ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

COMPUTER SCIENCE AND ENGINEERING

Ciclo 38

Settore Concorsuale: 09/H1 - SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

Settore Scientifico Disciplinare: ING-INF/05 - SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

SCARCE DATA AND COMPLEX TEXTS: THE ROLE OF LLMS IN AUTOMATIC TEXT CLASSIFICATION

Presentata da: Elena Palmieri

Coordinatore Dottorato

Paola Salomoni

Supervisore

Paolo Torroni

Co-Supervisore

Giuliano Franceschi

Esame finale anno 2026

Abstract

This thesis analyzes potential and limitations of LLMs for automatic text classification in domains where annotated data is scarce or unavailable. Although recent progress in NLP produced models trained on vast corpora and with great generalization capabilities, it remains unclear the extent to which these systems can complement or replace traditional supervised approaches, particularly in settings where domain knowledge, reasoning capabilities and context interpretation are essential.

This research was guided by three domain questions. Firstly, to what extent can LLMs handle classification tasks that require domain knowledge that is not completely represented in training data? Secondly, how do LLMs perform in settings where the domain knowledge is explicit and well-documented compared to those where reasoning on long and unstructured text is required? Finally, is it possible to use LLMs as substitutes to human annotators generating "silver" datasets that are suitable for the training of domain-specific lightweight models?

The results highlight that in domains where terminology is standard and knowledge is accessible, models such as GPT-4o and Gemini achieve high levels of accuracy. However, the performance drops when the task requires deeper reasoning or understanding of administrative or legal frameworks. This suggests that domain knowledge still represents a crucial constraint and that prompt engineering alone is not sufficient to compensate for its absence.

To address data scarcity, this thesis introduces an approach based on the creation of silver datasets, automatically annotated through carefully designed prompts, and their use for training lightweight classifiers such as Logistic Regression or Support Vector Machines. In the Emilia-Romagna case study, the models trained on the silver dataset outperformed zero-shot Llama and Gemini, offering a sustainable and cost-effective solution for public administrations.

Overall, this thesis presents empirical and methodological contributions to the role of LLMs in real-world classification, positioning them as assistive tools that enhance human expertise.

Contents

Chapter 1

Introduction

The role of text classification in the organization and analysis of textual information pervades a broad series of fields and applications [1] [2]. From personal to highly specialized institutional settings, the need to automatically assign and categorize text in meaningful ways becomes increasingly important for managing large collections of text [3]. Nevertheless, the growing application of data science methodologies has brought with it an apparent gap and contradiction between methodological advances and practical constraints, particularly in environments where texts are long, heterogeneous, and deeply embedded in specific legal, administrative, or scientific contexts [4]. This chapter introduces the broader problem addressed by the thesis, situating it at the intersection of natural language processing, domain-specific knowledge, and real-world institutional needs. It outlines the motivations that guided the research, the challenges that arise when annotated data is scarce or costly to obtain, and the reasons why recent developments in large language models invite both optimism and caution when applied to complex classification tasks [5].

1.1 Context and Motivation

Text classification has been one of the core challenges in natural language processing (NLP) for decades [1] [2]. Essentially, the task is that of assigning one or more pre-specified labels to an input text based on its content. On a first glance, this might look like and easy problem: for instance, determining whether an email is "spam" or "not spam" can be handled well with standard supervised approaches, as the two categories are diverse and the language features characteristic of each are relatively easy to tell apart from one another. The task becomes much more difficult when the categories are not so clear, the

terminology more nuanced, or the documents longer and more heterogeneous.

In the majority of real-world applications, text categorization is by no means straightforward. There are tasks that require single-label classification, where every document must be classified into one category only, as in the case of annotating administrative documents by subject or type. While others involve multi-label classification, where the same document can belong to multiple categories — a news item on a change in medical policy, say, can belong to "healthcare" and "public administration" simultaneously [6]. Each of these scenarios present their own difficulties, both in category definition and in how models must interpret texts in order to classify them correctly.

