOf’ so that ‘crime_1 enablesBlending crimeAsInfectionBlend_1’ is entailed. Crucially, the non-conservative-extension axiom (5c) is respected: HermiT proves that every blended entity is logically novel with respect to its inputs. Together, these results confirm that the ontology is both logically sound and capable of supporting the inferential tasks required for metaphor analysis.

Contexts of use. To demonstrate the Blending Ontology’s inferential capabilities, we also present three contexts of use: first, analyzing an isolated conceptual metaphor and its linguistic instances; second, examining metaphors within news discourse contexts; third, using the Blending ontology to formalize studies on transformational creativity and multimodality.

Isolated metaphors.

The Blending Ontology can be used to understand the selective projection of the roles among Blendable elements and the relationship between them within single metaphors. As an example, we focus on the metaphor CRIME IS AN INFECTION.

Historically, crime has often been compared to a disease. In the 20th century, this analogy typically portrayed crime as a kind of psychological disturbance. However, a more enduring perception views crime as an infectious, contagious disease. This metaphor gained traction in the mid-1980s, paralleling the widespread impact of HIV (CRIME IS A VIRUS) suggesting that, like viruses, crime self-replicates, mutates, and inflicts harm. It is often expressed through phrases such as “crime is a virus”, “the blight of crime”, “national epidemic”, and others (Oliynyk and Nauhenko, 2018).

In our model, CRIME and INFECTION are the two Blendable, each with defined roles. In fact, depending on the context, Crime can have as roles NormViolation, CriminalAct, and CrimeSeverity. On the other hand, an Infection implies the concepts of SpreadRate, Severity, Population, Process and Agent, among others. In both cases, we can note that the Blendable and their roles can be abstracted to the concept of HarmfulSpreading, which implies the roles of Rate, Event, Medium, HarmSeverity, and Entity. Therefore, in the Blended space CRIME AS INFECTION, the two Blendable INFECTION and CRIME subsume roles from the Blending. With these assertions in our ontology, we can infer:

1. The Blended space CRIME AS AN INFECTION inherits roles from each frame (input space) through selective projection of their roles;
2. In the Blending, roles like CrimeSeverity from CRIME map to Severity of HARMFUL SPREADING, while SpreadRate from INFECTION maps to Rate;
3. The Blending space constrains which roles can be projected into the blend.

Following this view, when CRIME is conceptualized as INFECTION, selected properties of the non-dominant frame (e.g. associated with contagion) become available for reasoning about criminal behavior. The Blended space thus allows for novel inferences not available in either input space alone, such as “crime requires quarantine measures”. This example shows how single isolated metaphors can be analyzed, but the Blending Ontology also supports to analyze metaphors in real-life use cases.

Metaphors in context.

In addition to isolated metaphors, the Blended Ontology aligns with the Cognitive Perspectivisation Ontology (Gangemi and Presutti, 2022) for the analysis of metaphorical expressions as viewpoints within broader contexts, such as news narratives, a current line of research (Motta et al., 2024). Let’s take as an example two portrayals in news about immigrants:

(1) *If the victorious party should follow up vigorously its successes and press the vanquished, these islands may be **inundated** with refugees to an **alarming** extent.* – Times, 1852

(2) *It was **good** for the West Indies and the Anti-Slavery party if immigrants could be **induced to pour into** those colonies.* – Times, 1859

Both sentences evoke the conceptual metaphor MIGRANTS ARE LIQUID. Although they share the same Background, a commonsense knowledge framing of immigrants and liquid, they differ in the sentiment on the target frame. The first sentence elicits a negative attitude (framing the flow as DANGEROUS FLOW), whereas the second conveys a positive sentiment (framing the flow as BENEFICIAL FLOW). Through our ontology’s alignment with the Perspectivisation ontology and its reasoning, we can infer:

1. Both examples (1) and (2) involve the same blendable inputs (MIGRANTS and LIQUID) but project different attitudinal elements;
2. The Background knowledge remains constant across different instantiations of the same metaphor;
3. The Lens computes different metaphorical interpretations as the composition of the Attitude and Blending frames.

Our inferential system shows that a metaphorical PerspectivizationSituation inherits its structure from the Blend but introduces attitudinal constraints. Furthermore, the emergent properties of the Blend are predictable given the input spaces and the specific Attitude value. In contexts such as public discourse, a Conceptualizer can adopt a collective attitude, further influencing whether a social phenomenon (like migration) is perceived positively or negatively (Gangemi, 2008).

Transformational creativity and multimodality.

The blending property operationalized in the Blending Ontology can also be extended to another context, that of scientific transformational creativity, as scientific breakthroughs often work through metaphorical thinking. In fact, transformational creativity, following the definition of Boden (2004), is a distinct type of the creative process that hides behind groundbreaking scientific discoveries and entails restructuring a conceptual hypothesis space basing on the emergence of new data. Thus, unlike ordinary creativity that works within existing rules, transformational creativity directly changes the rules.

Following Schapiro, Black, and Varshney (2025), in a formalization of transformational creativity, a conceptual space S is transformed to S' by changing its enabling constraints, which maximally alters downstream rules. Therefore, in Schapiro, Black, and Varshney (2025), transformations do not extend the domain but change admissibility via new axioms. If we adapted Schapiro, Black, and Varshney (2025)'s work to the Blending Ontology, we could model conceptual transformation through domain combination as follows: when a `Blend` combines conceptual spaces, it creates a dedicated context via Axiom 3, while Axiom 5c ensures a non-conservative extension that generates genuinely novel blended situations. This makes the resulting `Blended` context formally equivalent to Schapiro, Black, and Varshney (2025)'s transformed space S' .

In fact, in the Blending Ontology, each `Blend` instance induces such a transformation: Axiom 3 creates a dedicated sub- Δ context C for the blend, while Axiom 5c enforces a non-conservative extension so that every `Blended` situation is novel with respect to its `Blendable` inputs. Selective role projection via `inheritsRoleFrom` and the shared Blending schema acts as the new axiom set for C . Thus, the `Blended` context is precisely the transformed space S' , and modifications to dominance or projection constraints correspond to axiom-level edits that, by Theorem 4 of Schapiro, Black, and Varshney (2025), have the greatest transformative potential.

Another context of application of the Blending Ontology is that it can process multimodal signal, as CBT has been shown to have in previous research (Ox, 2014). In Ox (2014), the author treats visualization as cross-modal reasoning built from two operations: structure-mapping (analogy) that preserves a relational skeleton, and conceptual blending that yields emergent structure, and operationalizes them through music-to-image translations (e.g., Stravinsky; Kurt Schwitters' *Ursonate*). The goal is not musicology but a general recipe for richer qualitative visualization: music is chosen because it offers stable, rule-governed structure (keys, rhythm, dynamics, phonetics) that cleanly maps onto visual dimensions (time-to-space, pitch-to-verticality, dynamics-to-scale, vowel features-to-color, consonant features-to-cut/flip operations). In the cognitive space transfer, the creator's background imagery (e.g., fjords, façades) is injected so the mapping is filtered through a concrete point of view. In the Blending Ontology, each such visualization can be considered a `Blend`: the musical and visual inputs are `Blendable` descriptions; the shared relational skeleton, inclusive of plastic properties (See Section 2.2.2.7) is captured by the `Blending`; and the resulting `Blended`, by Axiom 3, induces a dedicated sub-context and, by Axiom 5c, is a non-conservative extension. What was not so far supplied was a formal account of the enabling constraints and clash-handling that make these mappings sound: there is no typed role system, no explicit dominance/asymmetry settings, no mechanism to declare or reason over context creation and novelty. The Blending Ontology allows contribution to these pieces, and the use of *Perspectivisation* makes cognitive space transfer explicit via *Conceptualiser-Background-Lens* classes.

5.2.1.6 Discussion

The Blending Ontology represents a significant advancement in the computational formalization of metaphorical reasoning, addressing fundamental gaps in current approaches to metaphor processing. This work establishes new theoretical and methodological foundations for understanding how metaphorical meaning emerges through dynamic conceptual integration.

The ontology's primary theoretical contribution lies in its formalization of emergent metaphorical meaning through non-conservative extensions (Axiom 5c). Unlike previous computational approaches that treat metaphors as static mappings between pre-existing domains, the framework captures the intrinsically creative aspect of metaphorical thinking, the emergence of novel conceptual structures that cannot be reduced to their constituent inputs. This formalization aligns with empirical findings from cognitive science showing that metaphors do not merely transfer properties but create new ways of thinking and acting in the world. Such findings are also extended to other fields of interest, such as visual and musicological interpretation, as well as transformational creativity in science, allowing the ontology to be used for such cases.

However, the evaluation of the blending ontology through its computational implementation, and particularly of the blending property, is not a trivial task. As shown in the literature, even LLM-based methodologies cannot work alone but need support from a neurosymbolic infrastructure (see Section 3.1.1.3). Furthermore, while the ontology formalizes the structural constraints of blending, the practical operationalization of blending properties must be determined contextually for each metaphorical instance, requiring integration with broader computational frameworks capable of handling heterogeneous, multimodal data that often lacks corresponding knowledge base representations. To approach this issue, in the next section, we move onto the discovery and analysis part of metaphorical operationalization.

5.3 Computational implementation of the blending property

As seen in Section 2.2.2.3, the Blending property is a crucial part of blending operations. In fact, it is what actually motivates the metaphor, and is a theoretical contribution that differentiates CBT from other theories. Despite its importance, to the best of our knowledge, there has never been a standard way to represent it or extract this property, leading to few computational studies taking this structure into consideration (see Section 3.0.2). The aim of this section is thus to investigate the operationalization and extraction of this property through these main approaches: frame-based, embedding-based, and LLM-based.

5.3.1 Problem formulation

Starting from the plausibility problem (§3.0.2), we focus on the extraction of the Blending property motivating the metaphor. The problem formulation can be simply described as follows.

Given:

- Source concept: S
- Target concept: T
- Candidate blending property: B

Find B such that $(S \sqsubseteq B) \wedge (T \sqsubseteq B)$ and B maximizes $\text{coherence}(S, T, B)$.
 We instantiate the coherence function $\text{coherence}(S, T, B)$ in three parallel ways:

1. **Frame-based:** $\text{coherence}_{\text{KB}}(S, T, B)$ - structural compatibility score
2. **Embedding-based:** $\text{coherence}_{\text{emb}}(S, T, B)$ - distributional similarity score
3. **LLM-based:** $\text{coherence}_{\text{LLM}}(S, T, B)$ - language model plausibility score

The output is B^* , the discovered blending property.

Before proceeding with individual approaches, we establish our central complementarity hypothesis: effective metaphor processing requires the integration of structural knowledge and flexible reasoning (LLMs), as no single approach can capture all aspects of metaphorical cognition. The following three experiments test this hypothesis by systematically exploring the limitations of each approach in isolation, building the empirical foundation for our integrated neurosymbolic solution.

5.3.2 Frame-based extraction of the blending property

Frame knowledge bases are comprehensive resources for information, indeed, they have been a source of inspiration for contemporary metaphor studies.

For the frame-based extraction, we test whether systematic mappings can be recovered along three dimensions: semantic similarity (shared synsets), structural commonality (shared frame elements), and relational similarity (regularities in how roles pair across frames). Since Framester already includes and links MetaNet metaphors, we evaluate whether these resources alone surface the principles that underwrite plausible blends.

5.3.2.1 Methodology

We queried Framester's SPARQL endpoint for metaphors with explicit source and target frames, and we retrieved entailments and, when present, mapped roles (see Gangemi, Alam, and Presutti, 2018). We used a Python script⁹ to perform:

- Frame-to-frame links via `metanet:hasSourceFrame` and `metanet:hasTargetFrame`, along with entailment descriptions and examples optional joint roles `metanet:sourceRole` and `metanet:targetRole`.
- A frame typing step that checks whether a frame appears as source or target, broadened with `skos:closeMatch` and `schema:subsumedUnder` to improve recall.
- A similarity probe that aggregates frame elements and WordNet synset labels per frame from prepared CSV queries and computes their intersections across paired frames.

Given a set of ten metaphors, chosen according to the presence of information, the relations between them,

```
GOVERNMENT_INSTITUTION_IS_A_BUILDING,
GOVERNMENT_IS_A_PERSON,
GOVERNMENT_IS_AN_ORGANISM,
GOVERNING_ACTION_IS_MOTION,
```

⁹Code is available at <https://github.com/dersuchendee/frame-based-metaphors/tree/main>.

GOVERNMENT_INSTITUTION_IS_A_PHYSICAL_STRUCTURE,
 CAUSED_CHANGE_OF_STATE_IS_CAUSED_CHANGE_OF_LOCATION,
 ANALYZING_IS_DISSECTING,
 MACHINES_ARE_PEOPLE,
 CHANGE_IN_CONTROLLER_IS_TRANSFER_OF_POSSESSION,
 DISEASE_TREATMENT_IS_WAR

we retrieved their source–target frames, any available role correspondences, and entailment texts, then computed, per source–target pair: the count and list of shared frame elements and shared synset labels. Finally, we typed frames as source or target and to expand near-equivalences before scoring (`closeMatch`, `subsumedUnder`).

5.3.2.2 Results and discussion

Across the selected metaphors, we observe heterogeneous surface overlap. For example, *AFFECTED MOTION* and *LEADERSHIP* share multiple frame elements such as *Place*, *Type*, *Role*, *Governed*, *Leader*, and *Descriptor*, while *PERSON IS LEADERSHIP* reduces to a single shared element (*Descriptor*). Likewise, *ACTION IS CAUSATION* shows overlap on *Explanation*, *Frequency*, *Manner*, *Place*, *Means*, and *Time*. At the same time, canonical pairs like *DISEASE TREATMENT IS WAR* and *MACHINES ARE PEOPLE* may share few or no synset labels or frame elements, despite being intuitively plausible.

These outcomes confirm two points. First, productive metaphors often recruit distant domains, so lexical and frame-element overlap alone is not a reliable signal of blending. Second, even when surface overlap exists, it does not guarantee that role–filler constraints align across frames. In the literature, the works confirm this claim (see Section 3.0.3). *Amnestic Forgery* (Gangemi, Alam, and Presutti, 2018) addresses plausibility by rejecting blends that violate role–filler constraints or inherit mutually incompatible attributes. By distinguishing conservative from non-conservative compositions, it can flag implausible mixes before propagation. Nevertheless, some implausible blends persist. Likewise, the MET^{CL} system by Lieto, Pozzato, and Zoia (2025) reports coverage limitations and gives only implicit treatment to the blending property (essentially as a weighted ensemble).

When a domain is sparsely represented, the system cannot compute a credible blending property, limiting scalability beyond fixed frame inventories such as *Framester* or *ConceptNet*. Furthermore, as we have seen with the dominance hypothesis (§ 2.2.2.1), as the source warps the target by infusing some features, it can be that this process might not be represented in the knowledge base.

In practice, we find that identifying a robust blending property requires integrating relational constraints from mappings and roles with distributional evidence. This motivates the embedding-based and LLM-assisted variants we present next, which combine learned priors with KB structure rather than relying on structure alone.

5.3.3 Embedding-based extraction of the blending property

A possible and popular approach to extract common structure among two concepts is word embeddings. Embedding-based methods (see Section 3.0.4.2) have been successfully applied to various metaphor-related tasks, particularly in identifying

metaphorical word usage through vector space similarities. These approaches typically rely on the distributional hypothesis, i.e., that words appearing in similar contexts have related meanings, and use metrics like cosine similarity to identify semantic deviations that might indicate metaphorical usage. However, these standard distributional approaches face fundamental limitations when applied to frame-based metaphor detection. First, cosine similarity between embeddings fails to capture the asymmetric nature of metaphorical mappings, where source and target domains play distinct roles. Second, two domains may be distributionally similar without forming plausible metaphors, or conversely, distributionally distant yet form highly conventional metaphors. For instance, as shown in the previous section, CRIME and DISEASE may not show high distributional similarity, yet CRIME IS A DISEASE is a common metaphor. Third, standard embeddings lack the cognitive grounding necessary to explain why certain cross-domain mappings feel intuitive while others do not.

A solution to these issues, we propose, is a novel method based on cognitive embeddings. Cognitive embeddings are vector representations that encode words or concepts according to neurologically-grounded semantic dimensions. There are three reasons why they could be beneficial. First of all, distributional embeddings do not particularly capture the difference between two domains in the sense that two domains can be different enough but also not in order to form a metaphor; secondly, explainability is an added value when working on metaphor. Third, we have seen that dominance is an important factor in metaphor (§ 2.2.2.1), and embeddings can operationalize it.

In this experiment, we use cognitive embeddings to test two hypotheses¹⁰. H1 (*Plausibility window*): metaphorically plausible frame pairs cluster within an intermediate band of cognitive similarity, very dissimilar pairs are unlikely to be metaphors, and very similar pairs are often literal or metonymic. H2 (*Directional dominance*): within this intermediate band, high-to-low transitions from source to target contribute more to divergence than low-to-high transitions, such that the share of high-to-low crossings is a stronger predictor of D_{KL} than the share of low-to-high¹¹.

5.3.3.1 Methodology

Training data selection. The proposed cognitive embeddings are trained on the dataset produced by Binder et al. (2016), which represents one of the most comprehensive neurologically-grounded collections of semantic feature norms available for psycholinguistic research.

This dataset provides ratings for 65 distinct semantic dimensions across 535 concrete concept words, with each dimension corresponding to specific brain regions identified through neuroimaging studies of semantic processing. It is grounded in the hub-and-spoke model of semantic cognition, which proposes that conceptual knowledge emerges from the integration of modality-specific information across different brain regions. Unlike traditional semantic feature databases that rely on linguistically-derived categories, the Binder norms are explicitly designed to reflect the neural architecture of semantic representation in the human brain. Participants (N = 20-30 per dimension) provided ratings on a 7-point Likert scale indicating how

¹⁰Code and material are available in the Github repository <https://github.com/dersuchendee/cognitive-embeddings>

¹¹Our directional penalties (hi-to-lo > lo-to-hi) are also consonant with Tversky's Contrast Model (Tversky, 1977), where asymmetry follows from unequal weights on source- vs. target-unique features and from reference selection.

strongly each concept was associated with specific perceptual, motor, cognitive, and emotional experience selected based on neuroimaging meta-analyses. For example, participants rated how much a concept involves “auditory experience” (auditory), “manipulation with hands” (haptic), or “social interaction” (socialness).

Apart from having been used in similar tasks, the Binder dataset offers several advantages for cognitive vector modeling: (1) neurological grounding provides theoretical motivation for dimensional structure, (2) fine-grained dimensionality allows for detailed analysis of semantic relationships, (3) continuous ratings capture graded rather than binary semantic features, and (4) comprehensive coverage includes both concrete and abstract semantic properties, making it suitable for studying diverse conceptual domains including metaphorical mappings.

We thus start from the assumption that metaphors can combine because of the commonality of a specific feature or norm: for instance, in the metaphor CRIME IS A DISEASE, both domains presuppose something harmful, but also presuppose a societal dimension, and the connections between these aspects cannot be completely captured by average embeddings.

Model training. A series of autoencoder architectures were systematically evaluated to establish the optimal framework for mapping word embeddings to cognitive attribute predictions.

Three distinct autoencoder variants were implemented and trained: one trained using the common mean squared error (MSE) loss or cognitive reconstruction; one trained using cross entropy loss by considering the cognitive scores as a multinomial probability distribution; and finally a Variational Autoencoder where hidden latent variables are explicitly learned. Each architecture maintained encoders and decoders for word embeddings and cognitive attributes, constrained to learn a shared latent representation through cross-modal alignment objectives.

Multiple model configurations were trained with varying hyperparameters, such as number of hidden layers and hidden dimension. The autoencoder trained using MSE was ultimately selected as the optimal architecture based on empirical performance metrics. This selection reflects the superior capacity of MSE-based loss functions to capture the continuous nature of Likert-scale cognitive ratings compared to the probabilistic framing imposed by the other approaches. For the base word embeddings, we selected fastText (Bojanowski et al., 2016) over alternatives like Word2Vec or BERT embeddings. This choice was motivated by fastText’s subword tokenization approach, which enables representation of out-of-vocabulary terms through character n-grams—critical for handling the diverse lexical realizations of frames in Framester. While contextualized embeddings like BERT could provide richer representations, they would require sentential context that is not available when working with isolated frame labels. Since our analysis operates at the frame level rather than in sentential contexts, static embeddings that can reliably represent single lexical items or frame names were more appropriate for this application.

Evaluation setup. After training the model, it has been explored again on the metaphors present in Framester by constructing a pairwise alignment space between candidate source and target frames and quantifying their distributional divergence. Concretely, we loaded a table of frame-level scores and parsed them into numeric vectors for sources and targets. Treating each score vector as a discrete probability distribution, we computed the asymmetric Kullback–Leibler divergence for every observed source–target pair.

We employ KL divergence rather than symmetric measures for three reasons. First, metaphorical mapping is inherently asymmetric and KL divergence naturally captures this directionality. Second, KL divergence quantifies the information loss when approximating the target frame’s cognitive profile using the source frame’s profile, operationalizing the cognitive effort required to understand the target through the source lens. Third, this measure aligns with theories of metaphorical processing where the source domain structure is projected onto the target domain, making the source distribution the reference against which the target is evaluated.

To explore the full combinations of potential metaphorical mappings, we performed a Cartesian product of the dataset with itself, and used KL divergence to measure how much the cognitive profile of each frame, possibly a source, differs from another frame, possibly the target. Higher divergence indicates greater cognitive distance between the frames.

The resulting augmented matrix of divergences over the Framester frame space provides a dense landscape of cross-frame dissimilarities that could then be mined. This procedure operationalizes “metaphor exploration” as probabilistic divergence over frame-conditioned score distributions, enabling systematic, reproducible assessment of the model’s preferences for particular mappings within Framester.

In addition, we visualized the resulting divergences to diagnose the model’s mapping behaviour. For a fixed target frame (e.g., *Idea*), we ranked all candidate source frames by increasing $D_{KL}(p||q)$ and produced a line/point plot where the x -axis reports the KL divergence and the y -axis lists source frames. Points associated with non-empty, matching metaphoric labels were highlighted to surface possible alignments. In the visualization, a target can be searched to find possible sources.

In this context, we privilege a qualitative evaluation precisely for the fact that we cannot assess plausibility without close reading. We thus chose 10 exemplary metaphors for this evaluation, the same ones we identified in §5.3.2.

Additionally, building on these pairwise divergences, we introduced a conceptual movement analysis to quantify how many psycholinguistic dimensions undergo extreme sign reversals between source and target. All dimensions were standardized by z -scoring over the union of source and target values. For each source–target pair and each dimension, we then assigned a ternary category: *high* if $z \geq 1$, *low* if $z \leq -1$, and *neutral* otherwise. A hi-to-lo event is recorded when the source is high and the target is low on the same dimension; lo-to-hi is the converse. The total number of crossings is $k = \text{cnt\_hi\_lo} + \text{cnt\_lo\_hi}$, which we relate to the observed divergence.

Consider, for example, the conventional expression “I sank into thought”, which instantiates the frame mapping *Downward_motion* (source domain)-to-*Thinking_about* (target domain). This metaphoric pair has $k = 4$ with one hi-to-lo and three lo-to-hi transitions: there is one cognitive dimension on which *Downward_motion* is strongly activated (high) while *Thinking_about* is strongly suppressed (low), and three dimensions on which *Downward_motion* is suppressed while *Thinking_about* is activated. Intuitively, the single hi-to-lo transition marks a property that does not carry over cleanly from motion to cognition, whereas the three lo-to-hi transitions capture properties that become more salient on the target side of the metaphor. We refer to the asymmetry whereby hi-to-lo transitions contribute more to misalignment than lo-to-hi as the hi-to-lo dominance hypothesis.

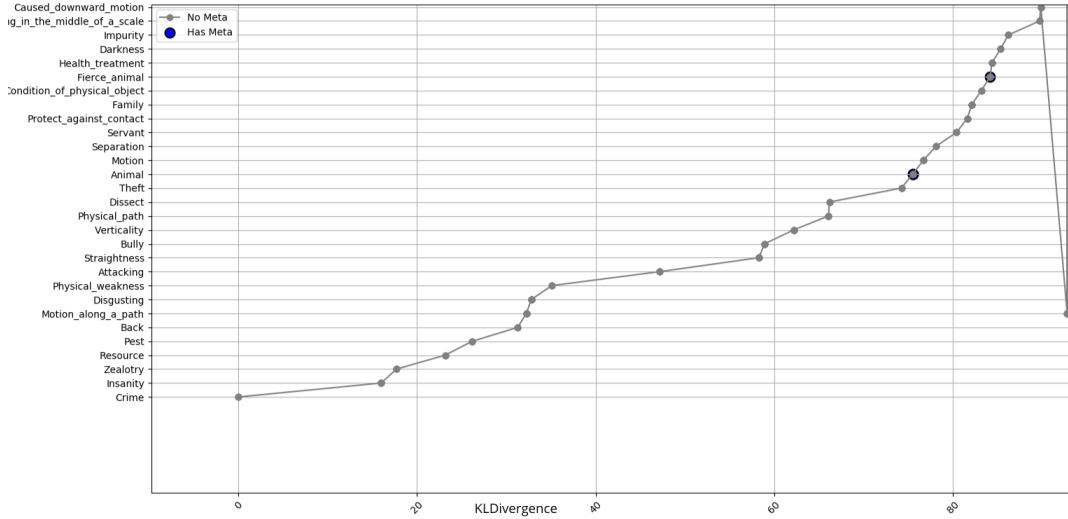


FIGURE 5.1: Example of KL divergence between source frames and the target frame CRIME. Gray markers trace all candidates; blue markers highlight pairs with non-empty, matching metaphoric labels.

5.3.3.2 Results and discussion

After the close reading analysis, we concluded that cognitive embeddings alone are not enough to assess metaphorical plausibility or operationalize the Blending property. For instance, for the target IDEAS, known sources are FOOD, POSSESSION, OBJECT. But this method also finds that an idea is MAGIC or that is ADDICTION or a RESOURCE, which is plausible. Likewise, as shown in a snapshot in Figure 5.1, CRIME is already associated with animal, but the tool discovers INSANITY, ZEALOTRY, PEST, BULLY and other frames that could be used in real world. At the same time, implausible words remain. Health treatment associated to crime, for instance, or servant.

This is also due to the fact that the number of synsets associated with frames varies. Therefore they can skew the results towards being more or less dissimilar.

For what concerns dominance, the results are the following. The global pattern is monotonic: average KL increases as k grows. Mean KL rises from 114.3 at $k = 0$ to 145.1 at $k = 1$, to 177.1 at $k = 2-3$, to 196.9 at $k = 4-7$, and to 201.4 for $k \geq 8$. The association is confirmed by a Spearman correlation of $r = 0.334$ ($p \approx 0$) between k and KL. The effect is directionally asymmetric: hi-to-lo is strongly associated with higher KL ($r = 0.479$, $p \approx 0$), whereas lo-to-hi has only a weak relationship ($r = 0.063$, $p \approx 10^{-92}$), suggesting that mappings from strongly activated source dimensions into suppressed target dimensions are particularly discordant, while the reverse is comparatively forgiving.

Illustrative examples are consistent with these aggregates. The pair SELF PROPELLED MOTION and BUREAUCRACY shows $k = 10$ with all ten crossings of the hi-to-lo type and a very large divergence, exemplifying an implausible mapping in this representation. By contrast, BACK and NATION exhibits many crossings ($k = 13$) that are predominantly lo-to-hi (1/12) yet maintains a modest KL, which is compatible with a conventional BODY IS INSTITUTION reading (e.g., backbone of the nation). Further low-divergence cases at moderate k likewise favor lo-to-hi-heavy patterns, for example DOWNWARD MOTION-to-THINKING ABOUT ($k = 4$) and SERVANT with ACTION ($k = 6$), consistent with familiar figurative paraphrases (sink into thought, actions serve ends). Even in the high-movement regime, plausible conventional metaphors persist when lo-to-hi dominates, as in BEING AT A HIGH LOCATION and

LEADERSHIP ($k = 8$, $KL \approx 20.65$), instantiating the POWER IS UP schema, and INCREASE_SIZE-to-NATION ($k = 9$), reflecting growth metaphors. We also observe in known metaphors (e.g. explicitly present as such on Framester) that this subset reproduces the same qualitative tendencies while exhibiting greater variance owing to smaller sample sizes in extreme k bins.

A structurally close (largely literal/metonymic) ($k = 0$) pair is RESOURCE IS CORPORATION. A sentence like “the corporation allocated resources to R&D” is not figurative: it states an organizational relation. The zero-crossing count in the pair tells us that no dimension sits at opposite extremes across the pair. Their low KL is thus explained by (too much) shared or compatible dimensions (e.g., economic/organizational salience) and by the absence of hi-to-lo or lo-to-hi flips, which is exactly what we would expect for a close, mostly literal relation.

Taken together, the results of this experiment suggest that cognitively grounded embeddings provide a useful lens on metaphorical mapping, but similarity in this space cannot represent, by itself, a decision rule for (plausible) metaphoricity. Very dissimilar pairs rarely yield plausible metaphors, whereas very similar pairs could license literal or metonymic readings, but this is not a given.

Still, the most interesting candidates tend to occupy an intermediate band where the proximity between proximity is neither too small nor too large. Within this band, the dominance pattern appears robust: increases in the number of extreme threshold crossings (k) are associated with higher D_{KL} , and this association is driven primarily by high-to-low transitions from source to target, which inflate divergence more than low-to-high transitions. Known metaphors annotated in the knowledge base reproduce the same qualitative tendencies, albeit with greater variance due to smaller sample sizes.

These findings recommend modeling metaphor plausibility as a joint function of graded similarity (e.g., D_{KL}) and directional crossing profiles (the mix of high-to-low versus low-to-high), rather than as a threshold on similarity alone, though this might still not be enough to deem plausibility given the need to situate metaphor in context.

An important limitation is that $k=0$ does not entail that a pair is metaphorical. By construction, $k=0$ only states that no dimension sits at opposite extremes (there are no high-to-low or low-to-high crossings); it is compatible with literal relations (e.g., RESOURCE-to-CORPORATION), with conventional metaphors that share aligned saliences (high-high or low-low), and with pairs that are largely neutral. Thus k is an indicator of movement, not of metaphoricity. Likewise, low D_{KL} signals structural closeness in the cognitive space but is neither necessary nor sufficient for a metaphorical reading; plausibility remains context-sensitive and often requires close reading. Two additional constraints qualify the evidence: coverage in the knowledge base is by definition always in progress, which introduces both false negatives (missing plausible metaphors) and false positives (idiosyncratic or polysemous entries), and frame-synset imbalance can bias divergences whenever a frame aggregates very many or very few lexical realizations. These factors should temper strong causal claims from the present analysis.

Methodologically, the results argue for a way towards the dynamic extraction of the blending property, that can be context-sensitive and that can also not rely necessarily and uniquely on a knowledge base.

5.3.4 LLM-based extraction of the blending property

In this section, the extraction of the implicit blending property accounting for the overarching structure of a metaphor begins with the assumption that LLMs can join distant parts of our world model, and with the idea that it would be interesting to understand to what extent these tools adhere to our own ideas about the world and conceptualization inner mechanisms¹². In that, LLMs could fill in the gaps seen in previous efforts to computationally extract the blending property. We probe this capacity by asking a *GPT-4* model queried through the API to propose candidate key concepts and role correspondences for conceptual metaphors contained in Framester and then validating those concepts with human rating. In particular, we test two hypotheses. H1 (Blending-dominance correlation): There is a systematic correlation between the fitness of blending properties for specific concepts and their roles as source or target frames, reflecting the inherent dominance asymmetry in metaphorical mappings. Specifically, we hypothesize that LLM-generated blending concepts will show differential fitness patterns depending on whether they bridge from source to target or target to source frames. H2 (Bidirectional metaphor completion): LLMs can not only determine the directionality of metaphorical mappings between frame pairs, but also generate plausible blending concepts that mediate these mappings. This predictive capability should enable the systematic identification of missing conceptual bridges in the Framester knowledge base, contributing to its completion through LLM-driven metaphor discovery.

Code and survey data are available on Github¹³.

5.3.4.1 Methodology

Starting from the Framester knowledge base, we use an LLM to propose up to ten key concepts per known metaphor together with role correspondences between source and target frames. This is done in two steps basing on the availability of frame roles for frames.

In the case the roles were present for both frames, one prompt type was proposed for direct extraction of the roles, otherwise a context engineering technique based on MP (see Section 4.2.2.4) encouraged the model to complete missing roles before giving an output. Outputs were then post-processed to keep concepts in the requested format and to remove duplicates and function words; a second pass with the model reduced each list to exactly ten distinct, filtered-out items. With an exploratory survey, users are required to assess the blending property's closeness to the source or target concept given.