The categories themselves generally govern the level of difficulty. In general-purpose applications, such as sentiment analysis or spam detection, labels tend to be broad and strongly differentiated, so less complex models can still achieve a good performance. In contrast, in scientific writing, legal reasoning, or public policy, there is much deeper semantic understanding of content required. In these cases, labels typically convey specialized concepts, ingrained patterns of inference, or institutional distinctions that are not explicitly mentioned in the text. A model that has only learned to recognize surface-level keywords will never be able to discover the underlying meaning or contextual relevance required to make accurate predictions [7]. This shift from lexical recognition to contextual and conceptual representation moves the problem from one of pattern matching to one of reasoning and knowledge representation.

Another persistent challenge is the availability of annotated data. It is often time- and resource-consuming to curate a high-quality dataset, which relies on human experts who are able to comprehend the text and assign labels consistently across documents [8]. Gold-standard datasets are hence typically small and are primarily suited for testing rather than training [9]. The lack of sufficiently large labeled corpora is a constraint for standard supervised learning, which is heavily dependent on annotated samples to achieve high accuracy and generalization.

In the recent years, the availability of Large Language Models (LLMs) has profoundly reshaped the way we approach this problem [5] [10]. GPT, Gemini, and Llama are only a few of the models that have demonstrated amazing capability to perform classification directly from natural-language prompts, at times without the need for task-specific training data. Their apparent ability to generalize across domains has held out promise that they might assist overcoming the challenges posed by the lack of labeled data. However, this potential does not come without limitations [11] [12]. LLMs are trained on large but general corpora, and their comprehension of specialized subject matter — such as national law, domain-specific technical terminology, or administrative classifications

— is far from perfect. Additionally, tasks involving structured reasoning, contextual inference, or domain knowledge may still remain outside of their capabilities.

Such issues put the overall motivation of this research into perspective. The tension between the general-purpose power of LLMs and requirements of domain-specific knowledge lies at the heart of this thesis. Through investigation of how LLMs operate across tasks of varying degrees of difficulty and inquiry into hybrid strategies combining human expertise with automated annotation, this work seeks to dispel the mystery surrounding LLMs' contribution to text classification — not as universal solutions, but as a further methodological strategy bridging the trilemma between accuracy, explainability, and efficiency.

1.2 Problem Statement

The introduction of LLMs has opened up new possibilities for text classification [5] [13]. Such models, having been trained on massive corpora, are in many instances capable of classifying texts without being explicitly trained [14]; provided with an appropriately designed prompt, an LLM can assign a document to one or more categories with surprising accuracy. This alleged ability to generalize across domains has positioned them as one potential solution to the ever-present problem of scarce annotated data.

However, this apparent versatility has important limitations. LLMs are not all-knowing: their training set, though vast, does not cover all fields to the same extent, making them familiar with very specialized topics only in part [11] [15]. In settings where the classification is based on a strong knowledge of domain-specific concepts—be they legal categories, policy configurations, or technical terminology—LLMs may be unable to provide sound outcomes [4]. Moreover, there are some tasks that require more than simple categorization. They need reasoning, the power to balance evidence throughout a lengthy and raucous report, or the interpretation of citations that are undefined within the report itself [16]. Such areas may reveal the constraints of models that were not specifically trained within the domain in question.

It is for these reasons, then, that LLMs' role in classification remains ambiguous. They perform wonderfully well under some conditions, but there is no guarantee that this will carry over to harder cases. Understanding where they shine, where they do not, and how they compare with more traditional approaches is therefore crucial. This tension between promise and limitation defines the central problem addressed in this thesis.

Originating from a concrete institutional setting, this doctoral project was funded by

the public administration and is closely tied to the needs of Regione Emilia-Romagna, which provided the main case study through its large collection of administrative acts. The decision to focus on this material is therefore not only of academic interest but also reflects the priorities of the funding body, which seeks practical solutions for managing its documentation. Working in this context also brings some clear constraints. Public administrations operate with limited budgets, and the extensive use of large language models is often not sustainable in the long term, both financially and computationally. For this reason, the thesis looks for a middle ground: methods that can benefit from LLMs when they add real value, but that also allow the development of lighter, more efficient, task-specialized models capable of being used in practice without excessive costs.

1.3 Research Questions

Against this background, the central aim of the thesis is to examine how far large language models can be relied upon for text classification in situations where data is scarce and the task itself requires more than superficial text matching. Several questions naturally follow from this aim.