LLM selection and use. We used *GPT-4* in the API (model: *gpt-4-1066-preview*, temperature 0) as it provided a good balance between state-of-the-art and LLM costs, and it being used in similar studies. Because Framester contains both fully specified mappings (170 metaphors with roles) and underspecified ones (431 metaphors with partial roles), we used two closely related prompts, as shown in the Appendix. The input always included the metaphor's source and target frames, any available roles, and a worked example to anchor the level of abstraction (see the C.1 Appendix for the prompt).

¹²There are now many experiments on the cognitive adherence of LLMs to our brain mechanisms, for example (Cui, Li, and Zhou, 2025; Zhou et al., 2024a).

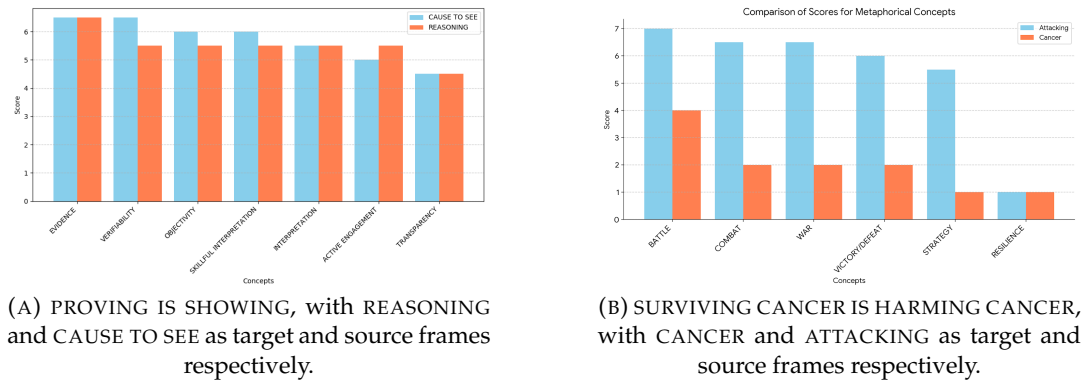
¹³<https://github.com/dersuchendee/implicit-grounding-metaphors>.

Survey setup. The source and target frames of the metaphors were separated and included apart and in random order within a survey on Qualtrics sent to University of Bologna members from student to researcher levels, with the same set of key concepts. The survey was sent to both English native speakers and speakers of other languages, for a total of 90 respondents¹⁴, for validation of the key concepts on a Likert scale from 1 to 7, assessing each concept’s importance in understanding the source or target frame. Evaluators were also asked to provide a one-sentence explanation of the relationship between the two domains in a metaphor. As a result, both source and target domains from 128 metaphors and 66 metaphor paraphrases were scored. While this sample size limits statistical power, it provides initial evidence for the feasibility of LLM-assisted blending property extraction and identifies patterns warranting larger-scale validation. We then calculated average score for each concept and determined the p-value to see the correlation for the relationships between the scores of each metaphor’s domains.

5.3.4.2 Results and discussion

Results indicate a slightly negative Pearson correlation (-0.02) across all metaphor domains. The percentage of strongly correlated domains in a metaphor ($r > 0.5$) is 15.20%, while negatively correlated domains constitute 16%.

FIGURE 5.2: Comparison of metaphorical scores for two metaphorical domains analyzed in the study



For example, in Fig. 5.2a, we can see the key concepts average scores for PROVING IS SHOWING metaphor. On the other hand, Fig. 5.2b shows the negatively correlated scores for the metaphor SURVIVING CANCER IS HARMING CANCER.

Upon reviewing the results, we can conclude that again, there is a middle ground for dominance given the correlation percentages. In some cases, key frames for both domains of a metaphors were evaluated similarly and given the same scores, indicating they were both crucial to understanding the metaphor. In other cases, this consistency was not present.

The observed negative-correlation subset naturally admits a directional reading: many of these are precisely the cases where concepts emphasized on the source side are de-emphasized on the target side (and vice versa), producing an imbalance that readers described in their paraphrases as infelicitous or strained. This qualitative fit is consistent with, though it does not by itself prove, a dominance asymmetry akin

¹⁴Where the demographic and non-mandatory question of English language proficiency has been answered, responders said they were: 7 Native Speakers, 22 Proficient speakers, 31 Advanced speakers, and 10 Intermediate speakers.

to the one quantified in Section 5.3.3: mappings appear less plausible when strongly activated source cues would have to map onto suppressed target cues. Conversely, when the concept set elevates target-side saliences without requiring strong downshifts from the source, respondents generally produced coherent explanations.

Interpreted through the lens of plausibility, the observed distribution is consistent with the earlier plausibility window intuition: many plausible metaphors appear to land in a middle ground where source and target give comparable weight to at least some of the same concepts (moderate, positive alignment), while clearly literal/metonymic relations would be expected to rate as very similar across domains and clearly incoherent pairings to diverge. There are two possible reasons for such a discrepancy: i) the stronger influence, or dominance, of the source on the target with respect to these metaphors, as seen in the case of the metaphor SURVIVING CANCER IS HARMING CANCER, or ii) the inability of the language model to provide effective key concepts to account of the metaphor. For what concerns this second case, for instance, in the metaphor CRIME IS AN INFECTION, high scores are given to concepts like *negative phenomena*, *damage*, and *prevention*, but the concept of SPREAD, defined as a gold standard among one or two examples, which along with the negative polarity is crucial for understanding the metaphor, was absent. This is not a failure of the plausibility idea so much as a reminder that plausibility depends jointly on which concepts are present and how their salience distributes across domains, which also has to be analyzed in context.

Furthermore, the context learning technique employed in this experiment follows the representation model of the Blending Ontology in natural language, which calls for a better heuristic method to extract it. This, however, also shows we would need a wider gold standard dataset for the annotation of the blending property before proceeding with survey. However, the construction of such dataset, which tends also to be quite subjective, is a problem in itself, also this dataset should have more context but this so far cannot exist, in turn raising more questions.

5.3.4.3 Toward a Neurosymbolic Approach for Metaphor

Our systematic exploration across three computational paradigms provides strong empirical support for the complementarity hypothesis. Each approach reveals distinct capabilities and limitations that, when considered together, motivate a neurosymbolic integration strategy.

Empirical validation of the complementarity thesis. Frame-based approaches suffered from critical coverage gaps, proving unable to process metaphors absent from existing knowledge bases or discover novel blending properties. The structural richness of knowledge graphs provides interpretable representations and enforces compositional constraints, yet their manual construction limits scalability and their static nature cannot accommodate the dynamic, contextual aspects of metaphorical reasoning. Embedding-based methods revealed systematic patterns such as plausibility windows and dominance asymmetries that capture important distributional regularities in metaphorical mappings. However, these approaches proved insufficient as standalone decision rules due to their inability to handle compositional reasoning and contextual sensitivity. While embeddings effectively quantify semantic similarity, they lack the structured representations necessary for explaining why certain mappings succeed while others fail. LLM-based approaches showed remarkable flexibility in generating human-interpretable candidates and adapting to novel contexts, yet they required representational scaffolding to achieve consistency and

reliability. Their probabilistic nature, while enabling creativity, introduces unpredictability that undermines systematic metaphor processing.

Integration strategy and LAG architecture design. Rather than treating these approaches as competing alternatives, LAG (§4.2.3.1) operationalizes complementary components within a unified framework. The architecture primarily integrates structural knowledge from frames with the flexible reasoning capabilities of LLMs. Frame-based knowledge graphs provide stable semantic structures and compositional rules through the Blending Ontology. LLM reasoning bridges coverage gaps inherent in manually constructed knowledge bases, generating missing conceptual links while remaining constrained by formal ontological heuristics. This design preserves the contextual adaptability that makes LLMs powerful while grounding their outputs in interpretable symbolic representations.

Importantly, while our experiments revealed the value of cognitively-grounded embedding patterns for understanding metaphorical plausibility, the LAG framework does not explicitly integrate a separate cognitive embedding component. Our cognitive embedding experiments demonstrated systematic patterns such as plausibility windows and directional dominance effects that provide valuable insights into the cognitive constraints governing metaphorical mappings. However, these findings primarily inform our understanding of metaphor evaluation and analysis: these insights complement but do not replace the structural scaffolding provided by knowledge graphs and the flexible reasoning capabilities of LLMs within the LAG architecture.

While standard distributional semantics may be implicitly encoded in LLM pre-training, the specific neurologically-grounded dimensions captured by cognitive embeddings offer a distinct analytical lens that could inform future developments in metaphor evaluation metrics and computational cognitive modeling. The structured knowledge provides stability and interpretability, while LLM reasoning enables flexible adaptation to novel contexts while remaining anchored in formal representations. This integration addresses the fundamental limitations identified in each isolated approach while maximizing their complementary strengths.

5.4 Recognising: a LAG approach to computational metaphor processing

5.4.1 Motivation

As we have seen in the previous chapters, metaphor is the most demanding form of analogical reasoning because it requires not only structural mapping, but also the creative blending of distinct conceptual domains (Wijesiriwardene et al., 2023). Current state-of-the-art methods such as LLMs struggle with this task (§3.1.1.3).

Therefore, despite repeated calls to enrich AI systems with deeper semantic and cognitive representations of the world, rather than mere word-level correlations, analogical reasoning, especially in the form of metaphor, remains a major open challenge for Natural Language Processing. Progress is further hampered by the absence of a shared problem definition for metaphor and by the scarcity of diverse, well-curated datasets and evaluation protocols, which we have seen, constitutes a metaphor crisis (see Section 3.1).

Since LLMs' performance tends to improve significantly with appropriate structured background material, and since we have seen frame-based methods, embedding-based methods and LLM-only-based methods have inherent limitations (§ 5.3), the combination of reliable but labor-intensive semantic technologies and quicker, context dependent, yet error-prone LLMs, as shown by LAG in Gangemi and Nuzzolese (2025) (§ 4.2.3.1), presents a new avenue for metaphor understanding and analogical reasoning in general.

Following this framework, we saw in Chapter 4 that traditional ontology engineering methodologies, while providing robust pattern-based workflows, struggled with scalability limitations and manual bottlenecks in requirements engineering and module implementation. The emergence of LLM-assisted approaches like Ontogenia demonstrated that computational methods could achieve performance comparable to novice ontology engineers, yet systematic evaluation remained a critical challenge. Chapter 4's exploration of context engineering techniques, multi-dimensional evaluation frameworks, and the CoLLM consistency assessment established the methodological foundation necessary for the comprehensive four-layer evaluation approach employed here.

In previous chapters, we have also discussed the need to draw a consensus on what metaphor understanding really means and how can it be operationalized computationally. In the framework, as shown in Figure 5.3, detection gives then space to understanding in the form of source and target domain identification, and reasoning, in the form of blending property extraction. Generalization completes this framework, as metaphor analysis involves the need to reason on different domains and modalities for it to be effective.

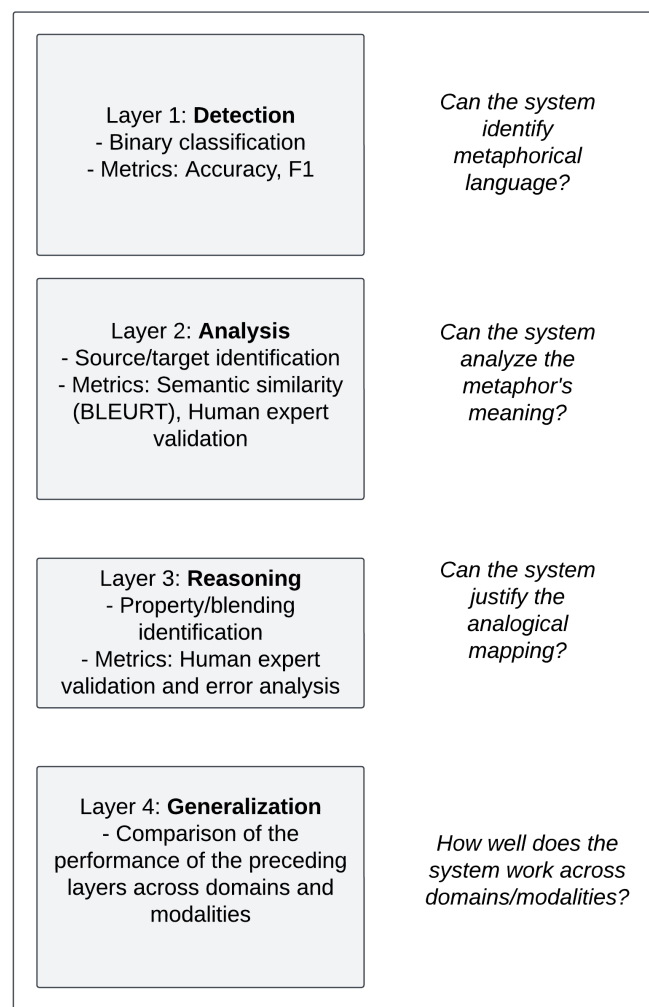
In this section, we test metaphor understanding across these four layers with a CBT-based LAG pipeline to capture the semantic implicit relationships underlying metaphors from unlabelled multimodal data. Integrating structured, machine-interpretable semantics through KGs with fast and dynamic LLMs systems allows the real-time flexible understanding and representation of the implied aspects of metaphor in data across various domains and modalities such as news, visual advertisements, memes, and medical sciences.

The pipeline projects multimodal data into a shared natural-language space; they are then merged into a Semantic Knowledge Graph (feature-level fusion in the JDL sense, i.e., combining low-level attributes before object formation, Hall and Llinas, 1997), and Logic-Augmented Generation injects LLM-derived triples to create an Extended KG (XKG).

5.4.2 Methodology

The proposed method is based on the general approach described in Section 4.2.3.1 and aims to enhance the detection and understanding of metaphor by establishing a standardized and explainable model for the semantic representation of implicit natural language. In this way, the tacit *know-how* knowledge, following Polanyi's definitions, can partly become *know-that* (Polanyi, 2009). Our approach has two main components: i) the representation of inputs, regardless of their modality, into natural language, ii) their conversion into a SKG with the Text2AMR2FRED knowledge extractor (Gangemi et al., 2023) and iii) the SKG enhancement driven by CBT (See Section 3.2) applied to metaphor.

FIGURE 5.3: Four-layer evaluation framework for metaphor processing. Layer 1 (§ 5.4.3.1) tests **detection**, a binary decision of whether an expression is metaphorical, using three text-based benchmark datasets in different domains and reporting accuracy and F1. Layer 2 (§5.4.3.1) probes **understanding**: the system must identify source and target domains; results are scored on two datasets of different domains according to BLEURT similarity and manual expert judgments. Layer 3 (§ 5.4.3.1) addresses **reasoning**, which requires the model to recover the blending property that justifies the mapping, in particular, this is done on an annotated visual metaphors dataset; evaluation relies on expert analysis. Layer 4 aggregates the preceding scores to assess **generalization**: how consistently the same architecture performs when those tasks are posed in different linguistic domains and across text-versus-image modalities. Experts complement this layer with an error analysis and an evaluation of the generated graphs' quality (§ 5.4.3.3).



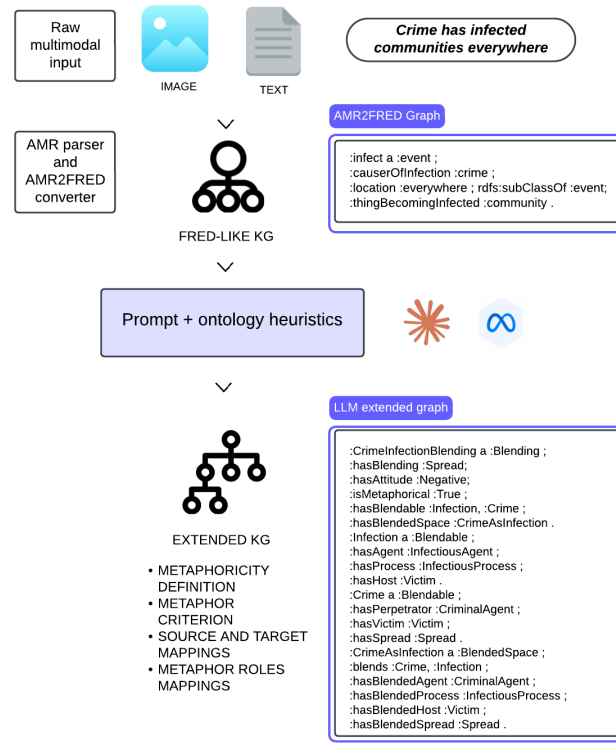


FIGURE 5.4: Proposed LAG-based pipeline for the sentence "Crime has infected communities everywhere".

5.4.2.1 Automated multimodal representation with Text2AMR2FRED

Text2AMR2FRED has been first introduced as a revised architecture of FRED's text-to-KG pipeline (Gangemi et al., 2023), using state-of-the-art Abstract Meaning Representation (AMR) parser with multilingual capabilities. In this way, relevant commonsense knowledge from knowledge bases linked by the Framester knowledge base (Gangemi et al., 2016) is retrieved and represented in a graph.

For example, as shown in Fig. 5.4, the sentence "Crime has infected communities everywhere" is first parsed into an AMR graph, using SPRING. The AMR graph is then converted into an RDF/OWL knowledge graph that follows FRED's knowledge representation patterns using AMR2FRED.

AMR identifies the cause of the infection in the sentence, which is, however, not a germ, but crime and the thing that gets infected (community) everywhere. In the FRED-like KG, this graph is enriched with knowledge from Framester such as frames, WordNet synsets and PropBank roles, along with alignment to the DOLCE foundational ontology (Gangemi et al., 2002). However, there is no representation of metaphorical knowledge yet at this stage. While a conceptual metaphor KG sourced from Metanet is contained in Framester, it is still necessary to be able to detect figurative language in any type of data. An image can also be used by being automatically described in natural language. In this case, language is considered an amodal way to bridge modalities in the resulting SKG (Gangemi and Nuzzolese, 2025). The SKG

can then be extended with implicit contextual knowledge with the support of an LLM (De Giorgis, Gangemi, and Russo, 2025).

5.4.2.2 Logic Augmented Generation with CBT-inspired heuristics

As shown in Fig. 5.4, the result of the framework that generates XKG triples is the extension of the SKG with new classes and properties according to a specific ontology in the prompt functioning as heuristics. In this work, we adopt the Blending Ontology for this purpose.

5.4.2.3 The Blending Ontology

Throughout the rest of the paper, we refer to the Blending Ontology as the overarching conceptual framework that guides our reasoning (detailed in Section 5.2.1). This ontology provides structured heuristics for LLM-based processing. Leveraging this framework, we design prompts that enable an LLM to take a given text along with its knowledge graph representation (generated via Text2AMR2FRED) and extend it into an augmented RDF graph in Turtle format.

5.4.3 Application and evaluation

In this section, we evaluate the LAG-based framework’s ability to extract implicit meaning by applying it to metaphor detection and understanding tasks. We first use three metaphor detection datasets to determine the best-performing configuration among several baselines. This optimal configuration is then applied to three metaphor understanding datasets.

5.4.3.1 Tasks

This section describes the application of the framework in three metaphor-related tasks: i) *metaphor detection*, where a SKG derived from a sentence in natural language is enhanced to determine if a sentence is metaphorical, sometimes with a target word as a reference; ii) *conceptual metaphor understanding*, which additionally aims to identify the correct source and target domains that make up the metaphor, and iii) *visual metaphor understanding*, where the aim is to identify the property linking the two domains in the visual metaphor.

Metaphor detection and understanding. In this task, the LLM is given a sentence, a SKG representing the sentence derived from Text2AMR2FRED, and a series of instructions. The prompt first requires identifying the possible source and target of the sentence; then, it is asked to define these two as conceptual frames containing roles and following a blending function that orchestrates the creation of a blended space. Following the Blending Ontology’s alignment to the Cognitive Perspectivization one (See Section 5.2.1), a Lens and an Attitude further refine the definition of the metaphor. An example of the output graph for a specific sentence is also provided.

Given the Text2AMR2FRED graph, and the presumed source and target of the sentence given a target word when available, along with blending-derived instructions, the LLM is prompted to analyze the sentence and to use at least these specific ontology elements according to the given definitions, to determine the possible metaphoricity of a sentence: `bl:Blending`, `bl:Blendable`, `bl:Blended`, `cp:Attitude`,

cp:Lens, and the datatype property `metanet:isMetaphorical` which the classification is based on. In the prompt, we also ask to determine blendable roles and their mapping to create a blended metaphorical space according to the Blending Ontology.

Visual metaphor understanding. Following the original study performed on human participants outlined in Petridis and Chilton (2019), the methodology for visual metaphor understanding involves the presentation of three examples and correct annotations of visual metaphor elements within an in-context learning technique, along with a new image to analyze. In this case, each image to analyze is first described in natural language through the LLM, and the text is passed to Text2AMR2FRED to generate a SKG. The SKG, image, accompanying sentence, and Blending Ontology heuristics are all included in the prompt with the three examples to generate the XKG, as shown in Fig. 5.5.

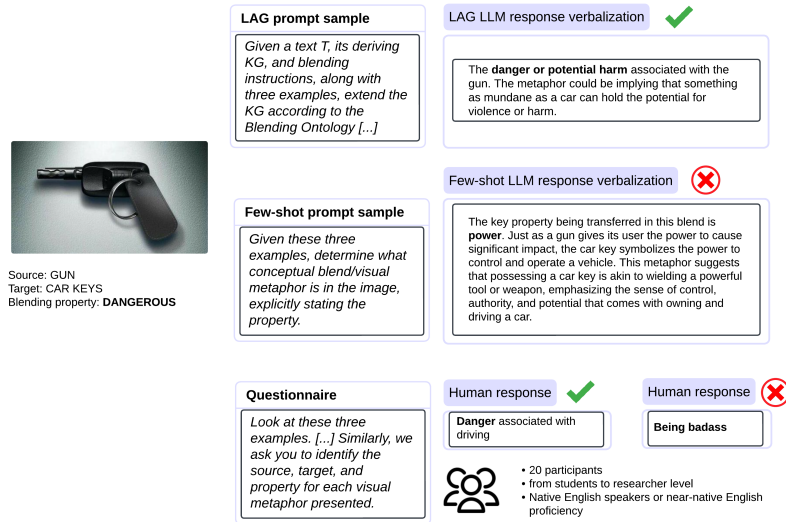


FIGURE 5.5: Example prompts and extracts of correct and wrong outputs according to manual evaluation for visual metaphor understanding.

5.4.3.2 Datasets

For the purpose of this study, and for comparison across the outlined tasks, for textual metaphor detection we conducted experiments on two publicly available metaphor datasets: (1) MOH-X (Mohammad, Shutova, and Turney, 2016), which comprises 647 sentences annotated as either metaphorical or literal; and (2) TroFi (Birke and Sarkar, 2006a), another metaphor detection dataset collected from the 1987-1989 Wall Street Journal Corpus. We sampled 300 balanced instances each for testing. For conceptual metaphors, we employ (3) the dataset refined by Wachowiak and Gromann (2023) of 447 sentences including conceptual metaphors annotated with source and target domain (for ease of reference, we will name it *WG*). We also introduced a fourth dataset to verify the capability to identify and explain domain-specific conceptual metaphors: (4) the Balanced Conceptual Metaphor Testing Dataset (BCMTD)

(see Section 5.1.2). This dataset consists of 147 sentences, equally distributed among general conceptual metaphors, scientific conceptual metaphors, and literal sentences. To stress-test external validity in scientific discourse, we further evaluate on the MMC (§5.1.1), which is imbalanced and contains borderline cases where metaphoricity is graded rather than clear-cut. MMC probes three pressures largely absent in the other datasets: domain specificity, label gradience, and skewed class ratios. The visual metaphors dataset is derived from Petridis et al. (Petridis and Chilton, 2019). It contains 48 images, 27 ads and 21 non-ads, and three examples, that have been given to human participants to test their ability to identify visual metaphors according to a conceptual blending framework. Table 5.6 shows statistics for all the datasets.

TABLE 5.6: Summary of datasets statistics, including instances, percentage of metaphorical sentences, and samples.

Dataset	# Instances	% Met.	# Samples
MOH-X	647	48.7	300
TroFi	3737	43.5	300
WG	447	100	447
MMC	792	51.4	792
BCMTD	49 (conceptual) 49 (scientific) 49 (VUA)	66.6	147
Visual metaphors	51	100	48

5.4.3.3 Evaluation methodology

The evaluation methodology for metaphor detection is both automated and manual depending on the type of task.

Automated evaluation metrics. We use Accuracy and F1 score to evaluate the model’s performance in metaphor detection. For conceptual metaphor understanding, evaluated via the correct identification of both source and target mappings according to a gold standard, the BLEURT score is used in order to evaluate the semantic distance between the candidate and reference source and target in the metaphor and considered to be successful if both achieve a positive score. This is an evaluation method similar to the one in Saakyan et al. (2025), where instead of only reporting label accuracy, it is reported according to defined thresholds. We also compute the correlation between the BLEURT score and human evaluation. In this case, the threshold is where the correctly predicted labels of source and target domains achieve a BLEURT positive score. For MMC, we follow the same analysis of the technical validation in Section 5.1.1. It contains graded or borderline labels. We pre-register a binary mapping for comparability: instances are marked metaphorical when the majority label indicates metaphor, otherwise literal. The same fixed decision rule and prompt templates used on BCMTD are applied unchanged to MMC.

Manual evaluation. To correctly determine metaphorical source and target mappings for domain identification, and to correctly compare gold standard and predicted blending properties for visual metaphor interpretation, it is also necessary to conduct a manual evaluation. The best performing solution for conceptual and visual metaphor understanding is therefore evaluated for source and target mappings

by three annotators. The annotators were graduated in linguistics, had proven expertise on metaphors, and had English Native or near-native language proficiency. They exhibited a Fair inter-annotator agreement (Fleiss' Kappa) in all cases, with an average on all tasks of 0.35, which is understandable for a task so subjective as metaphor understanding. To calculate the Accuracy, we compute an average of each Accuracy score.

In the visual metaphor understanding task, the manual evaluation compares the gold standard property with the predicted one connecting source and target. The evaluation directly follows the one in Petridis and Chilton (2019), which considered as correct responses directly related to the property in the image. We also repeated the study on a sample of 24 images on 20 participants of the University of Bologna, manually evaluating their responses as well. Participants ranged from student to researcher level, with native or near-native English fluency. This setup slightly differs from the original study, where participants were employed without specifying their study level. Participants' submissions were anonymous and informed consent was provided.

Furthermore, we provide a sample of 16 generated graphs, 8 derived from correctly identified metaphorical sentences, 8 derived from visual metaphors, to two knowledge engineers experts in conceptual blending and conceptual metaphor theories for an evaluation of the quality of the graphs. We evaluate the representation's adequacy, correctness, and completeness on a Likert scale from 1 to 7 concerning detected property, source, target and element mappings. Finally, one knowledge engineer manually evaluated errors (derived from negative majority voting) for conceptual metaphor understanding of the best performing LAG configuration. The errors follow the categorization proposed in Wachowiak and Gromann (2023) with the addition of a new category found, the *incorrect switching of source and target* (for a comprehensive account of these categories, see the Appendix, C.3.3).

5.4.3.4 Baseline methods

In Tian, Xu, and Mao (2024), the best-performing method for metaphor detection is a Theory Guided Scaffolding Instruction Framework based on Conceptual Metaphor Theory (TSI CMT). MetaPRO is a baseline among non-LLM methods for detecting source and target domains, serving as the state-of-the-art approach for identifying mappings without specifying a target word (Mao et al., 2023). For the MCC, we could not use MetaPRO, as it is temporarily closed. For conceptual metaphor understanding, no current baselines simultaneously detect source and target but only one of the two. We thus employ LLMs in different configurations, including zero-shot and few-shot, shown to be effective in previous research (Wachowiak and Gromann, 2023). No established baselines for visual metaphor understanding specify source, target, and blending properties. Therefore, we explore different configurations of the LAG framework using the same three few-shot examples used to test human participants in (Petridis and Chilton, 2019), and evaluate against the few-shot baseline without LAG.

5.4.3.5 Parameters setup

In our experiments, Claude 3.5 Sonnet¹⁵ serves as the primary LLM. We compare the best-performing results with Llama 3.1 70B-Instruct¹⁶ in all cases except for visual metaphor interpretation, where no current open models support few-shot image examples. Only the APIs are used for testing. We conducted tests using default parameters and set the temperature to 0 for greedy decoding.

5.4.3.6 Results

In this section, we outline the main results of the LAG framework applied to the selected tasks.

Metaphor detection: comparison with baselines. Table 5.7 shows comparative results of LAG compared with baselines for the sampled TroFi and MOH datasets. Table 5.8 shows comparative results of LAG with respect to the BCMTD. While BCMTD validates LAG’s competence on balanced data, the MMC presents a more challenging scenario: naturalistic scientific discourse where metaphoricity is gradient and instances are imbalanced.

TABLE 5.7: Performance comparison of various methods for metaphor detection on MOH-X and TroFi. Best performing results are in bold.

Method	MOH-X		TroFi	
	F1 (%)	Acc. (%)	F1 (%)	Acc.(%)
MetaPRO	84	81	79	70
TSI CMT*	82.5	82.9	66	66.8
LAG	89.7	87.3	89.7	84.6

TABLE 5.8: Performance metrics for the BCTMD Dataset for metaphor detection. Best performing results are in bold.

Method	Accuracy (%)	F1 Score (%)
LAG	80.1	84.1
MetaPRO	69.1	69.8
Few-Shot 12	59.0	48.9
Few-Shot 6	52.4	45.2
Few-Shot 3	47.5	42.8
Zero-shot	22.9	33.8

TABLE 5.9: Performance metrics for the MCC Dataset for metaphor detection. Best performing results are in bold.

Method	Accuracy (%)	F1 Score (%)
LAG	67	65
Few-Shot 12	69.41	74.88
Few-Shot 6	69.53	73.9
Few-Shot 3	69.65	75.52
Zero-shot	59.60	71.46

¹⁵<https://www.anthropic.com/news/claude-3-5-sonnet>

¹⁶<https://ai.meta.com/blog/meta-llama-3-1/>

Conceptual metaphor understanding. Conceptual metaphor understanding involves the correct identification of both source and target domains for a metaphor, which shows whether the system has understood it or not.

WG dataset. Manual evaluation reveals that the percentage of metaphors where source and target were correctly identified at the same time is 25.6%. BLEURT scores showed a moderate correlation between the human annotations and the predicted domains. For the target domain, a point-biserial correlation of 0.702 ($p < 0.001$) and a Spearman correlation of 0.698 indicate that BLEURT is reasonably sensitive to the human judgments. For the source domain, a point-biserial correlation of 0.707 ($p < 0.001$) and a Spearman correlation of 0.680 were observed.

BCTMD. Manual evaluation on the best performing method in metaphor understanding reveals that out of 96 sentences containing conceptual metaphors, the percentage of correctly identified source and target at the same time is 34.3%. Furthermore, the distribution of correctly identified metaphors is 51.6% for generic conceptual metaphors, while for scientific conceptual metaphors the score lowers to 8%. BLEURT scores showed a 36.9% positive correlation between reference and predicted source and target domains. A point-biserial correlation of 0.3525 ($p < 0.001$) and Spearman correlation of 0.2917 ($p < 0.001$) both indicate moderate positive correlations for BLEURT.

Visual metaphor Understanding For visual metaphor understanding, the identification of the blending property corresponds to the correct interpretation of the source and target employed in the metaphor (as shown by Petridis and Chilton (2019)). We experiment with five input “features”: (i) the Few-Shot examples, (ii) the sentence (*sent*), (iii) the image (*img*), (iv) the base graph (*kg*), and (v) blending heuristics (*blend*). In the table legends we indicate, in parentheses, how many of these features are included: *LAG (5f)* uses all five and outperforms a simpler prompt of Few-Shot (1f).

TABLE 5.10: Comparison of LAG (5 features: graph, blending heuristics, sentence, image, few-shot examples) for visual metaphor understanding against a Few-Shot baseline (no features).

Method	Accuracy (%)
LAG (5f)	65.0
Few-Shot (3) (1f)	54.7

TABLE 5.11: Ablation study on the visual metaphor dataset. In parenthesis, the number of features employed. Best result in bold.

Ablation	Acc.(%)
LAG: few-shot+kg+blend+img (4f)	67.0
LAG few-shot+kg+blend+sent (4f)	65.2
LAG few-shot+kg+img (3f)	68.6
LAG few-shot+blend+img (3f)	56.2

Metaphor detection and understanding: comparison with open-source LLM In Table 5.12, we compare the Claude-based best performing results of metaphor detection with the ones from Llama 3.1 70B-Instruct, showing Claude outperforms the open-source alternative.