A first question concerns the role of domain knowledge. LLMs are trained on general-purpose data, but many of the categories relevant to research or institutional practice are deeply tied to specific fields. It is therefore essential to ask: **to what extent can LLMs cope with tasks that require knowledge they may not have encountered during training?**

A second point relates to the conditions under which LLMs appear to do well. In some cases, such as the classification of scientific publications according to the Sustainable Development Goals, the knowledge required is relatively well documented and represented. Here, one might expect the models to perform strongly. In other cases, such as the classification of asylum requests, the task involves not only recognizing topical categories but also reasoning over long, unstructured narratives. **How do LLMs behave in these two very different scenarios, and what does this tell us about their strengths and limitations?**

The scarcity of annotated data raises a further question. Gold datasets are expensive to create and are often kept small. **Is it possible to use LLMs as a substitute for manual annotation, at least to generate larger "silver" datasets that can then be used to train lighter, domain-specific models?** And if so, **how do such models, trained indirectly through LLM-generated labels, compare with both the LLMs themselves and traditional**

supervised approaches trained on human-annotated data?

These questions are not only theoretical but also highly practical. Institutions and researchers increasingly face the need to classify large volumes of text without the resources to build extensive gold standards or pay expensive cloud services. Understanding what can realistically be expected from LLMs, and how they can best be integrated with more traditional methods, is therefore at the heart of the work carried out in this thesis.

1.4 Objectives and Contributions

The overall objective of this thesis is to investigate the potential and the limitations of large language models in text classification tasks where annotated data is scarce and domain knowledge plays a decisive role. Rather than treating LLMs as a final solution, the work explores how and when they can be positioned within a broader methodology that combines their strengths with the reliability of expert annotation and the efficiency of traditional classifiers.

The main contribution is the design and validation of a classification pipeline built around three elements:

1. a small gold dataset, created with minimal expert effort, used as a benchmark and for calibration;
2. a silver dataset, much larger in scale, generated automatically through LLM-based annotation;
3. lightweight supervised classifiers trained on the silver data and tested against the gold standard.

Through this pipeline, the thesis shows that in some settings it is possible to reduce dependence on costly annotation campaigns while still producing models that are accurate enough to be deployed in real-world settings. The approach is evaluated in several domains: scientific publications, administrative acts, asylum requests, and family case law. Each of these settings highlights different challenges, from multilingualism to reasoning over long documents, and together they provide a broad basis for assessing the adaptability of the method.

In addition to proposing the pipeline, the thesis contributes an empirical evaluation of LLMs in comparison with both traditional supervised models and hybrid approaches

based on silver data. This makes it possible to identify where LLMs can already serve as effective classifiers, where they fall short, and how they can be complemented by smaller, task-specific models.

Overall, the thesis contributes to debates regarding the applications of LLMs in natural language processing. It does so not just by testing their performance across challenging domains, but also by suggesting a cost-effective way of applying them to real-world classification tasks when there is incomplete annotated data or when the annotation is too costly to produce.

1.5 Thesis Structure

The thesis is organized into seven chapters. After this introduction, Chapter 2 provides the necessary background. It reviews the main approaches to text classification, the challenges posed by long and unstructured texts, and the different families of unsupervised methods, including similarity-based approaches, zero-shot classification, and knowledge distillation. Then concludes analyzing the role of LLMs and domain knowledge and reasoning in text classification.

Chapter 3 introduces the central case study: the automatic classification of administrative acts from Regione Emilia-Romagna. The chapter sets out the main case study, describes the obstacles posed by the lack of labeled data and the complexity of the documents, and reports the first unsupervised attempts at classification. The limitations of these initial methods motivate the turn to zero-shot classification.

Chapter 4 presents two controlled experiments with zero-shot methods in domains where knowledge is more accessible to us: the medical field and asylum requests in Italy. These experiments make it possible to assess how LLMs handle multilingual data and how they perform when tasks require more than simple topical classification. The chapter concludes with a comparison across experiments.

Chapter 5 introduces the proposed methodology. It explains the construction of small gold datasets and larger silver datasets, describes the training of lightweight classifiers, and reports results from experiments on scientific publications, where the task was to classify articles according to the UN Sustainable Development Goals.

Chapter 6 validates the pipeline on real-world legal texts. The first part revisits the Emilia-Romagna administrative acts, applying the methodology developed in Chapter 5. The second part extends the validation to Bulgarian family case law, showing how the pipeline adapts to a new legal domain and to a very different language.

Finally, Chapters 7 and 8 conclude the thesis. Chapter 7 draws together the main findings. It reflects on the research questions introduced in this chapter, highlights the contributions of the thesis, and outlines directions for future work. Chapter 8 draws the conclusion of the thesis, summarizing all the findings and focusing on what can be done in the future to improve the results obtained.

Chapter 2

Background

This chapter offers the theoretical and methodological foundation to discuss the work presented in this thesis. It discusses the concepts, techniques, and difficulties associated with automatic classification, focusing on the transition from the traditional supervised approaches to recent neural and large language model–based approaches. Furthermore, through this chapter, we situate classification tasks in the realm of natural language processing techniques and discuss how challenges associated with natural language processing, including availability, domain-specificity, and reasoning, affect these classification models. By reviewing the relevant literature and key methodological paradigms, this chapter establishes a common ground for understanding the experimental choices made in the subsequent chapters and clarifies how the proposed analyses relate to existing work in the field.

2.1 Text Classification

Text classification, also known as text categorization or document classification, is a fundamental NLP task that involves assigning one or more predefined category labels to a given piece of text based on its content [17]. In the past few decades, especially with the recent advancements in NLP and text mining, this topic has been widely studied and applied in numerous real-world applications [18] [19] [20]. Information systems [21] and search engines [22], for example, greatly benefit from text classification methods. Furthermore, text classification can also be used for information filtering, such as spam detection and recommender system [23], but also in domains such as public health [24] and human behavior analysis [25].

A typical text classification system often follows a pipeline consisting of several

phases: feature extraction, dimensionality reduction, classifier selection, prediction, and evaluation [17]. Text classification can be applied at different levels of scope: document level, paragraph level, sentence level, and sub-sentence level, each aiming to obtain relevant categories for the respective text segment. Raw text data, which is generally unstructured, must first be converted into a structured feature space for mathematical modeling. Common feature extraction techniques include Term Frequency-Inverse Document Frequency (TF-IDF) [26], Word2Vec [27] [28], and Global Vectors for Word Representation (GloVe) [29]. Dimensionality reduction, which is an optional step, can be applied to reduce computational time and complexity through techniques such as Principal Component Analysis (PCA) [30]. The selection of the classifier is a crucial step; the choice encompasses a wide range of algorithms, ranging from the more traditional methods like Rocchio classification [31] and ensemble-based techniques such as boosting [32] and bagging [33], to various machine learning methods including logistic regression [34], Naive Bayes [35], k-nearest Neighbor [36], Support Vector Machine (SVM) [37], decision trees [38], random forests [39].

During the recent years we have witnessed progress in NLP and availability of dense training data, as well as advancements in deep learning techniques, leading to significant success in supervised text classification [40]. However, the quest for suitable structures, architectures, and techniques for optimal text classification remains an open problem for researchers [17].

2.2 Challenges with Long and Unstructured Text

The exponential rise in the number of long and unstructured documents highlights a strong demand for NLP systems that are capable of automatically model complex text and extract relevant information [41] [42]. Long texts, such as academic articles, official reports, or meeting transcripts, can contain thousands or more characters, presenting unique challenges compared to short or normal texts [43]. The first challenge is the inherent length limitation of pre-trained language models (PLMs) based on Transformer architectures. For instance, models like BERT can only process input up to 512 tokens [44], and even bigger models like BART can only process up to 1024 tokens [45]. Preprocessing functions can be used to transform lengthy input into shorter sequences bypassing this problem. Common preprocessing techniques include truncation, content selection, and chunking [42]. However, the latter, consisting in splitting the documents into multiple non-overlapping parts, can lead to the context fragmentation problem, where the sentences are split into different chunks, resulting in incomplete contextual

understanding. This can be mitigated overlapping text segments, i.e. including the text at the end of one segment into the beginning of the next one [46].