For what concerns conceptual metaphor understanding, according to the evaluators, the average percentage of Accuracy is 37.5% against the method with Claude,

TABLE 5.12: Comparison of LAG for metaphor detection in each dataset with Claude and with Llama. Best results in bold.

LLM	MOH-X		TroFi		BCMTD	
	Acc.%	F1%	Acc.%	F1%	Acc.%	F1%
Claude	87.3	89.7	84.6	89.7	80.1	84.1
Llama	55.9	69	60.8	75	66.6	66.6

which yields an Accuracy of 34.3%. The results show a drop in performance when using Llama with metaphor detection, but metaphor understanding of both source and target mappings seems to perform slightly better with respect to Claude.

Ablation studies. We conduct the ablation study to evaluate the impact of the components in our method for metaphor detection by four configurations: (i) removing the blending instructions, and ii) removing the knowledge graph injection in the prompts. For visual metaphor understanding, we also ablate the insertion of a sentence describing the image to analyze (iii) and the image itself (iv). Table 5.13 shows the results of the ablation studies for metaphor detection for all the textual datasets, while Table 5.11 for visual metaphor understanding, removing one feature at a time.

TABLE 5.13: Ablation study results for MOH-X, TroFi, BCMTD, and MMC. Best results in **bold**.

Method	MOH-X		TroFi		BCMTD		MMC	
	Acc.%	F1%	Acc.%	F1%	Acc.%	F1%	Acc.%	F1%
LAG	87.3	89.7	84.6	89.7	80.1	84.1	67.0	65.0
No Blending	81.6	87.0	81.9	86.0	78.6	85.2	69	75
No Graph	78.6	82.0	83.9	87.0	70.0	73.0	77.0	77.6

For visual metaphor understanding, Table 5.10 shows LAG outperforms a Few-Shot baseline, though ablation studies 5.11 show an even higher performance in a 3-feature setting without the blending heuristics. Concerning human evaluation described in Petridis and Chilton (2019), which correctly interpreted the meaning of the image 41.32% of the time, the LLM correctly interpreted the meaning up to an average of 68.6% of the time. We then sampled the dataset, randomly taking 24 visual metaphors out of the 48 in the dataset, and repeated the original study (Petridis and Chilton, 2019) with 20 human participants. They have correctly interpreted the metaphorical property 59.25% of the time, with respect to the average LLM score of 73.5% for the same images.

Quality of the augmented graphs and error analysis. The results of the survey provided to two knowledge engineers on the quality of the augmented graphs highlight a consistent pattern where textual components generally outperform their visual counterparts across all evaluated metrics. Specifically, on a Likert scale from 1 to 7, correctness and adequacy in text scored at 5.1 and 6.3, suggesting a positive assessment, whereas the visual aspects scored slightly lower at 4.9 and 5 respectively, indicating room for improvement. The completeness of the text was rated at 5.2, while the visual completeness was lower at 4.4, showing a need for better integration or representation of visual elements to match the textual analysis.

For what concerns the error analysis, misalignment in conceptual metaphor mapping is predominantly due to *wrong subelement mapping*, which accounts for approximately 56.5% in the WG dataset and 57.1% in the BCMTD. Annotations that are too

specific contribute significantly as well, with 23.6% and 28.6% of errors in WG and BCTMD respectively, while overly general descriptions appear in 9.3% of WG and 14.3% of BCTMD cases. Less frequent errors concern switching the source and target (3.24% in WG) and identifying as literal something that was a metaphor (6.94% in WG). For visual metaphor understanding, the error analysis still follows the one done with human participants in Petridis and Chilton (2019), with errors mostly involving *incorrect objects* (21% of the errors) and *incorrect property* (57% of the errors). In 21% of cases, even when the LLM identifies source, target and connecting property, the role mapping explanation can show a misunderstanding of the metaphor (*incorrect target symbol*). For a more detailed analysis along with example errors, see the Appendix.

5.4.3.7 Discussion

In this section, we discuss the results, outline applications and further research directions for the proposed framework, and investigate limitations.

LAG for multimodal metaphor detection and interpretation. The results show that for the task of metaphor detection, the proposed LAG-based framework outperforms the current metaphor detection and understanding baselines.

However, the contrasting performance profiles across BCMTD and MMC reveal how LAG adapts to dataset characteristics. These effects are most evident on MMC, where domain-specific terminology, conventionalized scientific metaphors, and imbalanced prevalence jointly push LAG toward higher precision and lower recall. Three factors explain this shift: (1) *ambiguity sensitivity*—MMC contains boundary cases (Table 5.4) where humans disagree, triggering LAG’s conservative mode that requires concordant LLM-graph evidence; (2) *domain coverage asymmetry*—LAG’s 51.6% vs. 8% performance gap between generic and scientific metaphors on BCMTD directly predicts MMC’s low recall, as medical metaphors are predominantly domain-specific and graph evidence is sparse; (3) *dataset balance effects*—on balanced BCMTD, LAG achieves strong overall performance, while on imbalanced MMC, it prioritizes precision over recall. This adaptive behavior validates our complementarity hypothesis while revealing that neurosymbolic integration does not uniformly improve performance but rather shifts toward high-precision operation in high-stakes scientific contexts where false positives are costly.

This approach can be used to extract knowledge graphs from unstructured multimodal metaphorical data. As such, it could have effects in different applications, for example, improving hate speech, which often has metaphorical features, and analyzing debates in the press and online media.

For what concerns visual metaphors, since humans tend to struggle to identify them (Petridis and Chilton, 2019), the automated understanding of visual blends can bridge this gap. Additionally, using an open-source LLM like Llama reduces detection performance but not the understanding one.

Ablation studies show that while removing the graph reduces the performance in all cases, removing the blending heuristics in the prompt reduces it in all cases besides visual metaphor understanding and the detection of metaphors in the BCTMD. While this is a slight difference, it can be explained in two ways. First, visual metaphors and scientific metaphors might need a more specific ontological model. Furthermore, in visual metaphors, cross-modal interference can occur: textual descriptions produced by an LLM may foreground objects or relations that are only weakly expressed in the image, nudging the model toward a literal reading and masking the

metaphorical cue, or vice-versa. A second explanation is shortcutting: stock visuals—e.g. the classic handgun silhouette in a PSA or a light-bulb-shaped fruit—map so reliably to danger or creativity that the model picks the label without doing any cross-modal alignment. When we drop the description or the image, that one-channel cue remains and accuracy can even rise. Curbing this requires adversarial images that break the template and prompts that hide tell-tale keywords, forcing true visual–text fusion.

A cue for this slight performance drop emerges from the error analysis (see Section 5.4.3.6), where *incorrect property* was the most frequent issue both for humans (Petridis and Chilton, 2019) and LLMs: the objects are detected in the images, but there is still a struggle with extracting the correct properties. Because half of the images are advertisements, which typically present clearer context, the intended properties of the objects are usually clearly defined. By not specifying the purpose of the image and its context, we make it challenging for both humans and machines to accurately determine the metaphorical property via a blending mechanism that tends to be more open-ended without contextual specifications. Additionally, the low accuracy score observed in detecting scientific metaphors appears to reflect inherent limitations in current LLMs rather than a specific weakness of the LAG approach. Scientific metaphors often require domain-specific context and specialized knowledge that may not be well-represented in LLM training data. Also to be taken into account is the fact that scientific metaphors show multiple levels of conventionality, tending to result more literal as they come to be widespread in language.

Quality of the augmented graphs and error analysis The results from the questionnaire to analyze the quality of the augmented graphs suggest a positive performance with a disparity between text and visual outputs, probably because the chosen examples were the ones taken by Petridis and Chilton (2019) and they were not further annotated with an RDF syntax in order to be consistent with the human evaluation.

For what concerns error analysis, overly general descriptions show difficulty in identifying the correct taxonomical level of the metaphor. Additionally, less frequent errors, such as switching the source and target and identifying as literal something that was a metaphor, show further layers of complexity in the metaphor mapping process, such as misunderstanding the role of source and target domains and determining the metaphoricity level (a sentence might seem literal when the perceived conventionality of the metaphor is high).

In particular, misalignment in conceptual metaphor mapping is predominantly due to wrong subelement mapping, which accounts for approximately 56.5% in the WG dataset and 57.1% in the BCTMD. Annotations that are too specific contribute significantly as well, with 23.6% and 28.6% of errors in WG and BCTMD respectively, while overly general descriptions, appearing in 9.3% of WG and 14.3% of BCTMD cases, showing difficulty in identifying the correct taxonomical level of the metaphor. Additionally, less frequent errors, such as switching the source and target (3.24% in WG) and identifying as literal something that was a metaphor (6.94% in WG), show further layers of complexity in the metaphor mapping process, such as misunderstanding the role of source and target domains and determining the metaphoricity level (a sentence might seem literal when the perceived conventionality of the metaphor is high).

These results underscore the need to train and test automatic methods on domain-specific figurative datasets, but also that analogical reasoning is generally not properly performed in these tasks. On the other hand, visual metaphor understanding

shows different kinds of errors. Errors involving *incorrect objects* indicate poor image understanding skills, for example, a LLM will identify the shape of a gun as a whistle, or pink Nike shoes as fishes, usually skewing the interpretation. The most common error, however, is *incorrect property*: if the LLM correctly identifies the objects, source and target, but still misinterprets the property. For example, in the case of the blending of the car keys and the gun (displayed in Figure ??), with the annotated property being DANGEROUS, the predicted property is POWERFUL. This error is also observed in human participants and tied to their values and beliefs over the objects displayed in the blend, a further area of investigation. Furthermore, the presence of the *incorrect target symbol* error shows that the evaluation of the property alone as proposed in previous studies (Petridis and Chilton, 2019) is not sufficient to verify if the system has understood a metaphor. The error analysis shows the need to incorporate contextual meaning into the understanding of the metaphor, but also indicates that complete metaphor understanding requires a deeper level of semantic and cultural awareness that despite guidance towards relational analogical thinking, current systems struggle to achieve, also due to the absence of standardized gold-standard annotations that would establish clear metaphorical hierarchies. This multifaceted challenge highlights the gap between computational pattern recognition and the cognitive processes underlying human metaphor comprehension.

Interpretability of metaphorical datasets The proposed framework’s justifiability through XKG creation helps to outline some aspects that so far have not fully emerged in computational approaches to metaphor.

Context and culture.

Our research demonstrates that annotating the source and target elements in computational metaphor analysis is insufficient for fully grasping metaphorical meaning. For instance, as illustrated in Figure ??, the LLM accurately identifies all objects within the image that form the metaphor. However, although the blending property is annotated as DANGEROUS, the LLM interprets it as POWERFUL. In the absence of external context that can further frame the reference elements needed for comprehension, both interpretations are arguably valid.

This and other examples from our experiments reveal two key insights. First of all, the perspective adopted by a LLM in interpreting a metaphor reflects its underlying cultural background. Furthermore, for systems (and users) to reliably converge on metaphorical annotated properties, providing additional context is essential. This requirement for context may also help explain the relatively poor performance observed in human responses, such as distinguishing whether an advertisement is a public campaign, a comic strip, or a standard advertisement.

The dynamics of metaphor annotation. Beyond culture and context-dependence, the human annotation showed that some sentences’ gold standard metaphors in the WG, BCTMD and the Visual Metaphors dataset had multiple possible correct answers or even multiple answers in the same sentence. For example, for the sentence “This mini-controversy erupted when Republicans introduced a string of amendments in a final effort to obstruct passage of the reconciliation bill” the gold-standard metaphor is INHIBITING ENACTMENT OF LEGISLATION IS IMPEDING MOTION, while the predicted one is CONTROVERSY IS VOLCANO ERUPTION, which has been noted as correct by the evaluators but would be deemed as incorrect with automatic assessment. This phenomenon shows the need to keep both the human and the metaphor interpretation process flexible. However, current annotated metaphorical datasets

only contain one possible gold standard answer. Future work should address the possibility to incorporate multiple correct answers for a reference text.

5.4.4 Toward metaphor generation

Having established LAG's effectiveness for metaphor detection, understanding, and reasoning, a natural extension lies in metaphor generation, considered as the systematic creation of novel, plausible, contextually appropriate metaphors. The framework's ability to extract and formalize blending properties through the Blending Ontology provides the necessary scaffolding for generative tasks. While detection and analysis work with existing conceptual mappings, generation requires the inverse process: starting from desired communicative goals or domain constraints and systematically constructing plausible source-target pairings with appropriate blending properties. This generative capability would complete the operational cycle, enabling applications in creative writing, educational content generation, and therapeutic communication where novel metaphors must be both semantically coherent and contextually effective.

5.5 Implications for Computational Metaphor Treatment

This chapter has operationalized metaphor through three integrated components: domain-specific datasets revealing graded metaphoricity, a Blending Ontology formalizing conceptual integration, and a neurosymbolic LAG pipeline demonstrating practical applicability. Together, these contributions bridge theoretical understanding and computational implementation.

This operationalization directly answers our RQ, *How can metaphor be operationalized?*, through three integrated components responding to the requirements established in previous chapters: (1) datasets capturing graded metaphoricity and domain specificity, (2) a Blending Ontology encoding analogical structure via frames and non-conservative extensions, and (3) a LAG-based neurosymbolic pipeline executing detection, understanding, and constrained generation. These components realize the framework advocated in Chapter 3: domain-specific resources, formal representation methods, and consensus-aware evaluation reflecting human uncertainty.

The Medical Metaphor Corpus and Balanced Conceptual Metaphor Testing Dataset address a critical gap in metaphor research, that of lacking domain-specific datasets. Apart from providing medical metaphoric resources, the annotation of this dataset reveals metaphoricity as a spectrum rather than a binary property, with the confidence-weighted evaluation metrics showing that current LLMs perform systematically better on cases with strong human consensus while struggling with ambiguous instances. This finding has important implications for both theoretical understanding and practical applications, suggesting that metaphor processing systems must account for varying degrees of human agreement rather than assuming universal interpretive consensus.

The Blending Ontology provides the first formal computational framework for representing metaphors as dynamic conceptual integrations. By introducing non-conservative extensions and compositional principles, the ontology captures the genuinely creative aspect of metaphorical thinking: the emergence of novel conceptual structures that cannot be reduced to their constituent inputs. This formalization extends beyond metaphor to transformational creativity and multimodal reasoning,

offering a foundation for understanding how conceptual innovation occurs through structured blending processes.

Our exploration of blending property extraction, the property motivating the mapping between two domains, across frame-based, embedding-based, and LLM-based approaches reveals fundamental limitations in current computational methods. While each approach contributes valuable insights, none alone can reliably identify the conceptual bridges that enable metaphorical mappings. This failure motivates the neurosymbolic approach that follows.

The LAG framework successfully integrates these insights, demonstrating significant improvements over existing baselines across multiple evaluation dimensions. Following a four-layer evaluation framework—detection, understanding, reasoning, and generalization—we provide a comprehensive assessment methodology for metaphor processing. The LAG-based method’s superior performance on visual metaphors compared to human participants illustrates the potential for AI systems to complement human cognitive capabilities in complex interpretive tasks.

However, the results also expose important limitations. The substantial performance gap between generic and scientific metaphors highlights the domain-specificity challenge that must be addressed for practical applications. The error analysis reveals that current systems excel at surface associations but struggle with relational reasoning, suggesting that the challenge of metaphor understanding extends beyond pattern recognition to fundamental questions about analogical cognition.

The operational framework presented here makes several theoretical contributions to metaphor research. First, it shows that metaphor processing requires both symbolic structure and neural flexibility—a finding with implications for cognitive science and AI architecture. Second, it shows that metaphoricity perception varies systematically across domains and contexts, and provides resources that annotate this. Third, it reveals that successful metaphor understanding requires explicit modeling of perspectivization and cultural context, not merely structural mapping.

While challenges remain, particularly in domain transfer and deep relational reasoning, this framework demonstrates that neurosymbolic approaches can implement theoretical insights, establishing metaphor as a concrete testbed for advancing analogical reasoning in AI systems.

Chapter 6

Conclusions and future work

Throughout this thesis, we have approached metaphor as a meaning-making interworld that organizes how concepts meet, create, and transform one another. The central challenge has been to develop computational methods that can operationalize this space: to make explicit what metaphor reveals and what it obscures. Current approaches have proven insufficient. Symbolic systems offer interpretable structure but lack coverage and adaptability. Neural systems, including state-of-the-art LLMs, achieve broad coverage but struggle with reliability, constraint adherence, and interpretability, particularly for metaphors as they require relational reasoning.

This thesis addresses that gap by advancing what we call the complementarity hypothesis: no single paradigm can account for metaphor’s recognition, interpretation, and reasoning demands, but their combination can. We argue that knowledge patterns provide the interface between shared cognitive grounding and machine-interpretable structure, enabling systems to move beyond fixed definitions to process the flexible and dynamic nature of meaning. Metaphor, particularly in scientific discourse, serves as the decisive testbed for this claim because it demands relational mapping, blending, and context-sensitive reasoning while maintaining sufficient tractability for systematic analysis.

The thesis delivers four interconnected outcomes that instantiate this approach across data, representation, computation, and evaluation.

The first outcome establishes an empirical footing by making figurative meaning observable through domain-aware, consensus-weighted resources that capture graded metaphoricity and source-target structure (§5.1). The Medical Metaphor Corpus (MMC) supplies 792 sentences with crowd-validated perceived metaphoricity on a 0–7 scale, spanning scholarly articles, specialist handbooks, news, patient-oriented platforms, and crowdsourced data across 82 metaphor types, 24 target domains, and 38 source domains. The Balanced Conceptual Metaphor Testing Dataset (BCMTD) provides 147 sentences balanced across general conceptual metaphors, scientific conceptual metaphors, and literal controls, designed to minimize lexical artifacts and probe domain transfer. Both resources are linked to frame resources and explicitly track human agreement, addressing the shortage of domain-specific, fine-grained annotated datasets while ensuring that ambiguous and high-agreement cases are not treated alike in evaluation. These resources make human consensus explicit and enable confidence-weighted scoring that reveals when models succeed on clear surface associations and when they fail on relational reasoning.

The second outcome provides a formal structure for metaphorical conceptual integration. The Blending Ontology (§5.2.1) formalizes how blended structure and emergent properties arise by encoding a minimal set of operational primitives: input spaces, cross-space correspondences, a generic space, admissibility constraints, and provenance. It functions as a reusable pattern extended with other patterns, such as cognitive perspectivization, thus supporting multiple lenses within a single formal account. Critically, the ontology supports non-conservative emergence by explicitly distinguishing inherited structure from genuinely new relations and roles, addressing metaphor’s capacity to generate conceptual content that neither source nor target domain possessed independently. This representational layer is grounded in frame-based modeling and graph semantics, aligning with cognitive science while remaining computationally tractable.

The third outcome is an integrated neurosymbolic pipeline in which the Blending Ontology guides a Logic Augmented Generation (LAG) system to produce interpretable, constrained outputs across text and vision. LAG operates by using frames and the Blending Ontology as scaffolding and delegates adaptive filling to LLMs

within explicit constraints. Structured knowledge encodes roles, constraints, and admissibility, while LLMs provide flexible completion under guidance.

The studies in Chapter 5 formalize this integration and supply empirical support: frame-based signals alone miss novelty and coverage, cognitive embeddings alone reveal plausibility windows and dominance profiles without enforcing structure, and LLMs alone lack stability and provenance. Only the integration of structural knowledge through frame semantics in the form of KPs and LLMs as flexible reactors, which we see with the LAG-based pipeline, yields reliable, interpretable behavior. On scientific metaphor detection, LAG with graph and blending heuristics attains strong performance compared to baselines, with gains persisting under confidence-weighted metrics that respect human consensus and correct for ambiguous items that annotators dispute. Graph-only and blending-only ablations confirm that graph structure is necessary across textual tasks, with blending heuristics helping in most settings. In vision, LAG improves substantially over prior human baselines and few-shot prompts; image and graph features drive the gains, whereas textual descriptions can sometimes bias the model toward literal cues. Across text and vision, the main errors are role misbinding and granularity mismatches, missed scope and context markers, confusion with metonymy or literality, and lens misalignment across disciplines. Confidence-weighted evaluation of scientific metaphors shows many errors cluster in low-consensus items, but some high-consensus misses reveal reliance on topical collocations, suggesting shortcutting of LLMs. Graph constraints prevent most misbindings, and blending heuristics curb overly broad frame selection, yet gaps remain when background facts are absent. In vision, literal salience often overrides weak metaphorical cues, functional knowledge not visible in the image is required, and accompanying text can steer the model toward literal readings; adding graph features and visual relational cues mitigates these effects but does not remove them when domain knowledge is under-specified.

The fourth outcome addresses evaluation and reliability of LLM-based computational tasks in knowledge engineering. Evaluation cannot be unidimensional. The thesis develops and evaluates LLM-assisted methods that position LLMs as reactive generators within reproducible, human-in-the-loop workflows. Ontogenia demonstrates that LLM-assisted, pattern-guided ontology generation reaches high coverage of competency questions when steered by ontology design patterns, achieving quality comparable to novice ontology engineers under competency question verification and expert review, especially with reasoning LLMs. OE-Assist reveals that assistance improves accuracy when suggestions are correct but induces automation bias when they are wrong, quantifying gains and risks while motivating interface guardrails. CoLLM measures consistency in LLM-aided knowledge engineering, establishing empirical baselines on repeatability, sensitivity to model updates, and model replaceability. Results show that while most pipelines are repeatable, version updates can undermine the reliability of conclusions, which motivates strict versioning and explicit reporting of prompts, models, and parameters. LLMs propose mappings, constraints, and candidate triples, but only experts accept, revise, or reject them and certify evaluation outcomes. This boundary improves requirement coverage and reduces automation bias, as shown by evidence from Ontogenia and OE-Assist. User studies across knowledge engineering tasks reveal real benefits and real risks, underscoring the need for versioning, provenance, and human oversight in any pipeline that claims understanding. Thanks to these insights, for metaphor processing, we introduce a four-layer protocol that measures detection, understanding of source and target domains (source–target identification), reasoning via recovery of the blending property, and cross-domain and cross-modal generalization.

Item difficulty is modulated by human agreement to reflect practical reliability, and scores are weighted by annotator confidence so that reliability is explicit at every layer.

Together, these outcomes validate the complementarity hypothesis and demonstrate that neurosymbolic integration is necessary for robust implicit meaning processing. In this context, we argue LLMs can be considered as reactive generators anchored by structured knowledge, with human oversight and reproducibility practices as fundamental requirements.

Within the application of LAG for metaphor processing, much of what enables a metaphor to work is also not overtly stated. In the pipeline we propose, graphs state what must be explicit, LLMs conjecture what is missing while recording and testing those conjectures. The Blending Ontology makes that implicit substrate explicit through generic spaces, cross-space correspondences, admissibility constraints, and provenance. Empirically, the presented studies also recover two aspects that guide evaluation and design. First, a plausibility window: many plausible metaphor mappings live at intermediate cognitive similarity; pairs that are too close skew literal or metonymic, while pairs that are too far tend to be incoherent. Second, a directional dominance asymmetry: mappings that push highly salient source dimensions into suppressed target dimensions are markedly less plausible than the converse, which helps explain strain in certain readings and the stabilizing effect of conventional blends. These findings argue for modeling plausibility as a joint function of graded similarity and directional crossing profiles, and they justify LAG's constraint-first, completion-second architecture.

The outcomes of this thesis also connect to broader questions about conventionality and meaning. Despite how it is commonly treated in quantitative studies, conventionality in metaphor is graded rather than binary. The career of metaphor hypothesis (§2.2.3) holds that novel metaphors can become conventional and eventually terminological, a process particularly evident in science, where terminologization links directly to use and purpose within specific communities of practice. This explains why certain metaphors stabilize within a field while others remain fluid or fade away. Domain-sensitive computational approaches must therefore calibrate what counts as metaphor relative to user needs and contexts, a requirement that current systems struggle to meet. Yet conventionalization does not foreclose meaning. Semiotic theory reminds us that even so-called dead metaphors retain an open potential for reactivation in new settings, and their performative power persists in shaping how inquiry proceeds. This motivates sustained awareness of when a metaphor hardens into quasi-literal jargon and of how its entrenched use continues to guide attention and constrain reasoning, even or especially when it appears transparent. This awareness is sharpened by considering ease of processing: conventional metaphors are typically faster to understand and require less inferential work, which influences both how they spread through a community and how they direct cognitive effort. Conceptual Blending Theory helps bring these dynamics together by explaining how selective projection, emergent structure, and constraint satisfaction operate across the spectrum from novel to conventional metaphors, offering a framework for tracking both their stabilization and their ongoing capacity to generate new meanings.

In the view advanced here, we stress that knowledge patterns function as interfaces (Hookway, 2014), formal and structured mechanisms that enable connection across domains, while metaphors constitute interworlds, emergent spaces where meaning is produced through the encounter of source and target. KPs, aided by

LLMs, supply the shared relational scaffolding that makes these interworlds computationally tractable while preserving variability in use, and LAG provides the operational mechanism that turns scaffolding into testable graph updates and heuristics. This stance addresses fundamental limitations in current computational systems. Neurosymbolic methods, working with LLMs, address the previously unsolved problem of scalability. Traditional methodologies expose fundamental bottlenecks in requirements engineering and module implementation, relying primarily on manual systems. Human collaboration, which is fundamental in ontology engineering methodologies, is resource-intensive, and there is increasing difficulty for pattern discovery as repositories grow. A related problem concerns interoperability and eventual use. There is a gap between high-performing models and human interpretability, especially when users must understand how a system produced its outputs. Neurosymbolic approaches help bridge this gap by coupling statistical power with explicit structure, constraints, and provenance.

By addressing these challenges, the thesis answers its research questions. RQ1 asked what role knowledge patterns play in developing systems that can move beyond fixed definitions to process the flexible and dynamic nature of meaning as found in human thought and language, and metaphors in particular. The thesis argues that KPs function as a use-sensitive interface that preserves roles and relations while enabling flexible interpretation. They are supramodal, perspicuous layouts of roles and constraints that let systems move beyond fixed definitions by exploiting frame-based structure while leaving meaning to be fixed by use. This capability supports analogical mapping, blending, and multiple coexisting readings with provenance, acting as reusable templates that enable flexible yet testable interpretations of metaphor.

RQ2 asked what computational requirements are needed to study metaphor, either of the generic or scientific kind. The thesis demonstrates that effective computational study requires domain-aware datasets with graded metaphoricity and provenance, a blending-oriented representation that can host contextual and pragmatic cues while supporting executable constraints, and algorithms that combine symbolic constraints with adaptive generation, together with consensus-aware metrics across detection, interpretation, and reasoning, plus versioned graphs, prompts, and tests.

RQ3 asked how automation in knowledge engineering has evolved with LLMs and how LLMs can be systematically integrated into a reproducible, human-in-the-loop pipeline for knowledge engineering automation of metaphor processing. The thesis shows that LLMs can be systematically integrated as reactive generators within a reproducible, human-in-the-loop workflow—via Ontogenia for pattern-guided generation, OE-Assist for evaluation with guardrails, and CoLLM for consistency measurement—so long as oversight, version control, and explicit reporting of prompts, models, and parameters are enforced. Automation has evolved from manual methods to LLM-assisted workflows where models draft axioms, instantiate patterns, and evaluate outputs, with humans retaining oversight. The solution is a guarded integration that maintains human control, adopts multidimensional evaluation, and documents models, versions, and data flows for reproducibility.

RQ4 asked how metaphor can be operationalized. The thesis demonstrates that metaphor can be operationalized by using resources such as the MMC and BCMTD with graded metaphoricity and domain specificity, a Blending Ontology that formalizes conceptual integration, and a Logic Augmented Generation pipeline that detects metaphors, identifies source–target mappings, recovers blending properties

across text and images, and generalizes across four layers: detection, understanding, reasoning, and cross-domain and cross-modal generalization.

Taken together, these contributions address the contemporary critical literature’s desiderata for usage-based, intersubjective, and empirically constrained accounts of metaphor (Zlatev, Jacobsson, and Paju, 2021). They take seriously Barnden’s three aims for AI, where engineering progress is valuable insofar as it advances psychological and philosophical understanding (Barnden, 2008). In fact, on the engineering side, the thesis presents a reproducible neurosymbolic pipeline with knowledge pattern and blending constraints, together with pattern-guided knowledge engineering and evaluation protocols. On the psychological side, it introduces graded, consensus-weighted evaluations and user studies that reveal where models align with or diverge from human judgment, setting the basis for further experiments involving humans in the loop. On the philosophical side, it develops a use-sensitive account in which KPs function as interfaces and metaphors as interworlds, aligning formal, cognitive, and semiotic perspectives.

Implications and early uptake. The resources and methods presented here show concrete signs of adoption beyond this thesis. The Polanyi interface, which relies on Logic Augmented Generation, has integrated a dedicated coercion heuristic for metaphor that reuses the Blending Ontology¹. The ELMKE (Evaluation of Language Models in Knowledge Engineering) workshop, colocated with ESWC, has established itself as a recurring venue, with editions in 2025 and 2026². Pattern-guided ontology generation methods have been adopted in educational settings (including the University of Bologna’s Knowledge Engineering course) and are under discussion for adaptation to industry contexts. The GitHub repositories for the methodologies presented have garnered measurable community attention, traffic, and engagement. These early signals suggest that releasing versioned code, prompts, and tests lowers the barrier to independent reproduction and reuse, addressing reproducibility concerns that affect the broader field.

Limitations. The present studies are constrained by scale, scope, and setting. The metaphor datasets, while carefully constructed, remain relatively small and English-centric, with limited coverage of visual metaphors and pragmatic context, which narrows external validity and cultural generalizability. The Blending Ontology, though operational and reusable, encodes a minimal core and omits modules specific to scientific terminologization and semiotic features of visual metaphors, which restricts representational expressivity in some domains. LAG’s performance depends on hand-specified constraints and heuristics whose ablations show necessity but not sufficiency, and the pipeline inherits version sensitivity from LLM components, which affects repeatability across model updates. Finally, the knowledge engineering user studies, while informative, use bounded tasks and participant pools, limiting conclusions about real-world tooling and long-horizon collaboration. The proposed extensions directly target these limitations.

6.1 Future Work

The contributions introduced in this work are open to enrichment and systematic evaluation.

¹arco.istc.cnr.it/polanyi/.

²<https://sites.google.com/view/elmke>.

6.1.0.1 Resources

Regarding the expansion of resources, the MMC presented in Chapter 5.1.1 can be extended to include additional medical metaphors as well as metaphors from other domain-specific contexts. The dataset can also be enriched with visual scientific metaphors (for example, graphical abstracts, schematic cartoons, clinical diagrams), enabling annotation of plastic features following semiotic principles, together with pragmatic contextual information. On the pragmatic side, resources could include, for instance, the perlocutionary effect of a metaphor, which, as shown, is reported in medical metaphors but has not been evaluated systematically nor computationally. Other contextual elements could capture visual cues and visual tone of voice. We discuss this possible addition in Cappa, Lippolis, and Zoia (2025). Apart from that, the proposed datasets should be extended beyond English to include aligned subsets in other languages, preserving provenance and consensus scores, and supporting studies on cultural framing.

These annotations would not be useful without an ontology. The Blending Ontology presented in Chapter 5.2.1 can incorporate these aspects. As observed, scientific metaphors processed automatically may require separate treatment. The ontology could therefore be extended with a pattern on how scientific metaphorical terms are introduced and used (Chapter 2), which would also mitigate shortcutting effects.

6.1.0.2 Methods

Logic Augmented Generation has proven successful in processing metaphors in computation; however, its application to metaphors should be evaluated on additional existing datasets, and future studies should involve more qualitative annotators.

For LLM-assisted knowledge engineering, Ontogenia and the ontology-generation workflow can be further modified for the task ontology extension with retrieval-augmented generation. This line of work can also introduce new evaluation measures. Moreover, OE-Assist has shown that LLMs can struggle when used to evaluate ontologies. A natural next step is to improve the accuracy and to integrate LLM suggestions into a dedicated interface, such as a Protégé plugin. Just as today's coding environments provide suggestions, LLM-aided knowledge engineering should prioritize interpretable suggestions rather than solutions. Closely related is the question of how to preserve human autonomy while building these tools, which would require a separate discussion and analysis.

The work on consistency with LLMs is related to this agenda. In addition to the three existing tests, we envision the integration of further evaluations, in particular: (i) noise sensitivity, to investigate how noise in the input prompt affects the consistency and quality of LLM outputs; (ii) hyperparameter sensitivity, to examine how LLM hyperparameters affect different KE tasks, and (iii) biases and consistency of LLMs to find recurring error patterns across prompts, domains, representational formats, and demographic axes, and determines whether these biases cause inconsistent outputs under identical inputs.

Metacognitive stages	Ontogenia stages	Description
Comprehension clarification	1. Competency question understanding.	The LLM interprets the CQs, mirroring the interpretation of the text in MP. LLMs are prompted to contextualize the CQs.
Preliminary judgement	2. Preliminary identification of the context. 3. Division of the competency question in subject, predicate, object and predicate nominative.	A logical analysis of the CQs is requested to support the identification of classes and properties.
Critical evaluation	5. Starting from your knowledge, extension of the ontology with restrictions (existential, cardinality, disjointness).	Reflect on the CQs and incorporate knowledge to add rules and restrictions to the model.
Decision confirmation	8. Confirmation of the final answer and explain the reasoning.	Explain the reasoning again.
Confidence assessment	9. Confidence evaluation and explanation using the ontology reasoning and test it on some instances.	Request a reasoned evaluation of the performed actions, akin to the MP model, and ask for an evaluation of the entire process. Request test cases that provide evidence of the correctness of the created model.

TABLE A.1: Mapping between the Ontogenia methodology and the MP procedure.

Appendix A

Appendix A: ontology generation studies details

A.1 Ontogenia methodology details

Ontogenia stages.

Define an ontology using an overall methodology based on understanding each competency question using this procedure:

1. Competency question understanding.
2. Preliminary identification of the context.
3. Divide the competency question in subject, predicate, object, and predicate nominative. Assign each subject and object to a class, each predicate to an object property where the domain is the subject and the range is the object. If a predicate already exists, you can expand it by adding the new relevant classes, if they are not present, in the domain and the range. Each predicate nominative becomes a datatype property, referred to the subject or object.
4. If you can answer 'yes' to the question 'is A an element of B?', then A is a `rdfs:subClassOf B`. If not sure, try to reassess.
5. Starting from your knowledge, extend the ontology with these restrictions: if in your experience a class has a property that is true for all instances of another class, then you can add an `owl:allValuesFrom` restriction; if a class has a property that has a value, then you can add an `owl:hasValue` restriction; if a class has a property that is true for some instances of another class, then you can add an `owl:someValuesFrom` restriction. If a class has a property that can be defined, you can have an `owl:Restriction` `minCardinality` with zero; if a class has a property that must be defined, you can have an `owl:Restriction` `minCardinality` with one.
6. Starting from your knowledge, if all individuals of class A are individuals in the class B, and vice versa, then you can add an `owl:equivalentClass`; if all individuals of the class A are not in the class B, then A is disjoint to B so you can add an `owl:disjointClass`.
7. Now add all restrictions and cardinality and equivalent and disjoint classes for this context in the owl ontology.
8. Confirm the final answer and explain the reasoning.
9. Make a confidence evaluation and explanation using the ontology reasoning and test it on some instances.

A.2 Domain ontology generation prompt

```

<instruction>
You are a helpful assistant designed to generate ontologies. You receive
a Competency Question (CQ) and an Ontology Story (OS). \n
Based on CQ, which is a requirement for the ontology, and OS, which tells
you what the context of the ontology is, your task is generating one
ontology (O). The goal is to generate O that models the CQ properly.
This means there is a way to write a SPARQL query to extract the
answer to this CQ in O. \n
Use the following prefixes: \n
@prefix : <http://www.example.org/test#> .\n
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .\n
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .\n
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .\n

Don't put any A-Box (instances) in the ontology and just generate the OWL
file using Turtle syntax. Include the entities mentioned in the CQ.
Remember to use restrictions when the CQ implies it. The output should
be self-contained without any errors. Outside of the code box don't
put any comment.
</instruction>
<example>
I will give you an example now:
CQ: For an aquatic resource, what species, in what fishing areas,
climatic zones, bottom types, depth zones, geo forms, horizontal and
vertical distribution, and for what standardized abundance level,
exploitation state and rate, have been observed? for what reference
year? At what year the observation has been recorded?

OS: The aquatic resource for Skipjack tuna in Northern Atlantic in 2004
(as reported in 2008) was observed with low abundance level.

O:
@prefix : <http://www.example.org/test#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
<http://www.example.org/test#> a owl:Ontology .
:AquaticResource a owl:Class ;
    rdfs:label "Aquatic Resource" ;
    rdfs:subClassOf [
        a owl:Restriction ;
        owl:onProperty :hasObservation ;
        owl:someValuesFrom :AquaticResourceObservation
    ] .
:AquaticResourceObservation a owl:Class ;
    rdfs:label "Aquatic Resource Observation" ;
    rdfs:subClassOf :Observation ;
    rdfs:subClassOf [
        a owl:Restriction ;
        owl:onProperty :isObservationOf ;
        owl:onClass :AquaticResource ;
        owl:qualifiedCardinality "1"^^xsd:nonNegativeInteger
    ] ;

    rdfs:subClassOf [
        a owl:Restriction ;
        owl:onProperty :forReferenceYear ;
        owl:onDataRange xsd:int ;
        owl:qualifiedCardinality "1"^^xsd:nonNegativeInteger

```

```

] ;

rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :hasReportingYear ;
  owl:onDataRange xsd:int ;
  owl:qualifiedCardinality "1"^^xsd:nonNegativeInteger
] ;

rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofSpecies ;
  owl:onClass :Species ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;
rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofFishingArea ;
  owl:onClass :FAO_fishing_area ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;
rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofClimaticZone ;
  owl:onClass :ClimaticZone ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;
rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofBottomType ;
  owl:onClass :BottomType ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;
rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofDepthZone ;
  owl:onClass :DepthZone ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;
rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofGeoForm ;
  owl:onClass :GeoForm ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;
rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofHorizontalDistribution ;
  owl:onClass :HorizontalDistribution ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;
rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofVerticalDistribution ;
  owl:onClass :VerticalDistribution ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;
rdfs:subClassOf [
  a owl:Restriction ;
  owl:onProperty :ofStdAbundanceLevel ;
  owl:onClass :StdAbundanceLevel ;
  owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
] ;

```

```

    rdfs:subClassOf [
      a owl:Restriction ;
      owl:onProperty :ofStdExploitationState ;
      owl:onClass :StdExploitationState ;
      owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
    ] ;
    rdfs:subClassOf [
      a owl:Restriction ;
      owl:onProperty :ofStdExploitationRate ;
      owl:onClass :StdExploitationRate ;
      owl:minQualifiedCardinality "0"^^xsd:nonNegativeInteger
    ] .

:Species a owl:Class .
:FAO_fishing_area a owl:Class .
:ClimaticZone a owl:Class .
:BottomType a owl:Class .
:DepthZone a owl:Class .
:GeoForm a owl:Class .
:HorizontalDistribution a owl:Class .
:VerticalDistribution a owl:Class .
:StdAbundanceLevel a owl:Class .
:StdExploitationState a owl:Class .
:StdExploitationRate a owl:Class .
:ofSpecies a owl:ObjectProperty ;
  rdfs:label "ofSpecies" ;
  rdfs:domain :AquaticResourceObservation ;
  rdfs:range :Species .

:ofFishingArea a owl:ObjectProperty ;
  rdfs:label "ofFishingArea" ;
  rdfs:domain :AquaticResourceObservation ;
  rdfs:range :FAO_fishing_area .

:ofClimaticZone a owl:ObjectProperty ;
  rdfs:label "ofClimaticZone" ;
  rdfs:domain :AquaticResourceObservation ;
  rdfs:range :ClimaticZone .

:ofBottomType a owl:ObjectProperty ;
  rdfs:label "ofBottomType" ;
  rdfs:domain :AquaticResourceObservation ;
  rdfs:range :BottomType .

:ofDepthZone a owl:ObjectProperty ;
  rdfs:label "ofDepthZone" ;
  rdfs:domain :AquaticResourceObservation ;
  rdfs:range :DepthZone .

:ofGeoForm a owl:ObjectProperty ;
  rdfs:label "ofGeoForm" ;
  rdfs:domain :AquaticResourceObservation ;
  rdfs:range :GeoForm .

:ofHorizontalDistribution a owl:ObjectProperty ;
  rdfs:label "ofHorizontalDistribution" ;
  rdfs:domain :AquaticResourceObservation ;
  rdfs:range :HorizontalDistribution .

:ofVerticalDistribution a owl:ObjectProperty ;
  rdfs:label "ofVerticalDistribution" ;
  rdfs:domain :AquaticResourceObservation ;
  rdfs:range :VerticalDistribution .

```



```
:ofStdAbundanceLevel a owl:ObjectProperty ;
    rdfs:label "ofStdAbundanceLevel" ;
    rdfs:domain :AquaticResourceObservation ;
    rdfs:range :StdAbundanceLevel .

:ofStdExploitationState a owl:ObjectProperty ;
    rdfs:label "ofStdExploitationState" ;
    rdfs:domain :AquaticResourceObservation ;
    rdfs:range :StdExploitationState .

:ofStdExploitationRate a owl:ObjectProperty ;
    rdfs:label "ofStdExploitationRate" ;
    rdfs:domain :AquaticResourceObservation ;
    rdfs:range :StdExploitationRate .
:forReferenceYear a owl:DatatypeProperty ;
    rdfs:label "forReferenceYear" ;
    rdfs:domain :AquaticResourceObservation ;
    rdfs:range xsd:int .

:hasReportingYear a owl:DatatypeProperty ;
    rdfs:label "hasReportingYear" ;
    rdfs:domain :AquaticResourceObservation ;
    rdfs:range xsd:int .
:hasObservation a owl:ObjectProperty .
:isObservationOf a owl:ObjectProperty .
:Observation a owl:Class .
```

</example>

CQ: {CQ}
OS: {OS}
O:

Appendix B

Appendix B: CoLLM study details

B.1 Summary and details of the CoLLM studies

B.2 Results of studies 1, 3, 4, 5

B.3 Overall discussion on the studies

TABLE B.1: Summary of the selected studies used in the evaluation of CoLLM, listing their datasets, key assumptions or conclusions, benchmarks or baselines, the evaluation criteria employed, and the threshold pass/fail status.

Selected Studies	Datasets	Assumption or Conclusion	Benchmark or Baseline	Evaluation Criteria	Threshold Pass/Fail
Study 1: Ontology Generation with LLMs Saeedizade and Blomqvist, 2024	SemanticWeb Course (45 CQs & 3 ontology stories)	LLMs can generate ontology similar to novice ontology engineers	Students' submissions of the SemanticWeb Course	Proportion of correctly modeled CQs in the ontologies	Test scores less than 80% are fail
Study 2: Ontogenia Lippolis et al., 2024 (§4.3.1)	African Wildlife (1 ontology, 26 triples)	Ontogenia fosters a richer formalization and higher quality ontologies.	African Wildlife Ontology	OntomMetrics evaluation, qualitative evaluation by expert ontology engineer	Pass only if Ontogenia generates the highest OntoMetrics count
Study 3: LLMs4OL Paradigm for Ontology Learning Babaei Giglou, D'Souza, and Auer, 2023	UMLS – SNOMEDCT_US (Task A, B, and C)	Zero-shot seems promising for ontology learning, but task-specific fine-tuning is needed for practical viability.	UMLS – SNOMEDCT_US (Ground Truth)	Task A (Term Typing): MAP@1, Task B (Taxonomy Discovery) and C (Non-Taxonomic Relation Extraction), F1-Score	Pass if the score is comparable to or higher than those of the original study (here, fail if PD is less than -5%)
Study 4: LLMs4OM for Ontology Matching Giglou et al., 2024	MSE (MI-EMMO task), Biodiv (ALGAE-ZOOBENTHOS), Anatomy (Mouse-Human), Bio-ML (OMIM-ORDO), Phenotype (HP-MP)	RAG is optimal for ontology matching, with the right LLM and retriever models achieving performance that can surpass SOTA matching models	MI-EMMO (63 alignments), ALGAE-ZOOBENTHOS (18), Mouse-Human (1,516), OMIM-ORDO (3,721), and Phenotype (696)	For the alignments ground truth, F1-Score is used for evaluation.	Pass if the score is comparable to or higher than those of the original study (here, fail if PD is less than -4%)
Study 5: Retrofit-CQs Alharbi et al., 2024a	Video Game Ontology (57 triples), African Wildlife (26 triples), Dem@care (146 triples), Vicinity Core (226 triples)	CQs can be retrofitted from existing ontologies by instructing LLMs with in-context prompts accompanied by ontology triples.	Video Game Ontology (66 CQs), African Wildlife (14 CQs), Dem@care (107 CQs), Vicinity Core (57 CQs)	SBERT Similarity Reimers and Gurevych, 2019 measures the semantic similarity between retrofitted CQs and the ground-truth. Metrics are precision and recall.	Pass if precision & recall are comparable to or higher than those of the original study, with a low STD.

TABLE B.2: Summary of the experimental setup for all research studies used in the evaluation of CoLLM, listing their LLMs, temperature settings, prompts, and other settings.

Research Study	LLMs/Test	Temperature/LLM	Prompt	Other Settings
Ontology Generation with LLMs (Saeedizade and Blomqvist, 2024)	GPT-4-0613 for $\mathcal{T}_{\text{rep}}$, GPT-4-1106, GPT-4-Turbo 2024-04-09, and GPT-4-2024-05-13 for $\mathcal{T}_{\text{sen}}$, o1-preview for $\mathcal{T}_{\text{gen}}$	Temperature set to 0 for GPT-4 and default for o1	Details in supplementary materials	Penalty for GPT set to zero except for o1, which is set to default
Ontogenia (Lippolis et al., 2024)	GPT-4-1106 for $\mathcal{T}_{\text{rep}}$, GPT-4 1025, Turbo 2024-04-09 for $\mathcal{T}_{\text{sen}}$, o1-preview and Llama-3.1-405-Instruct_Turbo for $\mathcal{T}_{\text{gen}}$	Temperature set to 0 for GPT-4 and default for o1	Details in supplementary materials	Penalty for GPT set to zero except for o1, which is set to default
LLMs4OL (Babaei Giglou, D'Souza, and Auer, 2023)	Task A: BLOOM (3B) for $\mathcal{T}_{\text{rep}}$ and BLOOM (1B1, 1B7, 7B1) for $\mathcal{T}_{\text{sen}}$, and LLaMA-3.1-8B for $\mathcal{T}_{\text{gen}}$. B: GPT-4-0613 for $\mathcal{T}_{\text{rep}}$ and GPT-4 (1106, 0125, Turbo 2024-04-09) for $\mathcal{T}_{\text{sen}}$, and o1-preview for $\mathcal{T}_{\text{gen}}$. C: Flan-T5 (XL) for $\mathcal{T}_{\text{rep}}$ and Flan-T5 (base, L, XXL) for $\mathcal{T}_{\text{sen}}$, and Mistral-7B-V0.3 for $\mathcal{T}_{\text{gen}}$.	Temperature is set to 0 for GPT-4 and default for o1-preview.	Details in supplementary materials	
LLMs4OM (Giglou et al., 2024)	LLaMA-2-7B, Mistral-7B-V0.1, GPT-3.5-Turbo-1106 for $\mathcal{T}_{\text{rep}}$, LLaMA-3-8B, LLaMA-3.1-8B, Mistral-7B-V0.2, Mistral-7B-V0.3, and GPT-3.5-Turbo-0125 for $\mathcal{T}_{\text{sen}}$, Mistral-7B-V0.3 and LLaMA-3.1-8B for $\mathcal{T}_{\text{gen}}$. Ada text embedding and BERT for retriever.	Temperature is set to 0 for GPT-3.5.	Details in supplementary materials	For the retriever model, the top-k is set to 5.
Retrofit-CQs (Alharbi et al., 2024a)	GPT-4-0613 for $\mathcal{T}_{\text{rep}}$, GPT-3.5-0125 and GPT-3.5-1106 for $\mathcal{T}_{\text{sen}}$, o1-preview for $\mathcal{T}_{\text{gen}}$	Temperature set to 0.7 for GPT-4-0613, GPT-3.5-0125, and GPT-3.5-1106. For o1-preview, the temperature is fixed at 1 by OpenAI	[As an ontology engineer, generate a list of competency questions based on the <statement>. Definition of competency questions: the questions that outline the scope of the ontology and provide an idea about the knowledge that needs to be entailed in the ontology. Avoid using narrative questions + statement]	

TABLE B.3: Results of Study 1: CoLLM tests using evaluation metrics (reported as $\mu \pm \sigma$ of three runs) accompanied by a $PD_{\mathcal{T}}$ for ontology generation using task decomposition. The σ reported in the primary work Saeedizade and Blomqvist, 2024 was 10. The results are in %.

	LLM	Percentage of Modeled CQs ($PD_{\mathcal{T}}$)
Reported Score (r_{Org})	GPT-4 (0613)	89 $\pm$ 10
LLM Repeatability Test ($\mathcal{T}_{rep}$)	GPT-4 (0613)	91.11 $\pm$ 2.22 (+2.37%)
LLM Update Impact Test ($\mathcal{T}_{upi}$)	GPT-4 (1106)	91.85 $\pm$ 1.28 (+3.2%)
	GPT-4-Turbo (2024-04-09)	90.81 $\pm$ 2.67 (+2.02%)
	GPT-4-Turbo (2024-05-13)	95.53 $\pm$ 2.21 (+7.33%)
LLM Replacement Test ($\mathcal{T}_{rpl}$)	<i>o1-preview</i>	89.78 $\pm$ 3.07 (+0.87%)

TABLE B.4: Results of Study 3: CoLLM tests using evaluation metrics (reported as $\mu \pm \sigma$ of three runs) and $PD_{\mathcal{T}}$ for different tasks of the LLMs40L paradigm using UMLS datasets. The model $\star$ refers to GPT-4-turbo-2024-04-09 model. The results are given as percentages and results representing failed tests appear in italics.

Task A		Task B		Task C	
LLM	MAP@1 ($PD_{\mathcal{T}}$)	LLM	F1-Score ($PD_{\mathcal{T}}$)	LLM	F1-Score ($PD_{\mathcal{T}}$)
r_{Org}					
BLOOM (3B)	37.70	GPT-4 (0613)	78.10	Flan-T5 (XL)	49.50
LLM Repeatability Test ($\mathcal{T}_{rep}$)					
BLOOM (3B)	37.74 $\pm$ 0.0 (0%)	GPT-4 (0613)	78.02 $\pm$ 0.25 (0%)	Flan-T5 (XL)	41.39 $\pm$ 0.0 (-16.38%)
LLM Update Impact Test ($\mathcal{T}_{upi}$)					
BLOOM (1B1)	36.00 $\pm$ 0.0 (-4.50%)	GPT-4 (1106)	80.92 $\pm$ 0.23 (+3.61%)	Flan-T5 (Base)	49.00 $\pm$ 0.0 (-1.01%)
BLOOM (1B7)	37.00 $\pm$ 0.0 (-1.85%)	GPT-4 (0125)	79.66 $\pm$ 0.60 (+1.99%)	Flan-T5 (L)	41.17 $\pm$ 0.0 (-16.82%)
BLOOM (7B1)	35.62 $\pm$ 0.0 (-5.51%)	GPT-4 ($\star$)	80.87 $\pm$ 0.19 (+3.54%)	Flan-T5 (XXL)	39.46 $\pm$ 0.0 (-20.28%)
LLM Replacement Test ($\mathcal{T}_{rpl}$)					
LLaMA-3.1-8B	47.91 $\pm$ 0.05 (+27.05%)	<i>o1-preview</i>	87.27 $\pm$ 0.75 (+11.74%)	Mistral-7B-V0.3	51.86 $\pm$ 0.0 (+4.76%)

TABLE B.5: Results of Study 4: CoLLM tests using evaluation metrics (reported as $\mu \pm \sigma$ of three runs) and $PD_{\mathcal{T}}$ for different datasets/tracks of LLMs40M framework. C CC, and CP refer to concept, concept-children, and concept-parent representations, respectively. The model $\star$ refers to Mistral-7B-V0.2 and Mistral-7B-V0.3, and $\star\star$ refers to LLaMA-3-8B and LLaMA-3.1-8B, with both pairs yielding identical results. The results are given as percentages and results representing failed tests appear in italics.

Track	Retriever Model + LLM	r_{Org}	F1-Score ($PD_{\mathcal{T}}$)
LLM Repeatability Test ($\mathcal{T}_{rep}$)			
MSE	BERT (C) + LLaMA-2-7B (CC)	94.30	95.08 $\pm$ 0.0 (+0.61%)
Biodiv	ADA (C) + Mistral-7B-V0.1 (C)	56.00	56.00 $\pm$ 0.0 (0.0)
Anatomy	ADA (C) + GPT3.5-Turbo-1106 (C)	89.11	89.08 $\pm$ 0.02 (-0.03%)
Bio-ML	ADA (C) + GPT3.5-Turbo-1106 (CC)	64.14	63.55 $\pm$ 0.92 (-0.91%)
Phenotype	BERT (C) + Mistral-7B-V0.1 (CP)	85.01	86.12 $\pm$ 0.0 (+1.30%)
LLM Update Impact Test ($\mathcal{T}_{upi}$)			
MSE	BERT (C) + LLaMA-3-8B (CC) $\star\star$	94.30	94.30 $\pm$ 0.0 (0.0)
Biodiv	ADA (C) + Mistral-7B (C) $\star$	56.00	56.00 $\pm$ 0.0 (0.0)
Anatomy	ADA (C) + GPT3.5-Turbo-0125 (C)	89.11	88.18 $\pm$ 0.01 (-0.03%)
Bio-ML	ADA (C) + GPT3.5-Turbo-0125 (CC)	64.14	64.15 $\pm$ 0.13 (-0.01%)
Phenotype	BERT (C) + Mistral-7B (CP) $\star$	85.01	80.59 $\pm$ 0.0 (-4.94%)
LLM Replacement Test ($\mathcal{T}_{rpl}$)			
MSE	BERT (C) + Mistral-7B-V0.3 (CC)	94.30	95.16 $\pm$ 0.0 (+0.90%)
Biodiv	BERT (C) + Mistral-7B-V0.3(C)	56.00	80.00 $\pm$ 0.0 (+42.85%)
Anatomy	BERT (C) + Mistral-7B-V0.3(C)	89.11	88.18 $\pm$ 0.0 (-1.03%)
Bio-ML	BERT (C) + Mistral-7B-V0.3(CC)	64.14	63.06 $\pm$ 0.0 (-1.66%)
Phenotype	BERT (C) + LLaMA-3.1-8B (CP)	85.01	89.09 $\pm$ 0.0 (+4.78%)

TABLE B.6: Results of Study 5: CoLLM tests using evaluation metrics (reported as $\mu \pm \sigma$ of three runs) and $PD_{\mathcal{T}}$ for different ontologies for retrofitting competency questions using RETROFIT-CQ framework. The model V1 refers to 0613, V2 refers to 1106, V3 refers to 0125. The results are given as percentages.

Ontology	$r_{Org}(Pre, Rec) - LLM$	LLM	Precision ($PD_{\mathcal{T}}$)	Recall ($PD_{\mathcal{T}}$)
LLM Repeatability Test ($\mathcal{T}_{rep}$)				
Video Game	(82.1, 99.8)-GPT-4 V_1	GPT-4 V_1	70.87±0.4 (-13.68%)	99.82±0.1 (+0.02%)
African Wild	(49.9, 99.1)-GPT-4 V_1		78.14±3.3 (+56.59%)	99.70±0.5 (+0.61%)
Dem@care	(30.7, 99.4)-GPT-4 V_1		45.15±1.3 (+47.07%)	99.96±0.0 (+0.56%)
Vicinity Core	(62.3, 99.9)-GPT-4 V_1		68.40±0.3 (+9.79%)	99.98±0.1 (+0.08%)
LLM Update Impact Test ($\mathcal{T}_{upi}$)				
Video Game	(64.1, 99.1)-GPT-3.5 V_1	GPT-3.5 V_2	74.61 ±3.1 (+16.40%)	96.61±0.2 (-2.51%)
		GPT-3.5 V_3	70.72 ±1.9 (+10.33%)	97.85±0.0 (-1.27%)
African Wild	(75.6, 99.5)-GPT-3.5 V_1	GPT-3.5 V_2	89.63±1.1 (+18.56%)	100±0.0 (+0.50%)
		GPT-3.5 V_3	92.2 ±2.0 (+21.96%)	100±0.0 (+0.50%)
Dem@care	(42.2, 99.5)-GPT-3.5 V_1	GPT-3.5 V_2	57.01 ±1.9 (+35.09%)	99.57 ±0.4 (+0.07%)
		GPT-3.5 V_3	53.1 ±1.5 (+25.84%)	99.83 ±0.2 (+0.33%)
Vicinity Core	(42.2, 99.9)-GPT-3.5 V_1	GPT-3.5 V_2	75.97 ±1.4 (+78.75%)	99.92 ±0.1 (+0.02%)
		GPT-3.5 V_3	60.37 ±1.7 (+42.07%)	100 ±0.0 (+0.10%)
LLM Replacement Test ($\mathcal{T}_{rpl}$)				
Video Game	(82.1, 99.8)-GPT-4 V_1	o1-preview	63.85±0.5 (-22.23%)	99.73 ±0.0 (-0.07%)
African Wild	(49.9, 99.1)-GPT-4 V_1		47.72±4.5 (-4.37%)	100±0.0 (+0.91%)
Dem@care	(30.7, 99.4)-GPT-4 V_1		36.81±0.3 (+19.90%)	99.96 ±0.1 (+0.56%)
Vicinity Core	(62.3, 99.9)-GPT-4 V_1		51.71±1.4 (-17.00%)	100±0.0 (+0.10%)

TABLE B.7: Discussion over observations, evidence, impacts, and recommendations across studies.

Study	Observation	Evidence / Likely Causes	Impact / Consequences	Recommended Actions
Study 1: Ontology Generation	All three tests passed; main findings of the original study unchanged.	Higher empirical scores than the original study; task not affected by any variances.	Stronger support for the claim that LLMs generate ontologies comparable to student work.	Replicate across more datasets.
Study 2: Ontogenia	Some results show inconsistencies and variability.	Not repeating runs in the original study (ignores LLM stochasticity); high std. dev. in metrics; reliance on mixed qualitative/quantitative evaluations that fail to capture ontology richness. Variations in results are also influenced by the extent of memorised ontological information within the models, which can bias the performance towards well-documented concepts.	Reported benefits of MP+Patterns may not be robust; qualitative depth not fully assessed.	Repeat tests with variance reporting; adopt richer quantitative metrics; analyse memorisation bias; balance sample size with depth.
Study 3: LLMs4OL	Model updates (e.g., BLOOM) change conclusions (negative PD in Update Impact Test).	Parameter shifts across versions; some inconsistencies repeated in Tasks A and C; limited set of models tested so far.	Reduced reproducibility; conclusions may no longer hold after model updates.	Run update-impact tests routinely; expand experiments to larger models (o1-preview, Mistral, LLaMA-3).
Study 4: LLMs4OM	Two consistent failures (0% std. dev.) on the Phenotype dataset with BERT(C)+Mistral-7B (CP); inconsistent across others.	Architecture or training data in updated models; sensitivity to hyperparameters and datasets; heterogeneous behaviours across tasks.	Ontology-matching results are fragile; updates can either help or harm, depending on conditions.	Perform sensitivity analyses; validate across datasets; explore model/data factors linked to failures.
Study 5: Retrofitting Competency Questions	All three tests passed; scores exceeded original; low variability across tests.	Precision/recall comparable or better than original; low std. dev.; semantic similarity measures confirm quality vs. expert-authored CQs.	Confirms LLMs can retrofit competency questions for existing ontologies; stable and reliable results.	Extend to more ontology domains.

Appendix C

Appendix C: Details of Chapter 5

C.1 Prompt of experiment of Section 5.3.4

Examine the following data: ‘sentence’. Based on the metaphorical mapping provided, evaluate the given source and target frames: *sourceFrame* and *targetFrame*. Use frame semantic principles to identify and connect potential roles and elements from the source to the target domain that are made available. Your analysis should:

1. Enumerate any explicit roles or elements present within the source frame.
2. Map the identified source role to its corresponding target role, if one is readily apparent. You don’t need to map all roles if not relevant. Use the format *SOURCE_ROLE* → *TARGET_ROLE*.
3. Highlight any conceptual themes or actions that link these roles. Mark core ideas with asterisks to signify their importance to the metaphor’s underlying structure.
4. Where roles are missing or implicit, infer their presence and articulate the potential roles in the mapping, guided by semantic intuition and the principles of metaphorical transfer.

For instance, in the metaphor ACQUIRING IDEAS IS EATING, critical concepts such as **transformation** and **internalization** emerge. This metaphor extends beyond mere conversion, encompassing subtleties like **saturation** and **fulfillment**, indicative of the extent to which ideas are absorbed; **quality of consumption**, reflecting the discernment of ideas; **preferences** and **taste**, denoting the selective nature of cognitive engagement; and the act of **chewing** as *analogous to contemplation*, where ideas are deliberated upon. Apply this level of analysis to elucidate the metaphor at hand, spotlighting the interplay of its roles and concepts.

C.2 LAG Method and implementation details

C.2.1 Text2AMR2FRED Usage

The usage of Text2AMR2FRED has been carried out through its API¹. The input is a text or an image and the output is a RDF graph in Turtle format.

C.2.2 Dataset examples and statistics

Table C.3 summarizes the statistics for the three datasets according to number of instances, percentage of metaphorical sentences, and chosen number of samples. Table 5.6 shows examples for each dataset involved.

¹Documentation available at <http://framester.istc.cnr.it/txt-amr-fred/api/docs>.

C.2.3 Prompt designs and information extraction

All prompt designs, along with information extraction from the LLM's output are available on GitHub, both in the code and in a dedicated document².

C.3 Evaluation details

In this section, details about the evaluation are presented.

C.3.1 Inter Annotator Agreement

Table C.3 summarizes the inter-annotator agreement for each of the main tasks. For what concerns graph quality evaluation, done by two experts, the weighted Kappa is of 0.3, suggesting moderate agreement between the two annotators.

C.3.2 Quality of the enhanced graphs

Table C.1 shows the questions asked for the evaluation of the quality of the enhanced graphs and the final complete results. Sampling 16 correctly individuated metaphors for each dataset, we analyze the graph blends according to the quality of the enhanced graphs on the basis of i) Adequacy; ii) Correctness; iii) Completeness. Adequacy measures whether a graph includes all necessary and relevant elements of a sentence or image, ensuring that the representation is meaningful and complete enough for further analysis. Correctness assesses the accuracy of a graph's elements and their interrelationships, ensuring that they align precisely with the specified properties and relationships of the ontology, such as in accurately representing metaphorical mappings. Completeness evaluates whether a graph captures every essential component and relationship needed for a full understanding of the metaphor, covering both explicit and implicit aspects to provide a comprehensive view. Identifying not only the correct source domains but also precise element-wise mappings between the concepts of the target and source domain is useful for metaphor understanding.

C.3.3 Error analysis

We employ error categories outlined in Wachowiak and Gromann, 2023 and Petridis and Chilton, 2019 for metaphor understanding. More details on this categorization along with examples are available on GitHub. The categories are:

1. *Wrong with trigger*: incorrect prediction of the domain due to words in the sentence related to that domain;
2. *Wrong without trigger*: incorrect prediction of the domain without possible triggers;
3. *Too literal*: prediction is considered metaphoric but is actually literal;
4. *Should be non-metaphoric*: metaphoric prediction instead of literal;
5. *Should be metaphoric*: literal prediction instead of metaphoric;
6. *Too specific*: one or both domains are too specific;
7. *Too general*: one or both domains are too general;
8. *Wrong subelement mapping*: one or both domains predict an aspect of the correct domain, though not precisely;

²https://github.com/dersuchendee/metaphorical-logic-augmented-generation/blob/main/code/prompts_info.md

9. (for visual metaphors) *incorrect objects*: incorrect prediction of one or both the objects in the image;
10. (for visual metaphors) *incorrect direction of property transfer*: incorrect prediction of source and target objects;
11. (for visual metaphors) *incorrect property*: source and target are correctly identified, but not the property;
12. (for visual metaphors) *incorrect target symbol*: all objects are identified but misinterpretation remains.

TABLE C.1: Evaluation Form for Graph Representation.

Item	Question	Scale (1-7)
Adequacy	On a Likert scale from 1 (not at all adequate) to 7 (perfectly adequate), how would you rate the graph's overall adequacy in representing the sentence or image?	1-7
Correctness	On a Likert scale from 1 (not at all correct) to 7 (perfectly correct), how would you rate the graph's overall correctness in representing metaphorical mappings between the two domains?	1-7
Completeness	On a Likert scale from 1 (not at all complete) to 7 (perfectly complete), how would you rate the graph's overall completeness in representing metaphorical mappings and elements in the sentence or image?	1-7

TABLE C.2: Examples of sentence types with corresponding metaphors used in experiments.