Furthermore, long documents often exhibit complex discourse structures, including sections and paragraphs, which are crucial for the appropriate comprehension of the overall meaning. Leveraging this structural information can significantly boost model performance on downstream tasks. For example, scientific papers are divided in sections that typically cover a single topic, which makes it possible to treat each section as an individual segment [47]. Some documents may also present challenges that stem from complex layouts and noisy input resulting from Optical Character Recognition (OCR) systems, posing the further requirement of robust processing capabilities [41].

2.3 Unsupervised Text Classification

Traditional supervised text classification methods need a large amount of human-labeled data, which can be costly and difficult to obtain, especially in uncommon domains [48]. This scarcity of data has drawn attention to semi-supervised and unsupervised approaches, that aim to perform text classification without relying on vast amounts of annotated training data [49]. These approaches have the potential to significantly reduce annotation costs and allow the use of existing algorithms on unseen classes [50].

Within unsupervised text classification, two primary techniques have emerged: similarity-based approaches and zero-shot learning (ZSL) approaches [51].

2.3.1 Similarity-Based Approaches

Similarity-based approaches aim to classify text instances by measuring the similarities between text document representations and class description representations. They operate by generating semantic embeddings for both the input texts and the textual label descriptions. Semantic embeddings are vector representations of texts engineered to capture their meaning, which can then be utilized in various NLP downstream tasks [52] [53] [54]. Classification is achieved by matching the text representations to the label representations using similarity measures, such as cosine similarity [55] [56].

The concept behind similarity-based classification was previously investigated under the umbrella term "Dataless Classification" [57]. Early work utilized Explicit Semantic Analysis (ESA) [58] to embed text and label descriptions into a common semantic space, classifying the text based on the label with the highest matching score. Dataless

classification rests on the principle that semantic representations of labels hold equal relevance to learning semantic text representations. As text embeddings advanced, the term "Dataless Classification" waned in prevalence, giving way to the broader category of similarity-based approaches for unsupervised classification [51].

Various methods fall under the similarity-based category: (i) Traditional Embedding Methods: Earlier studies, such as those by Sappadla et al. [55], embedded text documents and textual label descriptions using Word2Vec and subsequently employed cosine similarity to predict instances of unseen classes. Other approaches incorporated Latent Semantic Analysis (LSA) similarities [59], sometimes enriching label descriptions with expert keywords [56]. (ii) Lbl2Vec [60]: This method is a recent, well-performing similarity-based approach that focuses on unsupervised text classification. Lbl2Vec works by jointly embedding word, document, and label representations, initially utilizing Doc2Vec [61]. The label vector for each class is generated by finding candidate document representations—those most similar via cosine similarity to the average of label keyword representations—and then averaging those candidate document representations. Classification assigns documents to the class corresponding to the highest cosine similarity between the document vector and the label vector. (iii) Transformer-Based Baselines: Recent research has proposed strong baselines for unsupervised text classification based on advanced embedding similarities derived from SimCSE [62] and SBERT (Sentence-BERT) [63], which outperform simpler baselines like Word2Vec and LSA.

2.3.2 Zero-Shot Learning

Zero-Shot Learning (ZSL) is a specific type of unsupervised text classification whose core idea is to generalize knowledge gained from a training task to classify instances belonging to unseen classes, without any new labeled data for the target classes [64]. Early forms of ZSL were explored under the paradigm of "Dataless Classification" [57]. They used ESA [58] to embed the text and label descriptions in a common semantic space and picked the label with the highest matching score for classification. This approach emphasizes that the representation of labels plays an equally crucial role as the representation learning of text [40]. Despite not needing a labeled training dataset, ZSL still requires labeled data for the seen label set and cannot be applied to cases where no labeled documents for any class is available [49]. Furthermore, traditional classifiers struggle to generalize to unseen classes because label names are often converted into mere indices, preventing the model from truly understanding the meaning of the labels [40] [51].