Type	Example	Metaphor
Conceptual metaphor	"The money is all dried up."	MONEY AS LIQUID
Scientific metaphor	"Tissues of cells independent of normal growth-controlling forces."	CELLS AS SOCIETY MEMBERS
Literal sentence	"Then I normally watch it with you."	None
Visual metaphor		Metaphor: EARTH IS ICE CREAM Property: Melt (of: Ice cream)

TABLE C.3: Table about inter-annotator agreement.

Task	Accuracy	Fleiss Kappa	Agreement in %
BCTMD LAG	34.3%	0.35 (Fair)	30%
WG LAG	25.6%	0.41 (Fair)	70.3%
Visual metaphor dataset LAG no img	65.06%	0.29 (Fair)	54%
Visual metaphor dataset LAG no sent	67.06%	0.306 (Fair)	65.28%
Visual metaphor dataset LAG	65.2%	0.274 (Fair)	63.68%
Visual metaphor dataset Few-Shot	54.7%	0.33 (Fair)	60.42%

Bibliography

- Abebe, Leah M et al. (2017). "Answers to common misconceptions about biological evolution". In.
- Achiam, Josh et al. (2024). *GPT-4 Technical Report*. arXiv: [2303.08774 \[cs.CL\]](https://arxiv.org/abs/2303.08774). URL: <https://arxiv.org/abs/2303.08774>.
- Agerri, Rodrigo (2008). "Metaphor in Textual Entailment". In: *Coling 2008: Companion Volume: Posters*, pp. 3–6.
- Agre, Philip E. (1992). "Formalization as a Social Project". In: *Quarterly Newsletter of the Laboratory of Comparative Human Cognition* 14.1, pp. 25–27.
- (1997). *Computation and Human Experience*. Cambridge: Cambridge University Press.
- Akula, Arjun R et al. (2023). "Metaclue: Towards comprehensive visual metaphors research". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23201–23211.
- Al-Aswadi, Fatima N, Huah Yong Chan, and Keng Hoon Gan (2020). "Automatic ontology construction from text: a review from shallow to deep learning trend". In: *Artificial Intelligence Review* 53.6, pp. 3901–3928.
- Alexander, Christopher (1977). *A pattern language: towns, buildings, construction*. Oxford university press.
- Algergawy, Alsayed et al. (2024). "Towards semantic annotation for scientific datasets". In: *European Semantic Web Conference*. Springer, pp. 164–167.
- Alharbi, Reham et al. (2024a). "An experiment in retrofitting competency questions for existing ontologies". In: *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, pp. 1650–1658.
- Alharbi, Reham et al. (2024b). "Characteristics and Desiderata for Competency Question Benchmarks". In: *The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings*.
- Alharbi, Reham et al. (2024c). "Exploring the Role of Generative AI in Constructing Knowledge Graphs for Drug Indications with Medical Context". In: *15th International Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS 2024) (to appear)*.
- Alharbi, Reham et al. (2024d). "Investigating open source LLMs to retrofit competency questions in ontology engineering". In: *Proceedings of the AAAI Symposium Series*. Vol. 4. 1, pp. 188–198.
- Alharbi, Reham et al. (2025). "A Review and Comparison of Competency Question Engineering Approaches". In: *Knowledge Engineering and Knowledge Management*. Cham: Springer Nature Switzerland, pp. 271–290. ISBN: 978-3-031-77792-9.
- Ali, Rohaid et al. (2023). "Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank". In: *Neurosurgery* 93.5, pp. 1090–1098. DOI: [10.1227/neu.0000000000002551](https://doi.org/10.1227/neu.0000000000002551).
- Allen, Bradley and Paul Groth (2024). "A Benchmark for the Detection of Metalinguistic Disagreements between LLMs and Knowledge Graphs". In: *The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings*.

- Alnajjar, Khalid, Mika Härmäläinen, and Shuo Zhang (2022). "Ring that bell: A corpus and method for multimodal metaphor detection in videos". In: *arXiv preprint arXiv:2301.01134*.
- Anand, Sanjay Kumar and Suresh Kumar (2022). "Uncertainty analysis in ontology-based knowledge representation". In: *New Generation Computing* 40.1, pp. 339–376.
- Ananny, Mike and Kate Crawford (2018). "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability". In: *New Media & Society* 20.3, pp. 973–989.
- Angeli, Gabor, Melvin Jose Johnson Premkumar, and Christopher D Manning (2015). "Leveraging linguistic structure for open domain information extraction". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 344–354.
- Antia, Mary-Jane and C Maria Keet (2023). "Automating the generation of competency questions for ontologies with AgOCQs". In: *Iberoamerican Knowledge Graphs and Semantic Web Conference*. Springer, pp. 213–227.
- Aranha, Claus et al. (2022). "Metaphor-based metaheuristics, a call for action: the elephant in the room". In: *Swarm Intelligence* 16.1, pp. 1–6.
- Aristotle (1920). *Poetics*. Trans. by Ingram Bywater. Oxford: Oxford University Press. URL: <https://archive.org/download/aristotleonarto00aris/aristotleonarto00aris.pdf>.
- Asprino, Luigi et al. (2017). "Frame-based ontology alignment". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1.
- Auer, Sören and Heinrich Herre (2006). "RapidOWL—An agile knowledge engineering methodology". In: *International Andrei Ershov Memorial Conference on Perspectives of System Informatics*. Springer, pp. 424–430.
- Auer, Sören et al. (2007). "Dbpedia: A nucleus for a web of open data". In: *international semantic web conference*. Springer, pp. 722–735.
- Austin, J. L. (1975). *How to Do Things with Words*. Ed. by J. O. Urmson and Marina Sbisa. 2nd ed. Cambridge, MA: Harvard University Press.
- Austin, John Langshaw (1962). *How to do things with words*. Cambridge, MA: Harvard University Press.
- Babaei Giglou, Hamed, Jennifer D'Souza, and Sören Auer (2023). "LLMs4OL: Large Language Models for Ontology Learning". In: *The Semantic Web – ISWC 2023*. Ed. by Terry R. Payne et al. Cham: Springer Nature Switzerland, pp. 408–427. ISBN: 978-3-031-47240-4.
- Babaei Giglou, Hamed, Jennifer D'Souza, and Sören Auer (2023). "LLMs4OL: Large language models for ontology learning". In: *International Semantic Web Conference*. Springer, pp. 408–427.
- Babaei Giglou, Hamed, Jennifer D'Souza, and Sören Auer (Oct. 2024). "LLMs4OL 2024 Overview: The 1st Large Language Models for Ontology Learning Challenge". In: *Open Conference Proceedings* 4, pp. 3–16. DOI: [10.52825/ocp.v4i.2473](https://doi.org/10.52825/ocp.v4i.2473). URL: <https://www.tib-op.org/ojs/index.php/ocp/article/view/2473>.
- Babaei Giglou, Hamed et al. (Oct. 2024). "LLMs4OL 2024 Datasets: Toward Ontology Learning with Large Language Models". In: *Open Conference Proceedings* 4, pp. 17–30. DOI: [10.52825/ocp.v4i.2480](https://doi.org/10.52825/ocp.v4i.2480). URL: <https://www.tib-op.org/ojs/index.php/ocp/article/view/2480>.
- Baker, Collin F, Charles J Fillmore, and John B Lowe (1998). "The berkeley framenet project". In: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

- Banaruee, Hassan et al. (2024). "Metaphors across cultures". In: *Poznan Studies in Contemporary Linguistics* 60.2, pp. 133–158.
- Barnden, John (2018). "Broadly reflexive relationships, a special type of hyperbole, and implications for metaphor and metonymy". In: *Metaphor and Symbol* 33.3, pp. 218–234.
- Barnden, John A (2008). "Metaphor and artificial intelligence: Why they matter to each other". In: *The Cambridge handbook of metaphor and thought*, pp. 311–338.
- Barnden, John A et al. (2004). "Varieties and directions of interdomain influence in metaphor". In: *Metaphor and Symbol* 19.1, pp. 1–30.
- Barsalou, Lawrence W (1999). "Perceptual symbol systems". In: *Behavioral and brain sciences* 22.4, pp. 577–660.
- Barsalou, Lawrence W. (2003). "Abstraction in perceptual symbol systems". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 358.1435, pp. 1177–1187.
- (2008). "Grounded cognition". In: *Annual Review of Psychology* 59, pp. 617–645.
- Barsalou, Lawrence W (2012). "Frames, concepts, and conceptual fields". In: *Frames, fields, and contrasts*. Routledge, pp. 21–74.
- Barthes, Roland (1964). "Rhétorique de l'image". In: *communications* 4.1, pp. 40–51.
- Bartlett, Frederic C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.
- Baudrillard, Jean (1990). *The transparency of evil: Essays on extreme phenomena*. Verso books.
- (1994). *Simulacra and simulation*. University of Michigan press.
- Becker, Lee A (2000). "Effect size (ES)". In.
- Bender, Emily M et al. (2021). "On the dangers of stochastic parrots: Can language models be too big?" In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.
- Benson, Carter et al. (2024). "My Ontologist: Evaluating BFO-Based AI for Definition Support". In: *arXiv preprint arXiv:2407.17657*.
- Benureau, Fabien C. Y. and Nicolas P. Rougier (2017). "Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions". In: *Frontiers in Neuroinformatics* 11, p. 69. ISSN: 1662-5196. DOI: [10.3389/fninf.2017.00069](https://doi.org/10.3389/fninf.2017.00069). URL: <https://www.frontiersin.org/articles/10.3389/fninf.2017.00069/full>.
- Benzi, Margherita and Marco Novarese (2022). "Metaphors we Lie by: our 'War' against COVID-19". In: *History and Philosophy of the Life Sciences* 44.2, p. 18.
- Berardinis, Jacopo de et al. (2023). "The polifonia ontology network: Building a semantic backbone for musical heritage". In: *International Semantic Web Conference*. Springer, pp. 302–322.
- Bergen, Benjamin K. (2004). "To Awaken a Sleeping Giant: Cognition and Culture in September 11 Political Cartoons". In: *Language, Culture, and Mind*. Ed. by Michel Achard and Suzanne Kemmer. Stanford, CA: CSLI Publications, pp. 23–35.
- Berners-Lee, Tim, Mark Fischetti, and Michael L. Dertouzos (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. 1st. Harper San Francisco. ISBN: 0062515861.
- Berners-Lee, Tim, James Hendler, and Ora Lassila (2001). *Scientific American: Feature Article: The Semantic Web*.
- Bhatt, Mehul, Joana Hois, and Oliver Kutz (2012). "Ontological modelling of form and function for architectural design". In: *Applied ontology* 7.3, pp. 233–267.
- Bi, Nanyi and Janet Yi-Ching Huang (2023). "I create, therefore I agree: Exploring the effect of AI anthropomorphism on human decision-making". In: *Companion*

- Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, pp. 241–244.
- Bianchini, Francesco (2016). “Analogy as categorization: A support for model-based reasoning”. In: *Model-Based Reasoning in Science and Technology: Logical, Epistemological, and Cognitive Issues*. Springer, pp. 239–256.
- Binder, Jeffrey R et al. (2016). “Toward a brain-based componential semantic representation”. In: *Cognitive neuropsychology* 33.3-4, pp. 130–174.
- Birke, Julia and Anoop Sarkar (2006a). “A clustering approach for nearly unsupervised recognition of nonliteral language”. In: *11th Conference of the European chapter of the association for computational linguistics*, pp. 329–336.
- (2006b). “A Simple Heuristic-Based Method for Detecting Metaphorical Usages in English Verb-Particle Constructions”. In: *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pp. 1–8.
- Bisang, Walter et al. (2006). *Corpus-based approaches to metaphor and metonymy*. Mouton de Gruyter.
- Black, Max (1962). *Models and Metaphors: Studies in Language and Philosophy*. Ithaca, NY: Cornell University Press.
- (1977). “More about metaphor”. In: *Dialectica* 31.3-4, pp. 431–457.
- Blakemore, Diane (2002). *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge Studies in Linguistics 99. Cambridge: Cambridge University Press. ISBN: 0521640075.
- Blanchette, Isabelle and Kevin Dunbar (2001). “Analogy use in naturalistic settings: The influence of audience, emotion, and goals”. In: *Memory & Cognition* 29.5, pp. 730–735.
- Bleakley, Alan (2017). *Thinking with metaphors in medicine: The state of the art*. Routledge.
- Blomqvist, Eva (2009a). “Ontocase-automatic ontology enrichment based on ontology design patterns”. In: *International Semantic Web Conference*. Springer, pp. 65–80.
- (2009b). *Semi-automatic ontology construction based on patterns*. Linköpings Universitet (Sweden).
- Blomqvist, Eva, Karl Hammar, and Valentina Presutti (2016). “Engineering ontologies with patterns—the eXtreme design methodology”. In: *Ontology Engineering with Ontology Design Patterns*. IOS Press, pp. 23–50.
- Blomqvist, Eva and Kurt Sandkuhl (2005). “Patterns in ontology engineering: Classification of ontology patterns”. In: *International Conference on Enterprise Information Systems*. Vol. 4. SCITEPRESS, pp. 413–416.
- Blomqvist, Eva, Azam Seil Sepour, and Valentina Presutti (2012). “Ontology testing-methodology and tool”. In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer, pp. 216–226.
- Blomqvist, Eva et al. (2010). “Experimenting with eXtreme design”. In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer, pp. 120–134.
- Bodd, Monica H. et al. (2023). ““More than conquerors”: A qualitative analysis of war metaphors for patients with cancer”. In: *Supportive Care in Cancer* 31.1, p. 87. DOI: [10.1007/s00520-022-07552-y](https://doi.org/10.1007/s00520-022-07552-y).
- Boden, Margaret A (2004). *The creative mind: Myths and mechanisms*. Routledge.
- Bohr, Niels (1985a). “Letter to Albert Einstein, 13 April 1927”. In: *Niels Bohr Collected Works, Volume 6: Foundations of Quantum Physics I (1926–1932)*. Ed. by Jørgen Kalckar. Amsterdam: North-Holland, pp. 418–421.

- (1985b). “Letter to Erwin Schrödinger, 2 December 1926”. In: *Niels Bohr Collected Works, Volume 6: Foundations of Quantum Physics I (1926–1932)*. Ed. by Jørgen Kalckar. Amsterdam: North-Holland, pp. 462–463.
- Bohrn, Isabel C, Ulrike Altmann, and Arthur M Jacobs (2012). “Looking at the brains behind figurative language—A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing”. In: *Neuropsychologia* 50.11, pp. 2669–2683.
- Bojanowski, Piotr et al. (2016). “Enriching Word Vectors with Subword Information”. In: *arXiv preprint arXiv:1607.04606*.
- (2017). “Enriching Word Vectors with Subword Information”. In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.
- Boldt, Joachim (2019). “Machine metaphors and ethics in synthetic biology”. In: *Life Sciences, Society and Policy* 15.1, pp. 1–20. DOI: [10.1186/s40504-019-0091-6](https://doi.org/10.1186/s40504-019-0091-6).
- Bollegala, Danushka et al. (2013). “Metaphor Interpretation Using Paraphrases Extracted from the Web”. In: *PLOS ONE* 8.9, e74304. DOI: [10.1371/journal.pone.0074304](https://doi.org/10.1371/journal.pone.0074304). URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0074304>.
- Bongard, Josh (2021). “Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior”. In: *Frontiers in Ecology and Evolution* 9, p. 650726. DOI: [10.3389/fevo.2021.650726](https://doi.org/10.3389/fevo.2021.650726).
- Boon, Errol (Sept. 2025). “One Last Miracle - Metaphysical Variations on Creativity and Artificial Intelligence”. In.
- Borghi, Anna M, Claudia Mazzuca, and Luca Tummolini (2025). “The role of social interaction in the formation and use of abstract concepts”. In: *Nature Reviews Psychology*, pp. 1–14.
- Boroditsky, Lera (2000). “Metaphoric structuring: Understanding time through spatial metaphors”. In: *Cognition* 75.1, pp. 1–28.
- Borsley, Stefan, David A Leigh, and Benjamin MW Roberts (2022). “Chemical fuels for molecular machinery”. In: *Nature Chemistry* 14.7, pp. 728–738.
- Borst, Willem Nico (1997). “Construction of engineering ontologies for knowledge sharing and reuse”. In.
- Bounegru, Liliana and Charles Forceville (2011). “Metaphors in editorial cartoons representing the global financial crisis”. In: *Visual Communication* 10.2, pp. 209–229. DOI: [10.1177/1470357211398446](https://doi.org/10.1177/1470357211398446).
- Bowdle, Brian F and Dedre Gentner (2005). “The career of metaphor.” In: *Psychological review* 112.1, p. 193.
- Boyd, Richard (1979). *Metaphor and Theory Change in Metaphor and Thought*.
- Brentano, Franz (1995). *Psychology from an Empirical Standpoint*. Trans. by Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister. London: Routledge.
- Brette, Romain (2019). “Is coding a relevant metaphor for the brain?” In: *Behavioral and Brain Sciences* 42, e215.
- Brito, Carlos and Victor Marques (2014). “The Rise and Fall of the Machine Metaphor: Organizational Similarities and Differences Between Machines and Living Beings”. In: *Verifiche* 43.1-3, pp. 77–111.
- Brooke, John et al. (1996). “SUS-A quick and dirty usability scale”. In: *Usability evaluation in industry* 189.194, pp. 4–7.
- Brown, Penelope and Stephen C Levinson (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.
- Brown, Theresa (Aug. 2014). *The Trouble with Medicine’s Metaphors*. The Atlantic. URL: <https://www.theatlantic.com/health/archive/2014/08/the-trouble-with-medicines-metaphors/374982/> (visited on 01/02/2025).

- Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*.
- Bundy, Jonathan et al. (2017). "Crises and crisis management: Integration, interpretation, and research development". In: *Journal of management* 43.6, pp. 1661–1692.
- Burrell, Jenna (2016). "How the machine 'thinks': Understanding opacity in machine learning algorithms". In: *Big Data & Society* 3.1, p. 2053951715622512.
- Buttigieg, Pier Luigi et al. (2013). "The environment ontology: contextualising biological and biomedical entities". In: *Journal of biomedical semantics* 4, pp. 1–9.
- Cabassa, Luca (2022). "Esistono buone metafore in scienza? Note in margine a" La linea e il circolo"". In: *Philosophy Kitchen-Rivista di filosofia contemporanea* 17, pp. 61–75.
- Cabr  Castellv , Maria Teresa (1999). *Terminology: Theory, Methods and Applications*. John Benjamins.
- Camus, Julia T Williams (2009). "Variation of cancer metaphors in scientific texts and press popularisations". In: *Corpus linguistics colloquium; Corpus-based approaches to figurative language*. Vol. 1, pp. 175–182.
- Cappa, Silvia, Anna Sofia Lippolis, and Stefano Zoia (2025). "Meanings are like Onions: A layered approach to metaphor processing". In: *EIKE Workshop at FOIS*.
- Carenini, Gaia et al. (2024). "Towards a Fully Interpretable and More Scalable RSA Model for Metaphor Understanding". In: *arXiv preprint arXiv:2404.02983*.
- Carnap, Rudolf (1950). "Logical foundations of probability." In.
- Carriero, Valentina Anita et al. (2019). "ArCo: The Italian cultural heritage knowledge graph". In: *International semantic web conference*. Springer, pp. 36–52.
- Carriero, Valentina Anita et al. (2020a). "Pattern-based design applied to cultural heritage knowledge graphs". In: *arXiv preprint arXiv:1911.07585*. DOI: [10.48550/arXiv.1911.07585](https://doi.org/10.48550/arXiv.1911.07585). arXiv: [1911.07585](https://arxiv.org/abs/1911.07585) [cs.AI]. URL: <https://doi.org/10.48550/arXiv.1911.07585>.
- Carriero, Valentina Anita et al. (2020b). "The Landscape of Ontology Reuse Approaches". In: *Applications and Practices in Ontology Design, Extraction, and Reasoning*. Ed. by Giuseppe Cota, Marilena Daquino, and Gian Luca Pozzato. Vol. 49. Studies on the Semantic Web. IOS Press, pp. 21–38. DOI: [10.3233/SSW200033](https://doi.org/10.3233/SSW200033).
- Carrillo, Nancy (2023). "Scientific Inquiry: From Metaphors to Abstraction". In: *Philosophy of Science* 90.2, pp. 233–257. URL: <https://direct.mit.edu/posc/article/31/2/233/112754/Scientific-Inquiry-From-Metaphors-to-Abstraction>.
- Carston, Robyn (2002). *Thoughts and utterances: The pragmatics of explicit communication*. Oxford: Blackwell.
- Carus, A. W. and Michael Friedman (2021). "Explication". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta.
- Carver, Rebecca, Ragnar Waldahl, and Jarle Breivik (2008). "Frame that gene: A tool for analysing and classifying the communication of genetics to the public". In: *EMBO reports* 9.10, pp. 943–947.
- Casasanto, Daniel (2009). "Embodiment of abstract concepts: Good and bad in right- and left-handers". In: *Journal of Experimental Psychology: General* 138.3, pp. 351–367.
- (2011). "Different bodies, different minds: The body specificity of language and thought". In: *Current Directions in Psychological Science* 20.6, pp. 378–383.
- Casasanto, Daniel and Lera Boroditsky (2008). "Time in the mind: Using space to think about time". In: *Cognition* 106.2, pp. 579–593.
- Casasanto, Daniel and Evangelia G. Chrysikou (2011). "When left is "right": Motor fluency shapes abstract concepts". In: *Psychological Science* 22.4, pp. 419–422.

- Casasanto, Daniel, Olga Fotakopoulou, and Lera Boroditsky (2010). "Space and time in the child's mind: Evidence for a cross-dimensional asymmetry". In: *Cognitive Science* 34.3, pp. 387–405.
- Casasanto, Daniel and Tania Henetz (2012). "Handedness shapes children's abstract concepts". In: *Cognitive Science* 36.2, pp. 359–372.
- Cauter, Zeno and Nikolay Yakovets (2024). "Ontology-guided knowledge graph construction from maintenance short texts". In: *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pp. 75–84.
- Chakrabarty, Tuhin et al. (2022). "FLUTE: Figurative Language Understanding through Textual Explanations". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 7139–7159.
- Chakrabarty, Tuhin et al. (2023). "I spy a metaphor: Large language models and diffusion models co-create visual metaphors". In: *arXiv preprint arXiv:2305.14724*.
- Chandler, Calvin et al. (2024). "Medical metaphors that may reinforce misconceptions are associated with increased trust in the clinician". In: *Quality Management in Healthcare* 33.4, pp. 246–252.
- Cheded, Mohammed, Chihling Liu, and Gillian Hopkinson (2022). "Dead metaphors and responsibilised bodies-in-transition: The implications of medical metaphors for understanding the consumption of preventative healthcare". In: *Transhumanisms and biotechnologies in consumer society*. Routledge, pp. 146–170.
- Chemero, Anthony (2003). "An Outline of a Theory of Affordances". In: *Ecological Psychology* 15.2, pp. 181–195. DOI: [10.1207/S15326969EC01502_5](https://doi.org/10.1207/S15326969EC01502_5).
- Chen, Lingjiao, Matei Zaharia, and James Zou (2024). "How is ChatGPT's behavior changing over time?" In: *Harvard Data Science Review* 6.2.
- Chen, Shengyuan et al. (2024). "Entity alignment with noisy annotations from large language models". In: *Advances in Neural Information Processing Systems* 37, pp. 15097–15120.
- Chen, Zhe et al. (2020). "Knowledge graph completion: A review". In: *Ieee Access* 8, pp. 192435–192456.
- Chu, Zheng et al. (2023). "A survey of chain of thought reasoning: Advances, frontiers and future". In: *arXiv preprint arXiv:2309.15402*.
- Chung, Hyung Won et al. (2022). *Scaling Instruction-Finetuned Language Models*. arXiv: [2210.11416](https://arxiv.org/abs/2210.11416) [cs.LG]. URL: <https://arxiv.org/abs/2210.11416>.
- Cimiano, Philipp (2006). *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer New York, NY. ISBN: 978-0-387-30632-2. DOI: [10.1007/978-0-387-39252-3](https://doi.org/10.1007/978-0-387-39252-3).
- Clark, Peter (2008). "Knowledge Patterns: (Extended Abstract)". In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer, pp. 1–3.
- Clusmann, Jan et al. (2023). "The future landscape of large language models in medicine". In: *Communications medicine* 3.1, p. 141.
- Coeckelbergh, Mark (2009). "Personal robots, appearance, and human good: a methodological reflection on roboethics". In: *International Journal of Social Robotics* 1, pp. 217–221.
- Cohen, Ted (1975). "Figurative Speech and Figurative Acts". In: *The Journal of Philosophy* 72.19, pp. 669–684. DOI: [10.2307/2024631](https://doi.org/10.2307/2024631).
- Cohn, Neil (2013). *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images*. London: Bloomsbury Academic.
- Collberg, Christian and Todd A. Proebsting (Feb. 2016). "Repeatability in computer systems research". In: *Commun. ACM* 59.3, 62–69. ISSN: 0001-0782. DOI: [10.1145/2812803](https://doi.org/10.1145/2812803). URL: <https://doi.org/10.1145/2812803>.

- Collins, Ben C. et al. (2017). "Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry". In: *Nature Communications* 8.1, p. 291. ISSN: 2041-1723. DOI: [10.1038/s41467-017-00249-5](https://doi.org/10.1038/s41467-017-00249-5). URL: <https://doi.org/10.1038/s41467-017-00249-5>.
- Collins, Harry (2010). *Tacit and Explicit Knowledge*. Chicago: University of Chicago Press.
- Confalonieri, Roberto and Oliver Kutz (2020). "Blending under Deconstruction: The roles of logic, ontology, and cognition in computational concept invention". In: *Annals of Mathematics and Artificial Intelligence* 88.5, pp. 479–516.
- Confalonieri, Roberto et al. (2016). "Conceptual blending in EL++". In: *CEUR Workshop Proceedings*.
- Confalonieri, Roberto et al. (2018). "Upward refinement operators for conceptual blending in the description logic ++". In: *Annals of Mathematics and Artificial Intelligence* 82.1, pp. 69–99.
- Coulson, Seana and Cyma Van Petten (2007). "A special role for the right hemisphere in metaphor comprehension?: ERP evidence from hemifield presentation". In: *Brain research* 1146, pp. 128–145.
- Cox, Michael (2016). "Metaphors in Medicine: Toward Transformation in Care". In: *Dignitas* 23.3, pp. 9–13.
- Crawford, L Elizabeth (2009). "Conceptual metaphors of affect". In: *Emotion review* 1.2, pp. 129–139.
- Cuccio, Valentina, Gerard Steen, et al. (2019). "Deliberate metaphors and embodied simulation". In: *Current approaches to metaphor analysis in discourse*, pp. 185–204.
- Cui, Lei, Furu Wei, and Ming Zhou (2018). "Neural open information extraction". In: *arXiv preprint arXiv:1805.04270*.
- Cui, Ziyang, Ning Li, and Huaikang Zhou (2025). "A large-scale replication of scenario-based experiments in psychology and management using large language models". In: *Nature Computational Science*, pp. 1–8.
- Dafermos, Manolis (2024). "Discussing the concept of crisis in cultural-historical activity research: A dialectical perspective". In: *Human Arenas* 7.2, pp. 273–292.
- Dancygier, Barbara and Lieven Vandelanotte (2017). "Internet memes as multimodal constructions". In: *Cognitive Linguistics* 28.3, pp. 565–598. DOI: [10.1515/cog-2017-0074](https://doi.org/10.1515/cog-2017-0074).
- Danesi, Marcel (2013). "A Note on the Notion of Metaform and Its Implications for Semiotics". In: *Signata: Annals of Semiotics* 4, pp. 397–407. DOI: [10.4000/signata.1028](https://doi.org/10.4000/signata.1028). URL: <https://journals.openedition.org/signata/1028>.
- d'Aquin, Mathieu and Enrico Motta (2011). "Watson, more than a Semantic Web search engine". In: *Semantic Web* 2.1, pp. 55–63.
- Davidson, Donald (1978). "What Metaphors Mean". In: *Critical Inquiry* 5.1, pp. 31–47.
- Dawkins, Richard (1981). "Selfish genes and selfish memes". In: *The mind's I: Fantasies and reflections on self and soul*, pp. 124–144.
- De Giorgis, Stefano, Aldo Gangemi, and Alessandro Russo (2025). "Neurosymbolic graph enrichment for Grounded World Models". In: *Information Processing Management* 62.4, p. 104127. ISSN: 0306-4573. DOI: <https://doi.org/10.1016/j.ipm.2025.104127>. URL: <https://www.sciencedirect.com/science/article/pii/S030645732500069X>.
- De Leenheer, Pieter and Christophe Debruyne (2008). "DOGMA-MESS: A tool for fact-oriented collaborative ontology evolution". In: *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, pp. 797–806.

- De Mauro, Tullio (1984). "Semantica". In: *Enciclopedia del Novecento*. Treccani. URL: https://www.treccani.it/enciclopedia/semantica_%28Enciclopedia-del-Novecento%29/.
- (1988). "Linguaggi scientifici e lingue storiche". In: *AR Guerriero (a cura di) L'educazione linguistica ei linguaggi delle scienze*. Firenze: La Nuova Italia.
- (1990). "Minisemantica dei linguaggi non verbali e delle lingue". In.
- (2018). *Linguaggi scientifici e lingue storiche*. URL: <https://giscel.it/wp-content/uploads/2018/04/Tullio-De-Mauro-Linguaggi-scientifici-e-lingue-storiche.pdf>.
- (2021). *Guida all'uso delle parole: parlare e scrivere semplice e preciso per capire e farsi capire*. Gius. Laterza & Figli Spa.
- Deacon, Terrence W (1998). *The symbolic species: The co-evolution of language and the brain*. WW Norton & Company.
- Debruyne, Christophe, Trung-Kien Tran, and Robert Meersman (2013). "Grounding ontologies with social processes and natural language". In: *Journal on Data Semantics* 2.2, pp. 89–118.
- DeepSeek-AI (2025). "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". In: *arXiv preprint arXiv:2501.12948*. DOI: [10.48550/arXiv.2501.12948](https://doi.org/10.48550/arXiv.2501.12948). arXiv: [2501.12948](https://arxiv.org/abs/2501.12948) [cs.CL]. URL: <https://arxiv.org/abs/2501.12948>.
- Degner, Lesley F et al. (2003). "A new approach to eliciting meaning in the context of breast cancer". In: *Cancer nursing* 26.3, pp. 169–178.
- Deignan, Alice and Elena Semino (2019). "Translating science for young people through metaphor". In: *The Translator* 25.4, pp. 369–384.
- Deignan, Alice, Elena Semino, and Shirley-Anne Paul (2019). "Metaphors of climate science in three genres: Research articles, educational texts, and secondary school student talk". In: *Applied Linguistics* 40.2, pp. 379–403.
- Del Corro, Luciano and Rainer Gemulla (2013). "Clausie: clause-based open information extraction". In: *Proceedings of the 22nd international conference on World Wide Web*, pp. 355–366.
- Demaidi, Mona Nabil and Mohamed Medhat Gaber (2019). "Tone: A method for terminological ontology evaluation". In: *Proceedings of the ArabWIC 6th Annual International Conference Research Track*, pp. 1–10.
- Demner-Fushman, Dina, Willie J Rogers, and James G Mork (2021). "Automatic recognition of figurative language in biomedical articles." In: *SDU@ AAAI*.
- Dentella, Vittoria, Fritz Günther, and Evelina Leivada (2023). "Systematic Testing of Three Language Models Reveals Low Language Accuracy, Absence of Response Stability, and a Yes-Response Bias". In: *Proceedings of the National Academy of Sciences* 120.51, e2309583120. DOI: [10.1073/pnas.2309583120](https://doi.org/10.1073/pnas.2309583120).
- Dev, Sunipa et al. (2020). "Measuring and Mitigating Unintended Bias in Text Classification". In: *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, pp. 92–98.
- Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.
- Di Nuzzo, Dario et al. (2024). "Automated Generation of Competency Questions Using Large Language Models and Knowledge Graphs". In: *SEMANTiCS Conference*.