To overcome these limitations, several approaches have been proposed. One approach models text categorization as a binary classification problem of finding relatedness (yes/no) between sentences and categories, rather than assigning a single class. This allows models trained on one dataset to be deployed for text categorization on other datasets without retraining, as they learn the concept of relatedness that can be extended across different datasets. Techniques like creating category trees are used to bridge the gap between atomic SEO tags and broader concepts found in standard datasets like UCI News Aggregator and Tweet Classification, allowing the model to function across different granularity levels of concepts [50]. A more recent and powerful approach frames zero-shot text classification as a textual entailment problem [51]. This is inspired by how humans classify: by mentally constructing a hypothesis (e.g., "this text is about sports?") and determining if it's true given the text [40]. By converting class labels into natural language hypotheses, models can leverage knowledge from entailment datasets and better understand the problem and label meanings, facilitating generalization. The Task-Aware Representation of Sentences (TARS) method [65] imbues the notion of the task directly into the transformer model by factorizing arbitrary classification problems into a generic binary "True/False" classification. The input to the transformer includes both the text and the class label, allowing the model to learn the semantics of new classes purely from the class label itself, even in zero-shot scenarios. The Label-Name-Only Text Classification (LOTClass) approach [49] utilizes pre-trained neural language models as both general linguistic knowledge sources for category understanding and as representation learning models for document classification, operating with only the label name of each class (often just 1-3 words) and unlabeled data. LOTClass constructs a category vocabulary for each class, collects category-indicative words from the unlabeled corpus, and generalizes through self-training. It has demonstrated comparable performance to strong semi-supervised and supervised models without requiring any labeled documents. SimPTC [66] is a framework that improves zero-shot classification by clustering texts in the embedding spaces of PLMs. SimPTC models texts in each class with a Gaussian distribution and fits text embeddings with a Bayesian Gaussian Mixture Model, using class names to initialize cluster positions and shapes. It expands class names using external knowledge bases (like ConceptNet Numberbatch) and constructs class anchor sentences using natural language templates, then encodes them into the PLM embedding space. SimPTC has shown superior or comparable performance on various datasets, particularly on unbalanced datasets. However, it still requires an unlabeled corpus or knowledge base to extract topic-related words. Finally Prompt-based methods reformulate classification tasks into a cloze task using prompts, where a template is filled with class names. This prompting can alter the input data distribution, encouraging a better separation of

data from different classes in the resultant feature space, thereby enhancing zero-shot capabilities. However, these methods can depend heavily on human engineering or external knowledge [67].

In recent years, zero-shot classification has gained renewed attention with the advent of large language models. Unlike traditional zero-shot methods that typically lean on semantic embeddings or other auxiliary information, LLMs make it possible to do zero-shot classification simply by using natural language prompts. In this new framework, the problem statement is framed in the form of an instruction in natural language: given the description of the classes to be predicted, the model is asked to select the best class for the given document. The essence of this new framework has been explored in [5], where it was proven that it is sufficient to have sufficiently large language models that achieve most general-purpose NLP tasks using only text prompts in either zero-shot or few-shot regimes. Subsequent work has highlighted the critical role of prompt design itself, with analyses showing that prompt complexity — such as the inclusion of detailed task and label descriptions — can significantly influence classification performance in zero-shot settings. For example, [68] evaluated zero-shot performance of instruction-tuned models across multiple social science classification tasks and found that different prompting strategies can cause variations in accuracy and F1 scores exceeding 10%, underscoring the sensitivity of LLM behavior to how tasks are described to the model.

More recent studies have also shown that LLMs can implicitly perform forms of reasoning when prompted appropriately, which is particularly relevant for complex classification tasks involving long or ambiguous texts [10]. These findings have positioned LLM-based zero-shot classification as an attractive option in low-resource scenarios, where annotated data is unavailable or costly to obtain. At the same time, the reliance on implicit knowledge and prompt sensitivity makes this approach inherently unstable: performance may vary significantly across domains, especially when classification requires specialized knowledge or the interpretation of context that is not explicitly encoded in the input. This tension between flexibility and reliability is central to the experiments presented in this thesis.

2.3.3 Knowledge Distillation

Knowledge Distillation (KD) has emerged as a representative and effective technique for model compression and acceleration [69]. While deep neural networks (DNNs) have achieved significant success in both academia and industry, particularly for computer vision tasks [70] and applications such as NLP [44] and reinforcement learning [71] [72]

[73], this success is largely attributed to their scalability in encoding large-scale data and managing billions of model parameters.