- Didolkar, Aniket et al. (2024). "Metacognitive capabilities of llms: An exploration in mathematical problem solving". In: *Advances in Neural Information Processing Systems* 37, pp. 19783–19812.
- Dijkstra, Edsger W et al. (1989). "On the cruelty of really teaching computing science". In: *Communications of the ACM* 32.12, pp. 1398–1404.
- Dingemanse, Mark and Nick J Enfield (2024). "Interactive repair and the foundations of language". In: *Trends in Cognitive Sciences* 28.1, pp. 30–42.
- Djulgovic, Benjamin and Iztok Hozo (June 2014). "Effect of initial conditions on reproducibility of scientific research". In: *Acta Informatica Medica* 22.3, pp. 156–159. ISSN: 0353-8109. DOI: [10.5455/aim.2014.22.156-159](https://doi.org/10.5455/aim.2014.22.156-159). URL: <https://doi.org/10.5455/aim.2014.22.156-159>.
- Do Dinh, Erik-Lân and Iryna Gurevych (2016). "Token-level metaphor detection using neural networks". In: *Proceedings of the Fourth Workshop on Metaphor in NLP*, pp. 28–33.
- Dodge, Ellen K, Jisup Hong, and Elise Stickles (2015). "MetaNet: Deep semantic automatic metaphor analysis". In: *Proceedings of the Third Workshop on Metaphor in NLP*, pp. 40–49.
- Donne, Carlo Delle (2020). "Searching for a Dialogue: The Need to Share a Common Language in Greek Medical Writings". In: *AION-SEZ. DI FILOLOGIA E LETTERATURA CLASSICA* 42, pp. 1–18. DOI: [10.1163/17246172-40010037](https://doi.org/10.1163/17246172-40010037). URL: <https://doi.org/10.1163/17246172-40010037>.
- Donoho, David L. et al. (2009). "Reproducible Research in Computational Harmonic Analysis". In: *Computing in Science & Engineering* 11.1, pp. 8–18. DOI: [10.1109/MCSE.2009.15](https://doi.org/10.1109/MCSE.2009.15).
- Dou, Jinteng, Meichun Liu, and Tong Chen (2023). "Embodiment of color metaphor: an image-based visual analysis of the Chinese color terms hēi 'black' and bái 'white'". In: *Language and Cognition* 15.3, pp. 574–600.
- Dragisic, Zlatan et al. (2017). "Experiences from the anatomy track in the ontology alignment evaluation initiative". In: *J Biomed Semant* 8.1, p. 56. DOI: [10.1186/s13326-017-0166-5](https://doi.org/10.1186/s13326-017-0166-5). URL: <https://doi.org/10.1186/s13326-017-0166-5>.
- Draicchio, Francesco et al. (2013). "Fred: From natural language text to rdf and owl in one click". In: *Extended semantic web conference*. Springer, pp. 263–267.
- Dunn, Jonathan (2010). "Gradient semantic intuitions of metaphoric expressions". In: *Metaphor and Symbol* 26.1, pp. 53–67.
- (2013). "Metaphor Identification with Cross-Linguistic Validation". In: *Proceedings of the First Workshop on Metaphor in NLP*, pp. 13–21.
- Dunn, Jonathan et al. (2014). "Language-independent ensemble approaches to metaphor identification". In: pp. 6–12.
- Eco, Umberto (1976). *A Theory of Semiotics*. Bloomington: Indiana University Press.
- (1986). *Semiotics and the Philosophy of Language*. Vol. 398. Indiana University Press.
- (2014). *From the Tree to the Labyrinth*. Harvard University Press.
- El Kadiri, Soumaya and Dimitris Kiritsis (2015). "Ontologies in the context of product lifecycle management: state of the art literature review". In: *International Journal of Production Research* 53.18, pp. 5657–5668.
- El Refaie, Elisabeth (2003). "Understanding visual metaphor: The example of newspaper cartoons". In: *Visual Communication* 2.1, pp. 75–95. DOI: [10.1177/1470357203002001755](https://doi.org/10.1177/1470357203002001755).
- (2009). "Metaphor in political cartoons: Exploring audience responses". In: *Multi-modal Metaphor*. Ed. by Charles J. Forceville and Eduardo Urios-Aparisi. Berlin/New York: Mouton de Gruyter, pp. 173–196. DOI: [10.1515/9783110215366.3.173](https://doi.org/10.1515/9783110215366.3.173).

- ElHassouni, Jalil and Abderrahim El Qadi (2021). "Ontology engineering methodologies: State of the art". In: *International Conference On Big Data and Internet of Things*. Springer, pp. 59–72.
- Erxleben, Fredo et al. (2014). "Introducing wikidata to the linked data web". In: *International semantic web conference*. Springer, pp. 50–65.
- Ethayarajh, Kawin (2019). "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings". In: *arXiv preprint arXiv:1909.00512*.
- ETSI (2023). SAREF4WATR: Smart Applications REFerence ontology for Water. <https://saref.etsi.org/saref4watr/v1.1.1/>. Accessed: 2024-07-10.
- European Union (2000). Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy. <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32000L0060&rid=2>. Accessed: 2024-07-10.
- Euzenat, Jérôme, Pavel Shvaiko, et al. (2007). *Ontology matching*. Vol. 18. Springer.
- Euzenat, Jérôme et al. (Jan. 2011). "Ontology Alignment Evaluation Initiative: Six Years of Experience". In: *J. Data Semantics* 15, pp. 158–192. DOI: [10.1007/978-3-642-22630-4_6](https://doi.org/10.1007/978-3-642-22630-4_6).
- Fader, Anthony, Stephen Soderland, and Oren Etzioni (2011). "Identifying relations for open information extraction". In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 1535–1545.
- Fainsilber, Lynn and Andrew Ortony (1987). "Metaphorical uses of language in the expression of emotions". In: *Metaphor and Symbol* 2.4, pp. 239–250.
- Fass, Dan (1991). "met*: A method for discriminating metonymy and metaphor by computer". In: *Computational linguistics* 17.1, pp. 49–90.
- Fatehi, Amirreza et al. (2022). "Medical metaphors: increasing clarity but at what cost?" In: *Archives of Bone and Joint Surgery* 10.8, p. 721.
- Fathallah, Nadeen, Steffen Staab, and Alsayed Algergawy (2024). *LLMs4Life: Large Language Models for Ontology Learning in Life Sciences*.
- Fathallah, Nadeen et al. (2024). "Neon-GPT: a large language model-powered pipeline for ontology learning". In: *European Semantic Web Conference*. Springer, pp. 36–50.
- Fauconnier, Gilles and Mark Turner (2003). "Conceptual blending, form and meaning". In: *Recherches en communication* 19, pp. 57–86.
- Fauconnier, Gilles and Mark B Turner (2008). "Rethinking metaphor". In: *Metaphor and Symbol* 23.4, pp. 257–281.
- Fellbaum, Christiane (2010). "WordNet". In: *Theory and applications of ontology: computer applications*. Springer, pp. 231–243.
- Fereralda, Iska Agnesya, Shanty AYPS Duwila, and Yulis Setyowati (2022). "The use of conceptual metaphor in cancer patient stories on a cancer center website". In: *EL2J (English Language and Literature Journal)* 1.2, pp. 1–11.
- Fernández-Izquierdo, Alba, María Poveda-Villalón, and Raúl García-Castro (2019). "CORAL: A Corpus of Ontological Requirements Annotated with Lexico-Syntactic Patterns". In: *Proceedings of the 16th Extended Semantic Web Conference, ESWC 2019*. Springer International Publishing, pp. 443–458. DOI: [10.1007/978-3-030-21348-0_29](https://doi.org/10.1007/978-3-030-21348-0_29).
- Fernández-López, Mariano, Asunción Gómez-Pérez, and Natalia Juristo Juzgado (1997). "Methontology: from ontological art towards ontological engineering". In: *Proceedings of the 1997 International Semantic Web Conference*. Springer, pp. 1–11.
- Fidler, Fiona and John Wilcox (2018). "Reproducibility of scientific results". In: *Proceedings of the 2018 ACM Conference on Empirical Methods in Natural Language Processing*. Association for Computing Machinery, pp. 1–11.
- Fillmore, Charles J (1975). "An alternative to checklist theories of meaning". In: *Annual Meeting of the Berkeley Linguistics Society*, pp. 123–131.

- Fillmore, Charles J. (2006). "Frame semantics". In: *Cognitive linguistics: Basic readings* 34, pp. 373–400.
- Finley, Kate (2025). "Embodied cognition and the grip of computational metaphors". In: *Ergo an Open Access Journal of Philosophy* 12.
- Floch, Jean-Marie (1985). *Petites mythologies de l'œil et de l'esprit. Pour une sémiotique plastique*.
- Floch, Jean-Marie, John F Sherry, and Robin Orr Bodkin (2001). *Semiotics, marketing, and communication: Beneath the signs, the strategies*. Springer.
- Flusberg, Stephen J et al. (2024). "The psychology of framing: How everyday language shapes the way we think, feel, and act". In: *Psychological Science in the Public Interest* 25.3, pp. 105–161.
- Forceville, Charles (2002). "The identification of target and source in pictorial metaphors". In: *Journal of pragmatics* 34.1, pp. 1–14.
- Forgács, Bálint (2024). "A medical language for climate discourse". In: *Frontiers in Climate* 6, p. 1384753. DOI: [10.3389/fclim.2024.1384753](https://doi.org/10.3389/fclim.2024.1384753). URL: <https://doi.org/10.3389/fclim.2024.1384753>.
- Frey, Johannes et al. (2024a). "Assessing the Evolution of LLM capabilities for Knowledge Graph Engineering in 2023". In: *Proceedings of the ESWC2024 Special Track: Large Language Models for Knowledge Engineering (to appear)*.
- Frey, Johannes et al. (2024b). "Assessing the evolution of LLM capabilities for knowledge graph engineering in 2023". In: *European Semantic Web Conference*. Springer, pp. 51–60.
- Fussell, Susan and Mallie Moss (1998). "Figurative language in descriptions of emotional states". In: *Social and cognitive approaches to interpersonal communication*. Lawrence Erlbaum.
- Galison, Peter (1997). *Image and Logic: A Material Culture of Microphysics*. Chicago, IL: University of Chicago Press. ISBN: 9780226279169.
- Gangemi, Aldo (2005). "Ontology design patterns for semantic web content". In: *International semantic web conference*. Springer, pp. 262–276.
- (2008). "Norms and plans as unification criteria for social collectives". In: *Autonomous Agents and Multi-Agent Systems* 17, pp. 70–112.
- (2020). "Closing the loop between knowledge patterns in cognition and the semantic web". In: *Semantic Web* 11.1, pp. 139–151.
- Gangemi, Aldo, Mehwish Alam, and Valentina Presutti (2018). "Amnestic forgery: An ontology of conceptual metaphors". In: *Formal Ontology in Information Systems*. IOS Press, pp. 159–172.
- Gangemi, Aldo and Peter Mika (2003b). "Understanding the semantic web through descriptions and situations". In: *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, pp. 689–706.
- (2003a). "Understanding the Semantic Web through Descriptions and Situations". In: *On The Move to Meaningful Internet Systems 2003: ODBASE*. Vol. 2888. LNCS. Springer, pp. 689–706. URL: <https://www.loa.istc.cnr.it/old/Papers/ODBASE-CONTEXT.pdf>.
- Gangemi, Aldo and Andrea Giovanni Nuzzolese (2025). "Logic augmented generation". In: *Journal of Web Semantics* 85, p. 100859.
- Gangemi, Aldo and Valentina Presutti (2009). "Ontology design patterns". In: *Handbook on ontologies*. Springer, pp. 221–243.
- (2010). "Towards a pattern science for the semantic web". In: *Semantic Web* 1.1-2, pp. 61–68.

- (2022). “Formal Representation and Extraction of Perspectives”. In: *Creating a More Transparent Internet*. Studies in Natural Language Processing. Cambridge University Press, 208–228.
- Gangemi, Aldo et al. (2002). “Sweetening ontologies with DOLCE”. In: *International conference on knowledge engineering and knowledge management*. Springer, pp. 166–181.
- Gangemi, Aldo et al. (2006). “Modelling ontology evaluation and validation”. In: *European semantic web conference*. Springer, pp. 140–154.
- Gangemi, Aldo et al. (2016). “Framester: A wide coverage linguistic linked data hub”. In: *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20*. Springer, pp. 239–254.
- Gangemi, Aldo et al. (2017). “Semantic web machine reading with FRED”. In: *Semantic Web 8.6*, pp. 873–893.
- Gangemi, Aldo et al. (2022). “Automatically Drafting Ontologies from Competency Questions with FrODO”. In: *Towards a Knowledge-Aware AI*. Studies on the Semantic Web. IOS Press. DOI: [10.3233/ssw220014](https://doi.org/10.3233/ssw220014).
- Gangemi, Aldo et al. (2023). “Text2AMR2FRED, a Tool for Transforming Text into RDF/OWL Knowledge Graphs via Abstract Meaning Representation.” In: *ISWC (Posters/Demos/Industry)*.
- Gao, Yunfan et al. (2023). “Retrieval-Augmented Generation for Large Language Models: A Survey”. In: *arXiv preprint arXiv:2312.10997*. URL: <https://arxiv.org/abs/2312.10997>.
- Garcia, Alexander et al. (2010). “Developing ontologies within decentralised settings”. In: *Semantic e-Science*. Springer, pp. 99–139.
- Garello, Stefana et al. (2024). “The role of embodied cognition in action language comprehension in L1 and L2”. In: *Scientific Reports* 14.1, p. 12781.
- Garijo, Daniel et al. (2024). “LLMs for ontology engineering: a landscape of tasks and benchmarking challenges”. In: *The Semantic Web-ISWC*.
- Gayathri, R et al. (2025). “Spatio-Temporal and Semantic Enhanced Context-Aware Navigation for Indoor Robots”. In: *IEEE Access*.
- Genovesi, Chris (2023). “Grice’s Café: coffee, cream, and metaphor comprehension”. In: *Frontiers in Communication* 8, p. 1175587.
- Gentner, Dedre (1983). “Structure-mapping: A theoretical framework for analogy”. In: *Cognitive science* 7.2, pp. 155–170.
- (Feb. 1988). “Metaphor as Structure Mapping: The Relational Shift”. In: *Child Development* 59.1, pp. 47–59. DOI: [10.2307/1130388](https://doi.org/10.2307/1130388). URL: <https://www.jstor.org/stable/1130388>.
- Gentner, Dedre, Keith J Holyoak, and Boicho N Kokinov (2001). *The analogical mind: Perspectives from cognitive science*. MIT press.
- GeoNames geographical database (2023). URL: <http://www.geonames.org/>.
- Gibbs Raymond W., Jr. (1999). *Intentions in the Experience of Meaning*. Cambridge: Cambridge University Press. ISBN: 0521572452.
- Gibbs, Raymond W (2015). “Counting metaphors: What does this reveal about language and thought?” In: *Cognitive Semantics* 1.2, pp. 155–177.
- Gibbs Jr, Raymond W (1984). “Literal meaning and psychological theory”. In: *Cognitive science* 8.3, pp. 275–304.
- (2000). “Making good psychology out of blending theory.” In: *Cognitive linguistics* 11.
- (2017). *Metaphor wars: Conceptual metaphors in human life*. Cambridge University Press.

- Gibbs Jr, Raymond W (2023). "Pragmatic complexity in metaphor interpretation". In: *Cognition* 237, p. 105455.
- Gibbs Jr, Raymond W and Herbert L Colston (2012). *Interpreting figurative meaning*. Cambridge University Press.
- Gibbs Jr, Raymond W and Heather Franks (2002). "Embodied metaphor in women's narratives about their experiences with cancer". In: *Health communication* 14.2, pp. 139–165.
- Gibbs Jr, Raymond W and Josie Siman (2021). "How we resist metaphors". In: *Language and Cognition* 13.4, pp. 670–692.
- Gibson, James J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
- Giglou, Hamed Babaei et al. (2024). *LLMs4OM: Matching Ontologies with Large Language Models*. arXiv: 2404.10317 [cs.AI]. URL: <https://arxiv.org/abs/2404.10317>.
- Gildea, Patricia and Sam Glucksberg (1983). "On understanding metaphor: The role of context". In: *Journal of verbal learning and verbal behavior* 22.5, pp. 577–590.
- Giora, Rachel (2003). *On our mind: Salience, context, and figurative language*. Oxford: Oxford University Press.
- Giora, Rachel and Ofer Fein (1999). "Irony comprehension: The graded salience hypothesis". In: *Humor: International Journal of Humor Research*.
- Giunchiglia, Fausto et al. (2012). "A facet-based methodology for the construction of a large-scale geospatial ontology". In: *Journal on data semantics* 1.1, pp. 57–73.
- Glebkin, Vladimir (2013). "A?ritical View on Conceptual Blending Theory". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 35. 35.
- Glucksberg, Sam (2008). "How metaphors create categories—quickly". In: *The Cambridge handbook of metaphor and thought*, pp. 67–83.
- Glucksberg, Sam and Boaz Keysar (1990). "Understanding metaphorical comparisons: Beyond similarity." In: *Psychological review* 97.1, p. 3.
- Glucksberg, Sam and Matthew S McGlone (2001). *Understanding figurative language: From metaphor to idioms*. 36. Oxford University Press.
- Goel, Ashok K and Sambasiva R Bhatta (2004). "Use of design patterns in analogy-based design". In: *Advanced Engineering Informatics* 18.2, pp. 85–94.
- Goel, Karan et al. (2021). "Robustness gym: Unifying the NLP evaluation landscape". In: *arXiv preprint arXiv:2101.04840*.
- Goertz, Gary (2006). *Social science concepts: A user's guide*. Princeton University Press.
- Goguen, Joseph (1998). "An introduction to algebraic semiotics, with application to user interface design". In: *International Workshop on Computation for Metaphors, Analogy, and Agents*. Springer, pp. 242–291.
- Gokcesu, Bahriye Selin (2009). "Comparison, categorization, and metaphor comprehension". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 31. 31.
- Gong, Hongyu et al. (2020). "IlliniMet: Illinois system for metaphor detection with contextual and linguistic information". In: *Proceedings of the second workshop on figurative language processing*, pp. 146–153.
- Goodman, Steven N., Daniele Fanelli, and John P. A. Ioannidis (2016). "What does research reproducibility mean?" In: *Science Translational Medicine* 8.341, 341ps12–341ps12. DOI: 10.1126/scitranslmed.aaf5027. eprint: <https://www.science.org/doi/pdf/10.1126/scitranslmed.aaf5027>. URL: <https://www.science.org/doi/abs/10.1126/scitranslmed.aaf5027>.
- Grady, Joseph E. (1997). *Foundations of Meaning: Primary Metaphors and Primary Scenes*. University of California, Berkeley.

- Gréa, Philippe and Guillaume Desagulier (2023). "Arbre formel ou forêt conceptuelle ?" fr. In: *Corela. Cognition, représentation, langue* HS-39. ISSN: 1638-573X. DOI: [10.4000/corela.15980](https://doi.org/10.4000/corela.15980). URL: <https://journals.openedition.org/corela/15980> (visited on 10/14/2025).
- Grice, H Paul (1975). "Logic and conversation". In: *Speech acts* 3, pp. 41–58.
- Group, Pragglejaz (2007). "MIP: A method for identifying metaphorically used words in discourse". In: *Metaphor and symbol* 22.1, pp. 1–39.
- Gruber, Thomas R (1993). "A translation approach to portable ontology specifications". In: *Knowledge acquisition* 5.2, pp. 199–220.
- Grüninger, Michael and Mark S Fox (1995). "The role of competency questions in enterprise engineering". In: *Benchmarking—Theory and practice*. Springer, pp. 22–31.
- Guarino, Nicola, Daniel Oberle, and Steffen Staab (2009). "What is an ontology?" In: *Handbook on ontologies*, pp. 1–17.
- Guizzardi, Giancarlo and João Paulo Andrade Almeida (2020). "Stability Patterns in Ontology-Driven Conceptual Modeling". In: *Proceedings of the XIII Seminar on Ontology Research in Brazil and IV Doctoral and Masters Consortium on Ontologies (ONTOBRAS 2020)*. Vol. 2728. CEUR-WS, pp. 148–160.
- Hall, David L. and James Llinas (1997). "An Introduction to Multisensor Data Fusion". In: *Proceedings of the IEEE* 85.1, pp. 6–23.
- Halpin, Harry (2013). *Social semantics*. Springer.
- Hamilton, Kyle (2021). "Towards an ontology for propaganda detection in news articles". In: *The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers* 18. Springer, pp. 230–241.
- Hammar, Karl (2014). "Ontology design patterns: Improving findability and composition". In: *European Semantic Web Conference*. Springer, pp. 3–13.
- Hammar, Karl, Feiyu Lin, and Vladimir Tarasov (2010). "Information reuse and interoperability with ontology patterns and linked data: first experiences from the expertfinder project". In: *International Conference on Business Information Systems*. Springer, pp. 168–179.
- Hanna, Elias and Alija Levic (2023). *Comparative analysis of language models: hallucinations in ChatGPT: Prompt Study*.
- Hannam, James (2011). *The Genesis of Science: How the Christian Middle Ages Launched the Scientific Revolution*. Washington, DC: Regnery.
- Harnad, Stevan (1990). "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1–3, pp. 335–346. DOI: [10.1016/0167-2789\(90\)90087-6](https://doi.org/10.1016/0167-2789(90)90087-6).
- Harris, Richard J (1976). "Comprehension of metaphors: A test of the two-stage processing model". In: *Bulletin of the Psychonomic Society* 8.4, pp. 312–314.
- Harrow, Ian et al. (2017). "Matching disease and phenotype ontologies in the ontology alignment evaluation initiative". In: *Journal of Biomedical Semantics* 8.1, p. 55. ISSN: 2041-1480. DOI: [10.1186/s13326-017-0162-9](https://doi.org/10.1186/s13326-017-0162-9). URL: <https://doi.org/10.1186/s13326-017-0162-9>.
- Hart, Christopher (2021). "Animals vs. armies: Resistance to extreme metaphors in anti-immigration discourse". In: *Journal of Language and Politics* 20.2, pp. 226–253.
- Hauser, David J. and Norbert Schwarz (2015). "The war on prevention: Bellicose cancer metaphors hurt (some) prevention intentions". In: *Personality and Social Psychology Bulletin* 41.1, pp. 66–77.
- He, Xiaoyu et al. (2024a). "VIEMF: Multimodal metaphor detection via visual information enhancement with multimodal fusion". In: *Information Processing & Management* 61.3, p. 103652.

- He, Yuan et al. (2022). "Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching". In: *The Semantic Web – ISWC 2022*. Ed. by Ulrike Sattler et al. Cham: Springer International Publishing, pp. 575–591. ISBN: 978-3-031-19433-7.
- He, Yuan et al. (2023). *Exploring Large Language Models for Ontology Alignment*. arXiv: 2309.07172 [cs.AI]. URL: <https://arxiv.org/abs/2309.07172>.
- He, Yuan et al. (2024b). "Deeponto: A python package for ontology engineering with deep learning". In: *Semantic Web 15.5*, pp. 1991–2004.
- Hearst, MA (1992). "Automatic acquisition of hyponyms from large text corpora in proc". In: *14th International Conference Computational Linguistics, Nantes France*.
- Heilig, Christoph (2024). *GPT-5 is a Terrible Storyteller (and That's an AI Safety Problem)*. URL: <https://www.christoph-heilig.de/en/post/gpt-5-is-a-terrible-storyteller-and-that-s-an-ai-safety-problem> (visited on 01/02/2025).
- Hertling, Sven and Heiko Paulheim (Dec. 2023). "OLaLa: Ontology Matching with Large Language Models". In: *Proceedings of the 12th Knowledge Capture Conference 2023. K-CAP '23*. ACM. DOI: [10.1145/3587259.3627571](https://doi.org/10.1145/3587259.3627571).
- Hesse, Jacob (2023). "Machines and metaphors: Challenges for the detection, interpretation and production of metaphors by computer programs". In: *Theoria* 89.5, pp. 607–624.
- Hesse, Mary (1965). *Models and analogies in science*.
- Hicke, Rebecca MM and Ross Deans Kristensen-McLachlan (2024). "SCIENCE IS EXPLORATION: Computational Frontiers for Conceptual Metaphor Theory". In: *arXiv preprint arXiv:2410.08991*.
- Hitzler, Pascal et al. (2009). "OWL 2 web ontology language primer". In: *W3C recommendation* 27.1, p. 123.
- Hoffman, Robert R (1995). "Monster analogies". In: *AI magazine* 16.3, pp. 11–11.
- (2018). "Metaphor in science". In: *Cognition and figurative language*. Routledge, pp. 393–424.
- Hoffman, Robert R and Richard P Honeck (1976). "The bidirectionality of judgments of synonymy". In: *Journal of Psycholinguistic Research* 5.2, pp. 173–184.
- Hogan, Aidan et al. (2021). "Knowledge graphs". In: *ACM Computing Surveys (Csur)* 54.4, pp. 1–37.
- Holtzman, Ari et al. (2020). "The Curious Case of Neural Text Degeneration". In: *International Conference on Learning Representations*.
- Holyoak, Keith J and Dušan Stamenković (2018). "Metaphor comprehension: A critical review of theories and evidence." In: *Psychological bulletin* 144.6, p. 641.
- Hookway, Branden (2014). *Interface*. Cambridge, MA: The MIT Press. ISBN: 978-0-262-52550-3. DOI: [10.7551/mitpress/9919.001.0001](https://doi.org/10.7551/mitpress/9919.001.0001). URL: <https://mitpress.mit.edu/9780262525503/interface/>.
- Houlsby, Neil et al. (2019). "Parameter-Efficient Transfer Learning for NLP". In: *Proceedings of ICML*.
- Hu, Edward J. et al. (2022). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv: 2106.09685 [cs.CL].
- Huang, Jie (2019). *Handbook of Cognitive Linguistics: Ewa Dąbrowska and Dagmar Divjak (eds)* Berlin and Boston: De Gruyter Mouton, 2015, 716+ viii pp.
- Huang, Jie and Kevin Chen-Chuan Chang (July 2023). "Towards Reasoning in Large Language Models: A Survey". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1049–1065. DOI: [10.18653/v1/2023.findings-acl.67](https://doi.org/10.18653/v1/2023.findings-acl.67). URL: <https://aclanthology.org/2023.findings-acl.67/>.

- Huetter-Almerigi, Yvonne et al. (2025). *Linguistic agency, Un'introduzione*. Biblioteca Clueb.
- Hummel, John E and Keith J Holyoak (2001). "A process model of human transitive inference". In: *Spatial schemas in abstract thought*, pp. 279–305.
- Hussain, Zaeem et al. (2017). "Automatic understanding of image and video advertisements". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1705–1715.
- Husserl, Edmund (1970). *The Crisis of European Sciences and Transcendental Phenomenology*. Trans. by David Carr. Evanston: Northwestern University Press.
- (1983). *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy: First Book (Ideas I)*. Trans. by Fred Kersten. The Hague: Kluwer Academic.
- (2001). *Logical Investigations*. Trans. by J. N. Findlay. 2nd ed. London: Routledge.
- Hutchby, Ian (2001). "Technologies, Texts and Affordances". In: *Sociology* 35.2, pp. 441–456. DOI: [10.1177/S0038038501000219](https://doi.org/10.1177/S0038038501000219).
- Huttar, George L. (1980). "Metaphorical Speech Acts". In: *Poetics* 9.5–6, pp. 383–401. DOI: [10.1016/0304-422X\(80\)90029-6](https://doi.org/10.1016/0304-422X(80)90029-6).
- Hütter-Almerigi, Yvonne (2020). "Two Forms of Realism: Making Sense of Rorty's Controversy with Brandom and Ramberg over Objectivity". In: *European Journal of Pragmatism and American Philosophy* 12.1. DOI: [10.4000/ejpap.1868](https://doi.org/10.4000/ejpap.1868). URL: <https://journals.openedition.org/ejpap/1868>.
- Indurkha, Bipin (1992). "Predictive analogy and cognition". In: *International Workshop on Analogical and Inductive Inference*. Springer, pp. 214–231.
- (2013). *Metaphor and cognition: An interactionist approach*. Vol. 13. Springer Science & Business Media.
- Jacquette, Dale, ed. (2004). *The Cambridge Companion to Brentano*. Cambridge: Cambridge University Press.
- Jahani, Eaman et al. (2024). "Prompt Adaptation as a Dynamic Complement in Generative AI Systems". In: *arXiv preprint arXiv:2407.14333*.
- James, William et al. (1890). *The principles of psychology*. Vol. 1. 2. Macmillan London.
- Jang, Hyeju et al. (2015). "Metaphor detection in discourse". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 384–392.
- Janowicz, Krzysztof and Michael Compton (2010). "The Stimulus-Sensor-Observation Ontology Design Pattern and its Integration into the Semantic Sensor Network Ontology." In: *SSN* 668.
- Jarrar, Mustafa and Robert Meersman (2009). "Ontology engineering—the DOGMA approach". In: *Advances in Web Semantics I*, pp. 7–34.
- Jauhar, Sujay Kumar et al. (2021). "MS-LaTTE: A Dataset of Where and When To-do Tasks are Completed". In: *arXiv preprint arXiv:2111.06902*.
- Jedrzej, Potoniec et al. (2020). "Dataset of ontology competency questions to SPARQL-OWL queries translations". In: *Data in Brief* 29, p. 105098. DOI: [10.1016/j.dib.2019.105098](https://doi.org/10.1016/j.dib.2019.105098).
- Jensen, Thomas Wiben (2018). "The world between us: The social affordances of metaphor in face-to-face interaction". In: *RASK Internationalt tidsskrift for sprog og kommunikation* 47, pp. 45–76.
- (2022). "The Indexical Affordance of Metaphor: Stain as a Case Example". In: *Metaphor and Symbol* 37.3, pp. 208–228.
- Jiang, Albert Q. et al. (2023). *Mistral 7B*. arXiv: [2310.06825](https://arxiv.org/abs/2310.06825) [cs.CL]. URL: <https://arxiv.org/abs/2310.06825>.
- Jiang, Xuhui et al. (2024). "Unlocking the power of large language models for entity alignment". In: *arXiv preprint arXiv:2402.15048*.

- Jiang, Yicun (2017). "What is "meta-" for? A Peircean Critique of the Cognitive Theory of Metaphor". PhD thesis. Hong Kong: Lingnan University. URL: https://commons.ln.edu.hk/eng_etd/14/.
- Jin, Xiongnan et al. (2025). "HLMEA: Unsupervised Entity Alignment Based on Hybrid Language Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 11, pp. 11888–11896.
- Johannsen, Wilhelm (1911). "The genotype conception of heredity". In: *The American Naturalist* 45.531, pp. 129–159.
- Johnson, Gerald J (1994). "Of metaphor and difficulty of computer discourse". In: *Communications of the ACM* 37.12, pp. 97–103.
- Johnson, Mark, George Lakoff, et al. (2002). "Why cognitive linguistics requires embodied realism". In: *Cognitive linguistics* 13.3, pp. 245–264.
- Johnstone, Caitlin and Elise Stickles (2024). "Navigating the climate change minefield: the influence of metaphor on climate doomism". In: *Frontiers in Communication* 9, p. 1380092.
- Jordan, Keller (2023). "On the variance of neural network training with respect to test sets and distributions". In: *arXiv preprint arXiv:2304.01910*.
- Julich-Warpakowski, Nina and Thomas Wiben Jensen (2023). "Zooming in on the notion of metaphoricity: Notions, dimensions, and operationalizations". In: *Metaphor and the Social World* 13.1, pp. 16–36.
- Jurafsky, Daniel and James H. Martin (2023). *Speech and Language Processing*. URL: <https://web.stanford.edu/~jurafsky/slp3/>.
- Kaddour, Jean et al. (2023). *Challenges and Applications of Large Language Models*. arXiv: 2307.10169 [cs.CL]. URL: <https://arxiv.org/abs/2307.10169>.
- Kahneman, Daniel (2011). "Thinking, fast and slow". In: *Farrar, Straus and Giroux*.
- Kaikarytė, Agnė (2020). "Conceptual Metaphors in Popular Medical Discourse". Bachelor's thesis. Šiauliai, Lithuania: Šiauliai University, Institute of Regional Development.
- Kalarani, Abisek Rajakumar, Pushpak Bhattacharyya, and Sumit Shekhar (2024). "Unveiling the Invisible: Captioning Videos with Metaphors". In: *arXiv preprint arXiv:2406.04886*.
- Kampourakis, Kostas (2020). "Why does it matter that many biology concepts are metaphors?" In: *Philosophy of science for biologists*, pp. 102–122.
- Kanthan, Rani and Sheryl Mills (2006). "Using metaphors, analogies and similes as aids in teaching pathology to medical students". In: *The Journal of the International Association of Medical Science Educators* 16.1, pp. 19–26.
- Karam, Naouel et al. (2020). "Matching biodiversity and ecology ontologies: challenges and evaluation results". In: *The Knowledge Engineering Review* 35, e9. DOI: 10.1017/S0269888920000132. URL: <https://doi.org/10.1017/S0269888920000132>.
- Karpathy, Andrej (2025). *Context engineering discussion*. URL: <https://www.llamaindex.ai/blog/context-engineering-what-it-is-and-techniques-to-consider> (visited on 08/21/2025).
- Kasneci, Enkelejda et al. (2023). "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and individual differences* 103, p. 102274.
- Kautz, Henry (2022). "The third AI summer: AAAI Robert S. Engelmore memorial lecture". In: *Ai magazine* 43.1, pp. 105–125.
- Kay, Lily E (2000). *Who wrote the book of life?: A history of the genetic code*. Stanford University Press.
- Keller, Evelyn Fox (2020). "Cognitive functions of metaphor in the natural sciences". In: *Interdisciplinary science reviews* 45.3, pp. 249–267.

- Khatin-Zadeh, Omid et al. (2023). "A review of studies supporting metaphorical embodiment". In: *Behavioral Sciences* 13.7, p. 585.
- Khorashadizadeh, Hanieh et al. (2024). *Research Trends for the Interplay between Large Language Models and Knowledge Graphs*. arXiv: 2406.08223 [cs.AI]. URL: <https://arxiv.org/abs/2406.08223>.
- Kim, Jeongyeon et al. (2023). "Metaphorian: Leveraging large language models to support extended metaphor creation for science writing". In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pp. 115–135.
- Kintsch, Walter (2001). "Predication". In: *Cognitive Science* 25.2, pp. 173–202. DOI: 10.1207/s15516709cog2502_1. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0364021301000349>.
- Kojima, Takeshi et al. (2022). "Large Language Models are Zero-Shot Reasoners". In: *Advances in Neural Information Processing Systems*. arXiv: 2205.11916.
- Koller, Sina, Nadine Müller, and Christina Kauschke (2022). "The elephant in the room: a systematic review of stimulus control in neuro-measurement studies on figurative language processing". In: *Frontiers in Human Neuroscience* 15, p. 791374.
- Kolluru, Keshav et al. (2020a). "Imojie: Iterative memory-based joint open information extraction". In: *arXiv preprint arXiv:2005.08178*.
- Kolluru, Keshav et al. (2020b). "Openie6: Iterative grid labeling and coordination analysis for open information extraction". In: *arXiv preprint arXiv:2010.03147*.
- Koptjevskaja-Tamm, Maria and Dmitry Nikolaev (2021). "Talking about temperature and social thermoregulation in the languages of the world". In: *International Review of Social Psychology* 34.1.
- Kotis, Konstantinos and George A Vouros (2006). "Human-centered ontology engineering: The HCOME methodology". In: *Knowledge and Information Systems* 10.1, pp. 109–131.
- Kotis, Konstantinos I, George A Vouros, and Dimitris Spiliotopoulos (2020). "Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations". In: *The Knowledge Engineering Review* 35, e4.
- Kövecses, Zoltán (2005). *Metaphor in Culture: Universality and Variation*. Cambridge: Cambridge University Press.
- (2008). "Conceptual metaphor theory: Some criticisms and alternative proposals". In: *Annual review of cognitive linguistics* 6.1, pp. 168–184.
- (2015). *Where Metaphors Come From: Reconsidering Context in Metaphor*. Oxford: Oxford University Press.
- (2020). "An extended view of conceptual metaphor theory". In: *Review of Cognitive Linguistics* 18.1, pp. 112–130.
- Krennmayr, Tina and Gerard Steen (2017). "VU Amsterdam metaphor corpus". In: *Handbook of linguistic annotation*. Springer, pp. 1053–1071.
- Kuhn, Thomas (1962). *The nature and necessity of scientific revolutions*. na.
- Kuhn, Thomas S et al. (1979). "Metaphor in science". In: *Metaphor and thought* 2, pp. 533–542.
- Lake, Brenden M. et al. (2017). "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40, e253. DOI: 10.1017/S0140525X16001837.
- Lakoff, George (1987). "Position paper on metaphor". In: *Theoretical Issues in Natural Language Processing* 3.
- (1990). "The Invariance Hypothesis: Is Abstract Reason Based on Image-Schemas?". In: *Cognitive Linguistics* 1.1, pp. 39–74. DOI: 10.1515/cogl.1990.1.1.39.
- (2014). "Mapping the brain's metaphor circuitry: metaphorical thought in everyday reason". In: *Frontiers in human neuroscience* 8, p. 958.

- Lakoff, George, Jane Espenson, and Alan Schwartz (1991). *Master Metaphor List (Draft 2nd Edition)*. Electronic source, Cognitive Linguistics Group, University of California at Berkeley.
- Lakoff, George and Mark Johnson (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.
- (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.
- (2020). “Conceptual metaphor in everyday language”. In: *Shaping entrepreneurship research*. Routledge, pp. 475–504.
- Lakoff, George and Srini Narayanan (2025). *The Neural Mind: How Brains Think*. University of Chicago Press.
- Langacker, Ronald W. (1987). *Foundations of Cognitive Grammar, Volume 1: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
- Langer, Susanne K (2009). *Philosophy in a new key: A study in the symbolism of reason, rite, and art*. Harvard University Press.
- Langlume, Sylvain et al. (2020). “Metaphor use in the ICU: rigor with words!” In: *Intensive Care Medicine* 46.11, pp. 2126–2127.
- Lantow, Birger (2016). “Ontometrics: Putting metrics into use for ontology evaluation.” In: *KEOD*, pp. 186–191.
- Lattmann, Claas (2012). “Icons of Novel Thought: A New Perspective on Peirce’s Definition of Metaphor (CP 2.277)”. In: *Semiotica* 2012.192, pp. 535–556. DOI: [10.1515/sem-2012-0074](https://doi.org/10.1515/sem-2012-0074).
- Leivada, Evelina, Vittoria Dentella, and Fritz Günther (2024). “Evaluating the Language Abilities of Large Language Models vs. Humans: Three Caveats”. In: *Biolinguistics* 18, e14391. DOI: [10.5964/bioling.14391](https://doi.org/10.5964/bioling.14391).
- Lent, Jeremy (2017). “The Dangerous Delusions of Richard Dawkins”. In: *Alternet*, August 3.
- Levinson, Stephen C (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Lewis, Mike et al. (2020a). “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of ACL*, pp. 7871–7880.
- Lewis, Patrick et al. (2020b). “Retrieval-Augmented Generation for Knowledge-Intensive NLP”. In: *Advances in Neural Information Processing Systems*.
- Lewis, Patrick et al. (2020c). “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. URL: <https://arxiv.org/abs/2005.11401>.
- Li, Yucheng et al. (2023). “FrameBERT: Conceptual metaphor detection with frame embedding learning”. In: *arXiv preprint arXiv:2302.04834*.
- Lieto, Antonio, Gian Luca Pozzato, and Stefano Zoia (2025). *The delta of thought: channeling rivers of commonsense knowledge in the sea of metaphorical interpretations*. URL: https://github.com/StefanoZoia/METCL/blob/main/IJCAI__25_Lieto_Pozzato_Zoia.pdf.
- Lightman, Hunter et al. (2023). “Let’s Verify Step by Step”. In: *arXiv preprint arXiv:2305.20050*. URL: <https://arxiv.org/abs/2305.20050>.
- Linell, Per (2009). *Rethinking language, mind, and world dialogically*. IAP.
- Lippolis, Anna Sofia, Giorgia Lodi, and Andrea Giovanni Nuzzolese (2025). “The water health open knowledge graph”. In: *Scientific Data* 12.1, p. 274.
- Lippolis, Anna Sofia et al. (2023). “Enhancing entity alignment between wikidata and artgraph using llms”. In: *Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH)*. Vol. 3540, pp. 1–12.

- Lippolis, Anna Sofia et al. (2024). "Ontogenia: Ontology generation with metacognitive prompting in large language models". In: *European Semantic Web Conference*. Springer, pp. 259–265.
- Lippolis, Anna Sofia et al. (2025a). "Assessing the Capability of Large Language Models for Domain-Specific Ontology Generation". In: *arXiv preprint arXiv:2504.17402*.
- Lippolis, Anna Sofia et al. (2025b). "Ontology generation using large language models". In: *European Semantic Web Conference*. Springer, pp. 321–341.
- Litovkin, Dmitrii et al. (2020). "Suitability of Object-Role Modeling Diagrams as an Intermediate Model for Ontology Engineering: Testing the Rules for Mapping". In: *Biologically Inspired Cognitive Architectures Meeting*. Springer, pp. 188–194.
- Liu, Pengfei et al. (2023). *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*. arXiv: 2107.13586 [cs.CL].
- Liu, Xingbing (2025). "Use of metaphor in provider-patient communication in medical settings: A systematic review". In: *Patient education and counseling* 137, p. 109184.
- Longino, Helen E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press. ISBN: 0691020515.
- Lopes, Alcides et al. (2023). "Using terms and informal definitions to classify domain entities into top-level ontology concepts: An approach based on language models". In: *Knowledge-Based Systems* 265, p. 110385. ISSN: 0950-7051. DOI: <https://doi.org/10.1016/j.knosys.2023.110385>.
- Lotman, Juri and Wilma Clark (2005). "On the Semiosphere". In: *Sign Systems Studies* 33.1, pp. 205–229. DOI: 10.12697/SSS.2005.33.1.09. URL: <https://ojs.utlib.ee/index.php/sss/article/view/SSS.2005.33.1.09>.
- Lu, Yaojie et al. (2022). "Unified structure generation for universal information extraction". In: *arXiv preprint arXiv:2203.12277*.
- Ma, Weicheng et al. (2020). "Improvements and extensions on metaphor detection". In: *arXiv preprint arXiv:2012.04540*.
- Machado, Marcelo et al. (2024). "LLM Store: Leveraging Large Language Models as Sources of Wikidata-Structured Knowledge". In: *International Semantic Web Conference*.
- Maedche, Alexander and Steffen Staab (2001). "Ontology learning for the semantic web". In: *IEEE Intelligent systems* 16.2, pp. 72–79.
- Malheiros, Yuri and Fred Freitas (2017). "Unification in EL for Competency Question Generation." In: *Description Logics*.
- Mangat, Ripan and Simon Dalby (2017). "Climate and wartalk: Metaphors, imagination, transformation". In: *Elementa: Science of the Anthropocene* 5. DOI: 10.1525/elementa.313. URL: <https://doi.org/10.1525/elementa.313>.
- Mao, Rui et al. (2023). "MetaPro Online:: A computational metaphor processing online system". In: *Proceedings of the 61st annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL).
- Mao, Rui et al. (2024). "Understanding public perception towards weather disasters through the lens of metaphor". In: *Proceedings of IJCAI*, pp. 7394–7402.
- Mao, Rui et al. (2025). "A comparative analysis of metaphorical cognition in ChatGPT and human minds". In: *Cognitive Computation* 17.1, p. 35.
- Marcus, Gary (2020). "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence". In: *arXiv abs/2002.06177*. arXiv: 2002.06177 [cs.AI]. URL: <https://arxiv.org/abs/2002.06177>.
- (2024). *AlphaProof, AlphaGeometry, ChatGPT, and why the future of AI is neurosymbolic*. Substack. Accessed February 14, 2026. URL: <https://garymarcus.substack.com/p/alphaproof-alphageometry-chatgpt>.

- Marcus, Gary and Ernest Davis (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.
- Margulis, Lynn and Dorion Sagan (1997). *Microcosmos: Four billion years of microbial evolution*. Univ of California Press.
- Marra, Giuseppe et al. (2024). "From Statistical Relational to Neurosymbolic Artificial Intelligence: A Survey". In: *Artificial Intelligence* 328, p. 104062. DOI: [10.1016/j.artint.2023.104062](https://doi.org/10.1016/j.artint.2023.104062).
- Marron, Jonathan M et al. (2020). *Waging war on war metaphors in cancer and COVID-19*.
- Martin, James H (1994). "Metabank: A knowledge-base of metaphoric language conventions". In: *Computational Intelligence* 10.2, pp. 134–149.
- Martín-Chozas, Patricia, Isam Diab-Lozano, and Elena Montiel-Ponsoda (2024). "Representing Metaphorical Terms in the Semantic Web: A Proposal from a Sociocognitive Perspective". In: *Aspects of Cognitive Terminology Studies: Theoretical Considerations and the Role of Metaphor in Terminology* 55, p. 133.
- Masis-Obando, Rolando, Kenneth A Norman, and Christopher Baldassano (2022). "Schema representations in distinct brain networks support narrative memory during encoding and retrieval". In: *Elife* 11, e70445.
- Mason, Zachary J (2004). "CorMet: a computational, corpus-based conventional metaphor extraction system". In: *Computational linguistics* 30.1, pp. 23–44.
- Matassi, Giorgio, Bud Mishra, and Pedro Martinez (2023). *Current thoughts on the brain-computer analogy—All metaphors are wrong, but some are useful*.
- Mateiu, Patricia and Adrian Groza (2023). "Ontology engineering with large language models". In: *2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. IEEE, pp. 226–229.
- McCusker, Jamie (2023). "LOKE: linked open knowledge extraction for automated knowledge graph construction". In: *arXiv preprint arXiv:2311.09366*.
- McNutt, Marcia (2014). *Reproducibility*.
- Mei, Lingrui et al. (2025). "A Survey of Context Engineering for Large Language Models". In: *arXiv preprint arXiv:2507.13334*.
- Melandri, Enzo (1968). "La linea e il circolo". In: — (1969). "Per una filosofia della metafora". In: *H. Blumenberg, Paradigmi per una Metaforologia*. Bologna: Il Mulino.
- (2025). "Is the Use of Analogy in Law Logically Correct?" Trans. by Alexander Ferguson. In: *Journal of Italian Philosophy* 8, pp. 175–192. URL: <https://research.ncl.ac.uk/italianphilosophy/currentissue/14.%20Melandri%20-%20Is%20the%20Use%20of%20Analogy%20in%20Law%20Logically%20Correct%2C%20FINAL.pdf>.
- Merleau-Ponty, Maurice (1968). *The Visible and the Invisible: Followed by Working Notes*. Ed. by Claude Lefort. Trans. by Alphonso Lingis. Evanston: Northwestern University Press.
- Metamia (n.d.). *Metamia*. <http://www.metamia.com/>. (Visited on 07/15/2025).
- Mey, Jacob (2009). *Concise encyclopedia of pragmatics*. Elsevier Science & Technology.
- Michelli, Gianluca (2025). "Why that metaphor? Steering Deliberate Metaphor to Safer Waters". In: *[Journal Name]*. forthcoming.
- Michelli, Gianluca, Xiaoyu Tong, and Ekaterina Shutova (2024). "A framework for annotating and modelling intentions behind metaphor use". In: *arXiv preprint arXiv:2407.03952*.
- Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv preprint arXiv:1301.3781*. URL: <https://arxiv.org/abs/1301.3781>.

- Miller, George A (1995). "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11, pp. 39–41.
- Min, Sewon et al. (2022). "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work". In: *Proceedings of EMNLP*.
- Minaee, Shervin et al. (2024). *Large Language Models: A Survey*. arXiv: 2402.06196 [cs.CL]. URL: <https://arxiv.org/abs/2402.06196>.
- Minsky, Marvin (1974). *A framework for representing knowledge*.
- Mitchell, Sandra D. (2002). "Integrative Pluralism". In: *Biology & Philosophy* 17.1, pp. 55–70. DOI: [10.1023/A:1012990030867](https://doi.org/10.1023/A:1012990030867).
- Mitrović, Jelena et al. (2017). "Ontological representations of rhetorical figures for argument mining". In: *Argument & Computation* 8.3, pp. 267–287.
- Mittelstadt, Brent, Chris Russell, and Sandra Wachter (2019). "Explaining explanations in AI". In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288.
- Mohammad, Saif, Ekaterina Shutova, and Peter Turney (2016). "Metaphor as a medium for emotion: An empirical study". In: *Proceedings of the fifth joint conference on lexical and computational semantics*, pp. 23–33.
- Mohler, Michael, Marc Tomlinson, and David Bracewell (2013). "Applying Textual Entailment to the Interpretation of Metaphor". In: *Proceedings of the 2013 IEEE 7th International Conference on Semantic Computing*. IEEE, pp. 118–125.
- Mohler, Michael et al. (2016). "Introducing the LCC metaphor datasets". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4221–4227.
- Mollo, Dimitri Coelho and Raphaël Milli re (2023). "The vector grounding problem". In: *arXiv preprint arXiv:2304.01481*.
- Moretti, Franco (2013). "'Operationalizing': or, the function of measurement in modern literary theory". In.
- Motta, Enrico et al. (2024). "Capturing the viewpoint dynamics in the news domain". In: *International Conference on Knowledge Engineering and Knowledge Management*. Springer, pp. 18–34.
- Mulligan, Kevin, Peter Simons, and Barry Smith (1984). "Truth-Makers". In: *Philosophy and Phenomenological Research* 44.3, pp. 287–321. DOI: [10.2307/2107475](https://doi.org/10.2307/2107475).
- Naamati-Schneider, Lior and Gillie Gabay (2022). "Metaphors of War in Effective and Ineffective Coping of Medical Directors of COVID-19 Wards in Public Hospitals". In: *Frontiers in Public Health* 10, p. 830266. DOI: [10.3389/fpubh.2022.830266](https://doi.org/10.3389/fpubh.2022.830266).
- Nacey, Susan et al. (2019). "What the MIPVU protocol doesn't tell you (even though it mostly does)". In: *Metaphor Identification in Multiple Languages*. John Benjamins Publishing Company, pp. 41–67.
- Nadler, Ethan O et al. (2025). "Statistical or embodied? Comparing colorseeing, colorblind, painters, and Large Language Models in their processing of color metaphors". In: *Cognitive Science* 49.7, e70083.
- Narayanan, Srinivas (1999). "Moving right along: A computational model of metaphoric reasoning about events." In: *AAAI/IAAAI* 121127.
- Narayanan, Srinivas Sankara (1997). *KARMA: Knowledge-based active representations for metaphor and aspect*. University of California, Berkeley.
- Nas, Engy and Martin Huschka (2023). *MSE Benchmark*. <https://github.com/EngyNasr/MSE-Benchmark>.
- National Academies of Sciences and Policy and Global Affairs (2019). *Reproducibility and replicability in science*. National Academies Press.

- Nazi, Zabir Al and Wei Peng (2024). "Large language models in healthcare and medical domain: A review". In: *Informatics*. Vol. 11. 3. MDPI, p. 57.
- NEMO-UFES (2023). *DOCE: Domain Ontology for Circular Economy*. <https://nemo-ufes.github.io/doce/>. Accessed: 2024-07-10.
- Nersessian, Nancy J (2008). "15 Mental Modeling in Conceptual Change Nancy J. Nersessian Georgia Institute of Technology". In: *International Handbook of Research on Conceptual Change*, p. 391.
- (2015). "The cognitive work of metaphor and analogy in scientific practice". In: *Philosophical Inquiries* 3.1, pp. 133–156.
- Nezhurina, Marianna et al. (2024). "Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models". In: *arXiv preprint arXiv:2406.02061*.
- Nguyen, Ngoc Thanh (2008). *Advanced methods for inconsistent knowledge management*. Springer.
- Nicholson, Daniel J. (2019). "Is the cell really a machine?" In: *Journal of Theoretical Biology* 477, pp. 108–126. DOI: [10.1016/j.jtbi.2019.06.002](https://doi.org/10.1016/j.jtbi.2019.06.002).
- Nietzsche, Friedrich (1980a). *Götzen-Dämmerung, oder, Wie man mit dem Hammer philosophiert*. Berlin and New York: de Gruyter.
- (1980b). "Über Wahrheit und Lüge im außermoralischen Sinne". In: *Sämtliche Werke: Kritische Studienausgabe (KSA), Bd. 1*. Ed. by Giorgio Colli and Mazzino Montinari. Berlin and New York: de Gruyter.
- Norman, Don (2008). "Signifiers, Not Affordances". In: *interactions* 15.6, pp. 18–19. DOI: [10.1145/1409040.1409044](https://doi.org/10.1145/1409040.1409044).
- (2013). *The Design of Everyday Things: Revised and Expanded Edition*. New York: Basic Books.
- Noy, Natalya F, Deborah L McGuinness, et al. (2001). *Ontology development 101: A guide to creating your first ontology*.
- Noy, Natalya Fridman and Carole D Hafner (1997). "The state of the art in ontology design: A survey and comparative review". In: *AI magazine* 18.3, pp. 53–53.
- Nuzzolese, Andrea Giovanni (2014). "Knowledge Patterns for the Web: extraction, tranformation and reuse". In.
- (2025). "Streamlining Knowledge Graph Creation with PyRML". In: *arXiv preprint arXiv:2505.20949*.
- Nuzzolese, Andrea Giovanni et al. (2011). "Encyclopedic knowledge patterns from wikipedia links". In: *The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I* 10. Springer, pp. 520–536.
- Nuzzolese, Andrea Giovanni et al. (2013). "Tipalo". In.
- Nuzzolese, Andrea Giovanni et al. (2016). "Aemoo: Linked data exploration based on knowledge patterns". In: *Semantic Web* 8.1, pp. 87–112.
- Nuzzolese, Andrea Giovanni et al. (2025). "Semantic Technologies for Global Governance: A Hybrid AI Approach to Tracking and Monitoring WHO Resolutions". In: *European Semantic Web Conference*. Springer, pp. 123–139.
- Oakley, Todd and Seana Coulson (1999). "Blending and metaphor". In: *Metaphor in Cognitive Linguistics*. Ed. Gerard Steen and Raymond Gibbs. Philadelphia: John Benjamins. On line: <http://cogweb.ucla.edu/CogSci/Grady> 99.
- Oliinyk, Oksana and Liudmyla Naumenko (2018). "Metaphoric representation of the concept 'crime' in the US media discourse". In: *Advanced Education*, pp. 134–140.
- Olza, Inés et al. (2021). "The# ReframeCovid initiative: From Twitter to society via metaphor". In: *Metaphor and the Social World* 11.1, pp. 98–120.

- OpenAI (Dec. 2024). *OpenAI o1 System Card*. Tech. rep. OpenAI. URL: <https://cdn.openai.com/o1-system-card-20241205.pdf>.
- Osborne, Francesco et al. (2016). "Automatic Classification of Springer Nature Proceedings with Smart Topic Miner". In: *Proceedings of the 15th International Semantic Web Conference (ISWC), Posters and Demos*.
- Osman, Inès, Salvatore Flavio Pileggi, and Sadok Ben Yahia (2024). "Uncertainty in Automated Ontology Matching: Lessons from an Empirical Evaluation". In: *Applied Sciences* 14.11, p. 4679.
- Ouyang, Long et al. (2022). "Training Language Models to Follow Instructions with Human Feedback". In: *Advances in Neural Information Processing Systems*.
- Ox, Jack (2014). "Analogy and conceptual blending are part of a visualization toolkit for artists and scientists: Introducing the cognitive space transfer". In: *Methods* 1.6, pp. 96–101.
- Palmer, M., D. Gildea, and P. Kingsbury (2005). "The Proposition Bank: An Annotated Corpus of Semantic Roles". In: *Computational Linguistics* 31.1, pp. 71–106.
- Pan, Feng and Jerry R Hobbs (2005). "Temporal Aggregates in OWL-Time." In: *FLAIRS*, pp. 560–565.
- Pan, Xueli et al. (2024). "A RAG Approach for Generating Competency Questions in Ontology Engineering". In: *Research Conference on Metadata and Semantics Research*. Springer, pp. 70–81.
- Panayiotou, Christiana (2024). "Ontological Foundation for Representing Metaphors as part of LOD". In: *IEEE Access*.
- Panzeri, Francesca, Simona Di Paola, and Filippo Domaneschi (2021a). "Does the COVID-19 war metaphor influence reasoning?" In: *Plos one* 16.4, e0250651.
- (2021b). "Does the COVID-19 war metaphor influence reasoning?" In: *PLOS ONE* 16.4, e0250651. DOI: [10.1371/journal.pone.0250651](https://doi.org/10.1371/journal.pone.0250651).
- Park, JM et al. (2008). "Product ontology construction from engineering documents". In: *2008 International Conference on Smart Manufacturing Application*. IEEE, pp. 305–310.
- Patel-Schneider, Peter F. (2014). "Analyzing Schema.org". In: *The Semantic Web – ISWC 2014*. Ed. by Peter Mika et al. Cham: Springer International Publishing, pp. 261–276. ISBN: 978-3-319-11964-9.
- Patil, Rajvardhan and Venkat Gudivada (2024). "A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)". In: *Applied Sciences* 14.5, p. 2074.
- Pearce, Z Reagan and Stephanie E Miller (2025). "Embodied cognition perspectives within early executive function development". In: *Frontiers in Cognition* 4, p. 1361748.
- Peijnenburg, Jeanne and David Atkinson (2003). "When are thought experiments poor ones?" In: *Journal for General Philosophy of Science* 34.2, pp. 305–322.
- (2007). "On poor and not so poor thought experiments. A reply to Daniel Cohnitz". In: *Journal for General Philosophy of Science/Zeitschrift für allgemeine Wissenschaftstheorie* 38.1, pp. 159–161.
- Hartshorne, Charles and Paul Weiss, eds. (1932). *Collected Papers of Charles Sanders Peirce, Volume 2: Elements of Logic*. Cambridge, MA: Harvard University Press.
- Peng, Roger D. (2011). "Reproducible Research in Computational Science". In: *Science* 334.6060, pp. 1226–1227. DOI: [10.1126/science.1213847](https://doi.org/10.1126/science.1213847). eprint: <https://www.science.org/doi/pdf/10.1126/science.1213847>. URL: <https://www.science.org/doi/abs/10.1126/science.1213847>.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on*

- Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.
- Pereira, Francisco Câmara and Amílcar Cardoso (2002). “Conceptual blending and the quest for the holy creative process”. In: *Second Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science, European Conference on Artificial Intelligence (ECAI 2002)*, Lyon, France, July.
- Perlerin, Vincent, Pierre Beust, and Stéphane Ferrari (2003). “Computer-assisted interpretation in domain-specific corpora: the case of the metaphor”. In: *Proceedings of NODALIDA’03, the 14th Nordic Conference on Computational Linguistics*.
- Peroni, Silvio (2016). “A simplified agile methodology for ontology development”. In: *International Experiences and Directions Workshop on OWL*. Springer, pp. 55–69.
- Peroni, Silvio and David Shotton (2012). “FaBiO and CiTO: Ontologies for describing bibliographic resources and citations”. In: *Journal of Web Semantics* 17, pp. 33–43.
- Perret, Nicole and Giuseppe Longo (2016). “Reductionist perspectives and the notion of information”. In: *Progress in Biophysics and Molecular Biology* 122.1, pp. 11–15.
- Peters, Matthew E. et al. (2018). “Deep Contextualized Word Representations”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237.
- Petridis, Savvas and Lydia B Chilton (2019). “Human errors in interpreting visual metaphor”. In: *Proceedings of the 2019 Conference on Creativity and Cognition*, pp. 187–197.
- Petrucci, Giulio, Marco Rospocher, and Chiara Ghidini (2018). “Expressive ontology learning as neural machine translation”. In: *Journal of Web Semantics* 52, pp. 66–82.
- Pettersson, [first name] (2025). “The Affordances of Metaphors for Meaning-Making in Science Education”.
- Pezzulo, Giovanni et al. (2024). “Generating meaning: active inference and the scope and limits of passive AI”. In: *Trends in Cognitive Sciences* 28.2, pp. 97–112.
- Peña, Gil Pena and José de Souza Andrade-Filho (2010). “Analogies in Medicine: Tools for Learning, Reasoning, and Communication”. In: *Medical Education* 44.5, pp. 485–491. DOI: [10.1111/j.1365-2923.2010.03665.x](https://doi.org/10.1111/j.1365-2923.2010.03665.x).
- Phillips, Jacob B, Lenore A Grenoble, and Peggy Mason (2023). “The unembodied metaphor: comprehension and production of tactile metaphors without somatosensation”. In: *Frontiers in Communication* 8, p. 1144018.
- Piantadosi, Steven T and Felix Hill (2022). “Meaning without reference in large language models”. In: *arXiv preprint arXiv:2208.02957*.
- Pineau, Joelle et al. (2021). “Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program)”. In: *Journal of machine learning research* 22.164, pp. 1–20.
- Pinto, Helena Sofia, Steffen Staab, and Christoph Tempich (2004). “DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolving InG Engineering of oNTologies”. In: *ECAI*. Vol. 16, p. 393.
- Plu, Julien et al. (2024). “A Comprehensive Benchmark for Evaluating LLM-Generated Ontologies”. In: *The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings*.
- Polanyi, Michael (2009). “The tacit dimension”. In: *Knowledge in organisations*. Routledge, pp. 135–146.

- Popa-Wyatt, (2017). "Compound Figures: Priority and Speech-Act Structure". In: *Philosophical Studies* 174.1, pp. 141–161. DOI: [10.1007/s11098-016-0629-z](https://doi.org/10.1007/s11098-016-0629-z). URL: <https://www.jstor.org/stable/26002646>.
- (2022). "Compound Figures: A Multi-Channel View of Communication and Psychological Plausibility". In: *Topoi* 41.3, pp. 527–538. DOI: [10.1007/s11245-021-09752-3](https://doi.org/10.1007/s11245-021-09752-3). URL: <https://link.springer.com/article/10.1007/s11245-021-09752-3>.
- Potoniec, Jędrzej et al. (2020). "Dataset of ontology competency questions to SPARQL-OWL queries translations". In: *Data in brief* 29, p. 105098.
- Poveda-Villalón, María, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa (2014). "Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation". In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 10.2, pp. 7–34.
- Poveda-Villalón, María et al. (2022). "LOT: An industrial oriented ontology engineering framework". In: *Engineering Applications of Artificial Intelligence* 111, p. 104755.
- Prashant, Upadhyay (2023). *RAKE: Rapid Automatic Keyword Extraction*. <https://github.com/u-prashant/RAKE>. Accessed: 2024-07-10.
- Presutti, Valentina et al. (2009). "eXtreme design with content ontology design patterns". In: *Proc. Workshop on Ontology Patterns*, pp. 83–97.
- Presutti, Valentina et al. (2016). "From hyperlinks to semantic web properties using open knowledge extraction". In: *Semantic Web* 7.4, pp. 351–378.
- Pulvermüller, Friedemann (2005). "Brain mechanisms linking language and action". In: *Nature Reviews Neuroscience* 6.7, pp. 576–582.
- Quine, Willard Van Orman et al. (1948). *On what there is*. Catholic University of America, Philosophy Education Society Washington, DC.
- Radford, Alec et al. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI technical report. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- Radford, Alec et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. Virtual: PMLR, pp. 8748–8763.
- Raffel, Colin et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67.
- Rai, Sunny et al. (2019). "Understanding metaphors using emotions". In: *New Generation Computing* 37, pp. 5–27.
- Rajan, Sai Sathiesh, Ezekiel Soremekun, and Sudipta Chattopadhyay (2024). "Knowledge-based consistency testing of large language models". In: *arXiv preprint arXiv:2407.12830*.
- Ramberg, Bjørn Torggrim (2000). "Post-Ontological Philosophy of Mind: Rorty versus Davidson". In: *Rorty and His Critics*. Ed. by Robert B. Brandom. Oxford: Blackwell, pp. 351–370.
- Rana, Manisha, Rita Chikkara, and Srishti Sharma (2024). "Computational Metaphor Processing: A Review". In: *2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit)*. IEEE, pp. 618–623.
- Rapp, Christof (2002). "Aristotle's rhetoric". In: — (2002). *Metaphors and Metonymies in Language and Thought*. Cambridge University Press, pp. 1–24.
- Rebboud, Youssra et al. (2024a). "Benchmarking LLM-based Ontology Conceptualization: A Proposal". In: *The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings*.
- Rebboud, Youssra et al. (2024b). "Can LLMs Generate Competency Questions?" In: *European Semantic Web Conference*. Springer, pp. 71–80.

- Rees, Charlotte E, Lynn V Knight, and Clare E Wilkinson (2007). "Doctors being up there and we being down here: a metaphorical analysis of talk about student/doctor-patient relationships". In: *Social science & medicine* 65.4, pp. 725–737.
- Reijnierse, W. Gudrun, Stephanie van Stee, and Margot van Mulken (2025). "The Differential Effects of Metaphor on Comprehensibility and Persuasion in Science Communication". In: *Journal of Science Communication* 24.4. URL: https://jcom.sissa.it/article/pubid/JCOM_2404_2025_A01/.
- Reijnierse, W Gudrun et al. (2015). "How viruses and beasts affect our opinions (or not): The role of extendedness in metaphorical framing". In: *Metaphor and the Social World* 5.2, pp. 245–263.
- Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, pp. 3982–3992.
- Resnik, Philip (1997). "Selectional Preference and Sense Disambiguation". In: *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?* Washington, DC: Association for Computational Linguistics, pp. 52–57.
- Reynolds, Laria and Kyle McDonell (2021). "Prompt programming for large language models: Beyond the few-shot paradigm". In: *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pp. 1–7.
- Richards, Ivor Armstrong (1936). *The Philosophy of Rhetoric*. New York: Oxford University Press.
- Ricoeur, Paul (2004). *The rule of metaphor: The creation of meaning in language*. Routledge.
- Rietveld, Erik and Julian Kiverstein (2014). "A Rich Landscape of Affordances". In: *Ecological Psychology* 26.4, pp. 325–352. DOI: [10.1080/10407413.2014.958035](https://doi.org/10.1080/10407413.2014.958035).
- Ro, Youngbin, Yukyung Lee, and Pilsung Kang (2020). "Multi-OIE: Multilingual Open Information Extraction Based on Multi-Head Attention with BERT". In: *arXiv preprint arXiv:2009.08128*.
- Roche, Christophe (2012). "Ontoterminology: How to unify terminology and ontology into a single paradigm". In: *LREC 2012, Eighth International Conference on Language Resources and Evaluation*. European Language Resources Association, p–2626.
- Rorty, Richard and Mary B. Hesse (1987). "Unfamiliar Noises: Hesse and Davidson on Metaphor". In: *Proceedings of the Aristotelian Society, Supplementary Volume*. Vol. 61, pp. 283–311. DOI: [10.1093/aristoteliansupp/61.1.283](https://doi.org/10.1093/aristoteliansupp/61.1.283).
- Rosen, Zachary (2018). "Computationally constructed concepts: A machine learning approach to metaphor interpretation using usage-based construction grammatical cues". In: *Proceedings of the workshop on figurative language processing*, pp. 102–109.
- Rossi, Maria Grazia, Fabrizio Macagno, and Sarah Bigi (2022). "Dialogical functions of metaphors in medical interactions". In: *Text & Talk* 42.1, pp. 77–103.
- Rothbart, Daniel (1984). "The semantics of metaphor and the structure of science". In: *Philosophy of science* 51.4, pp. 595–615.
- Rougier, Nicolas P et al. (2017). "Sustainable computational science: the ReScience initiative". In: *PeerJ Computer Science* 3, e142.
- Saakyan, Arkadiy et al. (2025). "Understanding Figurative Meaning through Explainable Visual Entailment". In: *Proceedings of the 2025 Conference of the Nations*

- of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pp. 1–23.
- Saba, Walid S. (2023). “Stochastic LLMs do not Understand Language: Towards Symbolic, Explainable and Ontologically Based LLMs”. In: *Conceptual Modeling*. Ed. by João Paulo A. Almeida et al. Cham: Springer Nature Switzerland, pp. 3–19. ISBN: 978-3-031-47262-6.
- Saeedizade, Mohammad Javad and Eva Blomqvist (2024). “Navigating Ontology Development with Large Language Models”. In: *The Semantic Web*. Ed. by Albert Meroño Peñuela et al. Cham: Springer Nature Switzerland, pp. 143–161. ISBN: 978-3-031-60626-7.
- Saif Mohammad Ekaterina Shutova, Peter D. Turney (2016). “A Corpus of English Verbs Annotated for Sense and Metaphoricity”. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pp. 1676–1683.
- Salatino, Angelo A., Francesco Osborne, and Enrico Motta (2018). “The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas”. In: *Proceedings of the 17th International Semantic Web Conference (ISWC)*.
- (2020). *The Computer Science Ontology: Applications in Scholarly Analytics*. arXiv:2003.12611.
- Salles, Arleen, Kathinka Evers, and Michele Farisco (2020). “Anthropomorphism in AI”. In: *AJOB neuroscience* 11.2, pp. 88–95.
- San Pietro, Chiara Barattieri di et al. (2023). “The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent”. In: *Sistemi intelligenti* 35.2, pp. 379–400.
- Sanchez-Bayona, Elisa and Rodrigo Agerri (2025a). *Meta4XNLI: A Crosslingual Parallel Corpus for Metaphor Detection and Interpretation*. arXiv: 2404.07053 [cs.CL]. URL: <https://arxiv.org/abs/2404.07053>.
- (2025b). “Metaphor and Large Language Models: When Surface Features Matter More than Deep Understanding”. In: *arXiv preprint arXiv:2507.15357*.
- Santoro, Adam et al. (2021). “Symbolic behaviour in artificial intelligence”. In: *arXiv preprint arXiv:2102.03406*.
- Schalk, Sami (2013). “Metaphorically speaking: Ableist metaphors in feminist writing”. In: *Disability Studies Quarterly* 33.4.
- Schank, Roger C. and Robert P. Abelson (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Lawrence Erlbaum.
- Schapiro, Samuel, Jonah Black, and Lav R Varshney (2025). “Transformational Creativity in Science: A Graphical Theory”. In: *arXiv preprint arXiv:2504.18687*.
- Scherer, Aaron M, Laura D Scherer, and Angela Fagerlin (2015). “Getting ahead of illness: using metaphors to influence medical decision making”. In: *Medical Decision Making* 35.1, pp. 37–45.
- Schilperoord, Joost and Alfons Maes (2009). “Visual metaphoric conceptualization in editorial cartoons”. In: *Multimodal Metaphor*. Ed. by Charles J. Forceville and Eduardo Urios-Aparisi. Berlin/New York: Mouton de Gruyter, pp. 213–242. DOI: 10.1515/9783110215366.3.213.
- Schloss, Karen B (2024). “Color semantics in human cognition”. In: *Current Directions in Psychological Science* 33.1, pp. 58–67.
- Schmitz, Michael et al. (2012). “Open language learning for information extraction”. In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 523–534.
- Scott, Kate (2021). “Memes as multimodal metaphors: A relevance theory analysis”. In: *Pragmatics & Cognition* 28.2, pp. 277–298. DOI: 10.1075/pc.21010.sco.

- Searle, John R (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press.
- Searle, John R. (1980). "Minds, brains, and programs". In: *Behavioral and Brain Sciences* 3.3, pp. 417–457. DOI: [10.1017/S0140525X00005756](https://doi.org/10.1017/S0140525X00005756).
- (1993). "Metaphor". In: *Metaphor and Thought*. Ed. by Andrew Ortony. 2nd ed. Cambridge: Cambridge University Press, pp. 83–111. DOI: [10.1017/CB09781139173865.008](https://doi.org/10.1017/CB09781139173865.008).
- Searls, David B (1997). "Linguistic approaches to biological sequences". In: *Bioinformatics* 13.4, pp. 333–344.
- Semino, Elena (2008). *Metaphor in discourse*. Cambridge University Press Cambridge.
- (July 2021). "Fire, waves and warfare: the way we make sense of Covid". In: *The Guardian*. URL: <https://www.theguardian.com/commentisfree/2021/jul/05/metaphors-covid-pandemic>.
- Semino, Elena, Zsófia Demjén, and Jane Demmen (2018). "An integrated approach to metaphor and framing in cognition, discourse, and practice, with an application to metaphors for cancer". In: *Applied linguistics* 39.5, pp. 625–645.
- Semino, Elena et al. (2017). "The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study". In: *BMJ supportive & palliative care* 7.1, pp. 60–66.
- Serianni, Luca (2011). *Linguaggi specialistici*. Treccani, § on *rideterminazione semantica*. URL: https://www.treccani.it/enciclopedia/linguaggi-specialistici_%28XXI-Secolo%29/.
- Sfard, Anna (2012). "Metaphors in education". In: *Educational theories, cultures and learning*. Routledge, pp. 39–49.
- Sharma, Vishwas B. (2023). *RAKE-NLTK Documentation*. https://csurfer.github.io/rake-nltk/_build/html/index.html. Accessed: 2024-07-10.
- Shimizu, Cogan, Karl Hammar, and Pascal Hitzler (2020). "Modular graphical ontology engineering evaluated". In: *European Semantic Web Conference*. Springer, pp. 20–35.
- (2023). "Modular ontology modeling". In: *Semantic Web* 14.3, pp. 459–489.
- Shutova, Ekaterina (2010a). "Automatic Metaphor Interpretation as a Paraphrasing Task". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Los Angeles, CA: Association for Computational Linguistics, pp. 1029–1037. URL: <https://aclanthology.org/N10-1147/>.
- (2010b). "Models of Metaphor in NLP". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pp. 688–697.
- (2015). "Design and evaluation of metaphor processing systems". In: *Computational Linguistics* 41.4, pp. 579–623.
- (2017). "Annotation of linguistic and conceptual metaphor". In: *Handbook of linguistic annotation*, pp. 1073–1100.
- Shutova, Ekaterina, Simone Teufel, and Anna Korhonen (2013). "Statistical Metaphor Processing". In: *Computational Linguistics* 39.2, pp. 301–353.
- Silva, Luis Miguel Vieira da et al. (2024). "Toward a method to generate capability ontologies from natural language descriptions". In: *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, pp. 1–4.
- Simons, Peter (1987). *Parts: A Study in Ontology*. Oxford: Oxford University Press.
- Singhal, Amit (2012). *Introducing the Knowledge Graph: Things, Not Strings*. <http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html>.

- Sinnenberg, Lauren et al. (2018). "Content analysis of metaphors about hypertension and diabetes on Twitter: exploratory mixed-methods study". In: *JMIR diabetes* 3.4, e11177.
- Sommerer, Thomas (2025). "Baudrillard and the Dead Internet Theory. Revisiting Baudrillard's (dis) trust in Artificial Intelligence". In: *Philosophy & Technology* 38.2, p. 54.
- Sontag, Susan (1990). *Illness as Metaphor and AIDS and Its Metaphors*. New York: Anchor Books.
- Sperber, Dan and Deirdre Wilson (1995). *Relevance: Communication and cognition*. Oxford: Blackwell.
- Stamenković, Dušan et al. (2023). "An individual-differences approach to poetic metaphor: Impact of aptness and familiarity". In: *Metaphor and Symbol* 38.2, pp. 149–161.
- Stanovich, Keith E and Richard F West (2008). "On the relative independence of thinking biases and cognitive ability." In: *Journal of personality and social psychology* 94.4, p. 672.
- Star, Susan Leigh and James R. Griesemer (1989). "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39". In: *Social Studies of Science* 19.3, pp. 387–420. DOI: [10.1177/030631289019003001](https://doi.org/10.1177/030631289019003001).
- Steen, Gerard (2008). "When is metaphor deliberate". In: *Selected papers from the Stockholm*, pp. 43–63.
- (2017). "Deliberate Metaphor Theory: Basic assumptions, main tenets, urgent issues". In: *Intercultural pragmatics* 14.1, pp. 1–24.
- (2023). "Thinking by metaphor, fast and slow: Deliberate Metaphor Theory offers a new model for metaphor and its comprehension". In: *Frontiers in Psychology* 14, p. 1242888.
- Steen, Gerard et al. (2010). "A method for linguistic metaphor identification". In: — (2015). "Developing, testing and interpreting deliberate metaphor theory". In: *Journal of pragmatics* 90.1, pp. 67–72.
- Steen, Gerard et al. (2019). "MIPVU: A manual for identifying metaphor-related words". In: *Metaphor Identification in Multiple Languages*. John Benjamins Publishing Company, pp. 23–40.
- Steinbock, Anthony J. (1995). *Home and Beyond: Generative Phenomenology after Husserl*. Evanston: Northwestern University Press.
- Stockwell, Peter (1999). "The inflexibility of invariance". In: *Language and literature* 8.2, pp. 125–142.
- Stowe, Kevin, Prasetya Utama, and Smaranda Muresan (2022). "IMPLI: Investigating NLI Models' Performance on Figurative Language". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Stuart, Michael T and Daniel Wilkenfeld (2022). "Understanding metaphorical understanding (literally)". In: *European Journal for Philosophy of Science* 12.3, p. 49.
- Su, Chang, Shuman Huang, and Yijiang Chen (2017). "Automatic detection and interpretation of nominal metaphor based on the theory of meaning". In: *Neurocomputing* 219, pp. 300–311.
- Su, Chang et al. (2024). "Efficient visual metaphor image generation based on metaphor understanding". In: *Neural Processing Letters* 56.3, p. 150.
- Su, Chuandong et al. (2020). "DeepMet: A reading comprehension paradigm for token-level metaphor detection". In: *Proceedings of the second workshop on figurative language processing*, pp. 30–39.

- Suárez-Figueroa, Mari Carmen, Asunción Gómez-Pérez, and Mariano Fernández-López (2011). "The NeOn methodology for ontology engineering". In: *Ontology engineering in a networked world*. Springer, pp. 9–34.
- Tan, Ting Fang et al. (2024). "A proposed score evaluation framework for large language models: Safety, consensus, objectivity, reproducibility and explainability". In: *arXiv preprint arXiv:2407.07666*.
- Tanev, Hristo and Bernardo Magnini (2006). "Weakly supervised approaches for ontology population". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 17–24.
- Tate, Tyler (2020). "Your father's a fighter; your daughter's a vegetable: a critical analysis of the use of metaphor in clinical practice". In: *Hastings Center Report* 50.5, pp. 20–29.
- Taylor, Cynthia and Bryan M Dewsbury (2018). "On the problem and promise of metaphor use in science and science communication". In: *Journal of microbiology & biology education* 19.1, pp. 10–1128.
- Temmerman, Rita (2000). *Towards New Ways of Terminology Description: The Sociocognitive Approach*. John Benjamins.
- Teng, Norman Y. (2009). "Image alignment in multimodal metaphor". In: *Multimodal Metaphor*. Ed. by Charles J. Forceville and Eduardo Urios-Aparisi. Berlin/New York: Mouton de Gruyter, pp. 197–212. DOI: [10.1515/9783110215366.3.197](https://doi.org/10.1515/9783110215366.3.197).
- Thibodeau, Paul H, Les Sikos, and Frank H Durgin (2018). "Are subjective ratings of metaphors a red herring? The big two dimensions of metaphoric sentences". In: *Behavior Research Methods* 50, pp. 759–772.
- Tian, Siqi et al. (2025). "A six-tiered framework for evaluating AI models from repeatability to replaceability". In: *Trends in Biotechnology*.
- Tian, Yuan, Nan Xu, and Wenji Mao (2024). "A Theory Guided Scaffolding Instruction Framework for LLM-Enabled Metaphor Reasoning". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7731–7748.
- Tian, Yuan et al. (2023). "Modeling conceptual attribute likeness and domain inconsistency for metaphor detection". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7736–7752.
- Tietge, David J. (1998). "The Role of Burke's Four Master Tropes in Scientific Expression". In: *Journal of Technical Writing and Communication* 28.4, pp. 317–324.
- Titchkosky, Tanya (2007). *Reading and writing disability differently: The textured life of embodiment*. University of Toronto Press.
- Tong, Xiaoyu et al. (2024). "Metaphor understanding challenge dataset for LLMs". In: *arXiv preprint arXiv:2403.11810*.
- Tongeren, Geraldine W Van Rijn-van (1997). *Metaphors in medical texts*. Vol. 8. Rodopi.
- Touvron, Hugo et al. (2023a). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: [2307.09288](https://arxiv.org/abs/2307.09288) [cs.CL].
- Touvron, Hugo et al. (2023b). "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288*.
- Touvron, Hugo et al. (2023c). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: [2302.13971](https://arxiv.org/abs/2302.13971) [cs.CL].
- Tsaneva, Stefani, Guntur Budi Herwanto, and Marta Sabou (2024). "Benchmarking Ontology Validation Capabilities of LLMs". In: *The Semantic Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings*.

- Tsaneva, Stefani, Stefan Vasic, and Marta Sabou (2024). "LLM-driven ontology evaluation: Verifying ontology restrictions with chatgpt". In: *The semantic web: ESWC satellite events 2024*.
- Tsvetkov, Yulia, Leonid Boytsov, and Anatole Gershman (2014). "Metaphor Detection with Cross-Lingual Model Transfer". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD: Association for Computational Linguistics, pp. 248–258. URL: <https://homes.cs.washington.edu/~yuliats/papers/metaphor-acl14.pdf>.
- Tufek, Nilay et al. (2024). "Validating semantic artifacts with large language models". In: *European Semantic Web Conference*. Springer, pp. 92–101.
- Turbayne, Colin Murray (1962a). "Berkeley's Two Concepts of Mind (Part II)". In: *Philosophy and Phenomenological Research*, pp. 383–386.
- (1962b). *The myth of metaphor*. Yale Univer. Press.
- Turner, Victor (1969). *The Ritual Process: Structure and Anti-Structure*. Chicago: Aldine Publishing Company.
- Turney, Peter D. et al. (2011). "Literal and Metaphorical Sense Identification through Concrete-to-Abstract Mapping". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 680–690. URL: <https://aclanthology.org/D11-1063.pdf>.
- Tversky, Amos (1977). "Features of similarity". In: *Psychological Review* 84.4, pp. 327–352.
- Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty." In: *science* 185.4157, pp. 1124–1131.
- Uceda-Sosa, Rosario et al. (2024). "Reasoning about concepts with LLMs: Inconsistencies abound". In: *arXiv preprint arXiv:2405.20163*.
- Uschold, Mike and Michael Gruninger (1996). "Ontologies: Principles, methods and applications". In: *The knowledge engineering review* 11.2, pp. 93–136.
- Valiant, Leslie G (2003). "Three problems in computer science". In: *Journal of the ACM (JACM)* 50.1, pp. 96–99.
- Varela, Francisco J. (1981). "Autonomy and Autopoiesis". In: *Self-Organizing Systems: An Interdisciplinary Approach*. Ed. by Gerhard Roth and Helmut Schwegler. Frankfurt am Main and New York: Campus Verlag, pp. 14–24.
- Varzi, Achille C (2008). "From language to ontology: Beware of the traps". In: *The categorization of spatial entities in language and cognition*. John Benjamins Publishing Company, pp. 269–284.
- (2011). "Boundaries, conventions, and realism". In: *Carving nature at its joints: Natural kinds in metaphysics and science*, pp. 129–153.
- (2014). "Realism in the Desert". In.
- Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30.
- Veale, Tony (2019). "From conceptual mash-ups to badass blends: A robust computational model of conceptual blending". In: *Computational creativity: The philosophy and engineering of autonomously creative systems*, pp. 71–89.
- Veale, Tony and Guofu Li (2012). "Specifying Viewpoint and Information Need with Affective Metaphors: A System Demonstration of the Metaphor-Magnet Web App/Service". In: *Proceedings of the ACL 2012 System Demonstrations*. Jeju Island, Korea: Association for Computational Linguistics, pp. 7–12. URL: <https://aclanthology.org/P12-3002/>.
- Veale, Tony and Diarmuid O'Donoghue (2001). "Computation and blending". In.

- Venkatasubramanian, Venkat (2025). "Do large language models "understand" their knowledge?" In: *AICHe Journal*, e18661.
- Verberne, Frank MF, Jaap Ham, and Cees JH Midden (2015). "Trusting a virtual driver that looks, acts, and thinks like you". In: *Human factors* 57.5, pp. 895–909.
- Vigo, Markel et al. (2014). "Overcoming the pitfalls of ontology authoring: Strategies and implications for tool design". In: *International Journal of Human-Computer Studies* 72.12, pp. 835–845.
- Volberda, Henk W (2004). "Crisis in strategy: fragmentation, integration or synthesis". In: *European Management Review* 1.1, pp. 35–42.
- Völker, Johanna (2009). *Learning expressive ontologies*. Vol. 2. IOS Press.
- Vrandečić, Denny and Aldo Gangemi (2006). "Unit tests for ontologies". In: *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*. Springer, pp. 1012–1020.
- Wachowiak, Lennart and Dagmar Gromann (2023). "Does GPT-3 Grasp Metaphors? Identifying Metaphor Mappings with Generative Language Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1018–1032.
- Walton, Kendall L. (1990). *Mimesis as Make-Believe: On the Foundations of the Representational Arts*. Cambridge, MA: Harvard University Press.
- Wan, Mingyu et al. (2024). "Perceptual and actional enrichment for metaphor detection with sensorimotor norms". In: *Natural Language Engineering* 30.6, pp. 1181–1209.
- Wang, Boshi et al. (2023a). "Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters". In: *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Wang, Chenguang et al. (2022). "DeepStruct: Pretraining of language models for structure prediction". In: *arXiv preprint arXiv:2205.10475*.
- Wang, Genshan (2022). "Multimodal metaphor construction and cognitive analysis in educational cartoons". In: *Theory and Practice in Language Studies* 12.3, pp. 543–550.
- Wang, Juanjuan et al. (2023b). "The psychological reality of Chinese deliberate metaphors from the reception side: an experimental approach". In: *Brain Sciences* 13.2, p. 160.
- Wang, Lei et al. (July 2023c). "Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 2609–2634. DOI: [10.18653/v1/2023.acl-long.147](https://doi.org/10.18653/v1/2023.acl-long.147). URL: <https://aclanthology.org/2023.acl-long.147>.
- Wang, Sitong et al. (2023d). "PopBlends: Strategies for conceptual blending with large language models". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.
- Wang, Yuqing and Yun Zhao (2023). "Metacognitive prompting improves understanding in large language models". In: *arXiv preprint arXiv:2308.05342*.
- Watson, David (2019). "The rhetoric and reality of anthropomorphism in artificial intelligence". In: *Minds and Machines* 29.3, pp. 417–440.
- Waytz, Adam, Joy Heafner, and Nicholas Epley (2014). "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle". In: *Journal of experimental social psychology* 52, pp. 113–117.
- Wedholm, William (2024). *Exploring the influence of data formats on the consistency of large language models outputs*.

- Wei, Jason et al. (2022). "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems*, pp. 24824–24837.
- Weiland, Hanna, Valentina Bambini, and Petra B Schumacher (2014). "The role of literal meaning in figurative language comprehension: Evidence from masked priming ERP". In: *Frontiers in Human Neuroscience* 8, p. 583.
- Werner, Luisa et al. (2024). "Reproduce, replicate, reevaluate. the long but safe way to extend machine learning methods". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 14, pp. 15850–15858.
- Wharton, Tim (2009). *Pragmatics and Non-Verbal Communication*. Cambridge: Cambridge University Press. ISBN: 0521870976.
- Wierzbicka, Anna (1986). "Human emotions: universal or culture-specific?" In: *American anthropologist* 88.3, pp. 584–594.
- Wijesiriwardene, Thilini et al. (2023). "ANALOGICAL—A Novel Benchmark for Long Text Analogy Evaluation in Large Language Models". In: *arXiv preprint arXiv:2305.05050*.
- Wilson, Deirdre and Dan Sperber (1988). "Representation and Relevance". In: *Mental Representations: The Interface between Language and Reality*. Ed. by Ruth M. Kempson. Cambridge: Cambridge University Press, pp. 133–153.
- Wilson, Deirdre and Tim Wharton (2006). "Relevance and Prosody". In: *Journal of Pragmatics* 38.10, pp. 1559–1579. DOI: [10.1016/j.pragma.2005.04.012](https://doi.org/10.1016/j.pragma.2005.04.012).
- Winograd, Terry (1977). "A Procedural Model of Language Understanding". In: *The Psychology of Computer Vision*. Ed. by Patrick H. Winston. McGraw–Hill, pp. 152–186.
- Winograd, Terry and Fernando Flores (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Ablex.
- Wisniewski, Dawid (2018). "Automatic translation of competency questions into sparql-owl queries". In: *Companion Proceedings of the The Web Conference 2018*, pp. 855–859.
- Witte, René, Ninus Khamis, and Juergen Rilling (2010). "Flexible ontology population from text: The owl exporter". In: *Proceedings of LREC 2010*.
- Wittgenstein, Ludwig (2002). *The Blue and Brown Books: Preliminary Studies for the 'Philosophical Investigations'*. Reprint of the 1958 edition; pagination is the same. Oxford: Blackwell.
- (2005). *The Big Typescript: TS 213*. Ed. and trans. by C. Grant Luckhardt and Maximilian A. E. Aue. Oxford: Blackwell.
- (2009). *Philosophical investigations*. John Wiley & Sons.
- Wnuk, Ewelina and Yuma Ito (2021). "The heart's downward path to happiness: Cross-cultural diversity in spatial metaphors of affect". In: *Cognitive Linguistics* 32.2, pp. 195–218.
- Wolbers, Jeroen, Kees Boersma, and Peter Groenewegen (2018). "Introducing a fragmentation perspective on coordination in crisis management". In: *Organization Studies* 39.11, pp. 1521–1546.
- Workshop, BigScience et al. (2023). "Bloom: A 176b-parameter open-access multilingual language model". In: arXiv: [2211.05100 \[cs.CL\]](https://arxiv.org/abs/2211.05100). URL: <https://arxiv.org/abs/2211.05100>.
- Xu, Bo et al. (2022). "Met-meme: A multimodal meme dataset rich in metaphors". In: *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 2887–2899.
- Yang, Fan-Pei Gloria et al. (2013). "Contextual effects on conceptual blending in metaphors: an event-related potential study". In: *Journal of Neurolinguistics* 26.2, pp. 312–326.

- Yang, Huifan et al. (2022). "Open Relation Extraction with Non-existent and Multi-span Relationships". In: *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*. Vol. 19. 1, pp. 373–383.
- Yang, Linyan, Jingwei Cheng, and Fu Zhang (2024). "Advancing Cross-Lingual Entity Alignment with Large Language Models: Tailored Sample Segmentation and Zero-Shot Prompts". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8122–8138.
- Yang, Linyao et al. (2025). "Integrating knowledge from knowledge graphs and large language models for explainable entity alignment". In: *Information Fusion*, p. 103587.
- Yang, Zi Yun, Ziqing Zhang, and Yisong Miao (2024). "The ELCo Dataset: Bridging Emoji and Lexical Composition". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 15899–15909.
- Yao, Shunyu et al. (2022). "ReAct: Synergizing Reasoning and Acting in Language Models". In: *arXiv preprint arXiv:2210.03629*. URL: <https://arxiv.org/abs/2210.03629>.
- Yao, Shunyu et al. (2023). "Tree of Thoughts: Deliberate Problem Solving with Large Language Models". In: *arXiv preprint arXiv:2305.10601*. URL: <https://arxiv.org/abs/2305.10601>.
- Yao, Zhao et al. (2024). "Conventionality and context jointly modulate the effect of inhibitory control on L2 metaphor comprehension". In: *Humanities and Social Sciences Communications* 11.1, pp. 1–11.
- Yates, Alexander et al. (2007). "Textrunner: open information extraction on the web". In: *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 25–26.
- Ye, Keren and Adriana Kovashka (2018). "Advise: Symbolism and external knowledge for decoding advertisements". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 837–855.
- Yetman, Cameron C (2025). "Metaphor and Metaphilosophy: Wittgenstein, MacDonald, and Conceptual Metaphor Theory". In: *European Journal of Philosophy*, e13038.
- Yosef, Ron, Yonatan Bitton, and Dafna Shahaf (2023). "Irfi: Image recognition of figurative language". In: *arXiv preprint arXiv:2303.15445*.
- Yu, Dengyu et al. (2023). "A survey on neural-symbolic learning systems". In: *Neural Networks*. DOI: [10.1016/j.neunet.2023.06.028](https://doi.org/10.1016/j.neunet.2023.06.028).
- Yu, Ning (2022). *The moral metaphor system: A conceptual metaphor approach*. Oxford University Press.
- Zaitoun, Antonio, Tomer Sagi, and Katja Hose (Apr. 2023). "Automated Ontology Evaluation: Evaluating Coverage and Correctness using a Domain Corpus". In: *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023*. ACM. Austin, TX, USA. DOI: [10.1145/3543873.3587617](https://doi.org/10.1145/3543873.3587617). URL: <https://doi.org/10.1145/3543873.3587617>.
- Zaitoun, Antonio et al. (2023). *Can Large Language Models Augment a Biomedical Ontology with missing Concepts and Relations?* arXiv: [2311.06858 \[cs.CL\]](https://arxiv.org/abs/2311.06858). URL: <https://arxiv.org/abs/2311.06858>.
- Zeng, Qiuhai et al. (2025). "An Analyst-Inspector Framework for Evaluating Reproducibility of LLMs in Data Science". In: *arXiv preprint arXiv:2502.16395*.
- Zhan, Junlang and Hai Zhao (2020). "Span model for open information extraction on accurate corpus". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 9523–9530.

- Zhang, Bohui et al. (2024). "OntoChat: a framework for conversational ontology engineering using language models". In: *European Semantic Web Conference*. Springer, pp. 102–121.
- Zhang, Bohui et al. (2025). "Trustworthy Knowledge Graphs: Practices and Approaches". In: *Handbook on Neurosymbolic AI and Knowledge Graphs*. IOS Press, pp. 363–384.
- Zhang, Dongyu et al. (2021). "MultiMET: A multimodal dataset for metaphor understanding". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3214–3225.
- Zhang, Dongyu et al. (2023a). "Multicmet: A novel chinese benchmark for understanding multimodal metaphor". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6141–6154.
- Zhang, Rui et al. (2023b). "Autoalign: Fully automatic and effective knowledge graph alignment enabled by large language models". In: *IEEE Transactions on Knowledge and Data Engineering* 36.6, pp. 2357–2371.
- Zhang, Weihang and Ovidiu Serban (2025). "LLM-based Reranking and Validation of Knowledge Graph Completion". In: *Sixth International Workshop on Knowledge Graph Construction@ ESWC2025*.
- Zhao, Yihang et al. (2024). *Improving Ontology Requirements Engineering with OntoChat and Participatory Prompting*. arXiv: 2408.15256 [cs.HC]. URL: <https://arxiv.org/abs/2408.15256>.
- Zhou, Yuchen et al. (2024a). "Divergences between language models and human brains". In: *Advances in neural information processing systems* 37, pp. 137999–138031.
- Zhou, Yujia et al. (2024b). "Metacognitive retrieval-augmented large language models". In: *Proceedings of the ACM Web Conference 2024*, pp. 1453–1463.
- Zlatev, Jordan, Göran Jacobsson, and Liina Paju (2021). "Desiderata for metaphor theory, the Motivation & Sedimentation Model and motion-emotion metaphoremes". In: *Figurative Language—Intersubjectivity and Usage*. John Benjamins Publishing Company, pp. 41–74.

Acknowledgements

Looking back on these three years that have passed so quickly, I am reminded that this achievement would not have been possible without the people who have been there with me. Many have shaped this path, and it is not possible to properly thank everyone in these few pages.

I would like first, and most of all, to express my deepest gratitude to my supervisors, Prof. Aldo Gangemi and Prof. Andrea Giovanni Nuzzolese, for trusting me, patiently supporting me from before I even started my PhD, and for being an endless source of knowledge and interesting discussions.

I would also like to acknowledge the support of the ISUE and the Scuola Normale Superiore, who awarded me the Tonielli prize, supporting my research; in particular, Prof. Buffardi and Prof. Ruberti.

Furthermore, I would like to acknowledge the Collegio Superiore as a fundamental support for my research and personal life. In particular, my tutor, Prof. Antonella Grassi, who was always available and with whom I deeply enjoyed sharing interests outside academia.

Dialogue with professors has also been very important to me. In particular, I thank Prof. Presutti, Prof. Bianchini, Prof. Blomqvist, Prof. Hogan, Prof. Gherri, Prof. Lugli, and Prof. Wulff, for wholeheartedly welcoming me in many different ways. Thanks also to Giorgia Lodi, who with great passion guided me as I entered this field.

I would also like to thank my friends of the Collegio, and in particular those of my cohort, who truly felt like a community I could share laughs and struggles with.

Thanks also to the STLab team for the many trips, coffee breaks, and friendship. In particular, sharing the office with Andrea, Francesco, Sara, Silvia, and Davide has made my days better and taught me a lot, both professionally and as a person. I also thank my other PhD colleagues more broadly, and among them, Arianna, Luana, Stefano, Nicolas, Nicola, Glenn, and Jeffrey, as well as my international colleagues—in particular, for the many calls and jokes, I must specially mention Javad, Reham, and Hamed.

Thanks to my friends, and in particular to Camilla, who has been there for me over many years, from close and from afar.

Finally, for the words that do not need to be said and the words that are now shared, I would like to thank my family. To my partner Matteo, who with the utmost love and patience made sure I was listened to, fed, and rested throughout this journey, and the ones before and ahead. And to my uncles, aunts, and cousins, and to Bianca, who is arriving, for creating a sense of home wherever we gathered. To my mom, and to my sister Elena, whose unwavering presence and love has forever carried me through every achievement and loss, no matter how absurd, big or small.