However, the huge computational complexity and massive storage requirements of large deep models make their deployment challenging in real-time applications, especially on devices with limited resources, such as mobile phones and embedded systems. To address this challenge, a variety of model compression and acceleration techniques have been developed [74] [75]. These categories include parameter pruning and sharing, low-rank factorization, and transferred compact convolutional filters. KD falls into these categories, aiming to distill knowledge from a larger deep neural network into a small network.

The basic idea of model compression, which consists of transferring information from a large model or an ensemble of models to train a small model without a significant drop in accuracy, was first proposed by Bucilua et al. [74]. Subsequently, the learning of a small model from a large model was formally popularized as knowledge distillation Hinton et al. [76].

Generally, KD operates within a generic teacher–student framework, where a small student model is supervised by a large teacher model [77] [78]. The main objective is for the student model to mimic the teacher model in order to obtain a competitive or even superior performance. The key problem in KD is how to effectively transfer knowledge from the large teacher model to the small student model. A knowledge distillation system is composed of three key components: knowledge, the distillation algorithm, and the teacher–student architecture.

Despite the notable practical successes, the theoretical and empirical understanding of KD is still developing [79]. Recent studies have sought to provide theoretical justification for KD and empirically analyze its efficacy [80][81].

In traditional offline distillation, the teacher model is typically pre-trained on a dataset, and that original dataset (or a transfer dataset) is used during the distillation phase [76]. However, various advanced KD schemes utilize synthetically generated data, often addressing scenarios where the original training data is unavailable or inaccessible due to security, privacy, legality, or confidentiality concerns [82] [83] [84].

Data-Free Distillation (DFKD) is explicitly designed to handle situations where the training data is not present ("data free"). In these methods, the student network is trained using newly or synthetically generated data. The key mechanism is that the pre-trained teacher model provides the necessary information—the "Knowledge for Generating Data"—which is then used to create this "Synthesis Data" for the student's training. Methods of data generation within DFKD include: (i) GAN-based Generation:

The transfer data can be generated by a Generative Adversarial Network (GAN) [82] [85] [86] [87]. (ii) Feature Reconstruction: Some techniques reconstruct the transfer data by leveraging the teacher network's layer activations or layer spectral activations [83]. (iii) Zero-Shot KD: This approach produces the transfer data by mathematically modeling the softmax space using the parameters of the teacher network, requiring no existing data [84].(iv) DeepInversion: This method uses knowledge distillation to generate synthesized images specifically for data-free knowledge transfer [88].

Finally, there is a newer type of distillation called "data distillation", which is similar to data-free distillation, where new annotations of unlabeled data are generated from the teacher model and then used to train the student model[89] [90] [91].

2.4 Large Language Models for Text Classification

Over the past few years, LLMs have become increasingly important in text classification tasks, especially when there is not enough annotated data. Their ability to generalize across domains, follow instructions (prompts), and use knowledge implicitly acquired during pre-training makes them promising tools. But as with all powerful tools, they come with both advantages and significant limitations, especially when domain specificity, document complexity, or reasoning are involved.

What makes LLMs appealing are first of all their zero- and few-shot capabilities; LLMs are often able to perform classification without fine-tuning, simply by being given instructions (prompts) and definitions of the categories. For instance, [92] shows that in several datasets, models like GPT can achieve strong zero-shot performance, especially when label definitions are clear and the prompts well-designed. Another strength of LLMs is their flexibility. Because they are not anchored to fixed feature sets or manual engineering, they can adapt to different tasks and domains simply by changing the prompt. Prompt engineering allows models to balance performance and computational costs [93], and with only a few examples models can generalize or produce synthetic data for new tasks without retraining [94]. Finally, they reduce the need for large annotated training sets. Several studies have shown that LLM-generated silver or synthetic datasets, when combined with a small gold set, can enable supervised models to perform well even in low-resource settings. For example, [95] shows that mixing gold and silver (or bronze) data boosts performance particularly in rare classes. Lighter models trained on LLM-generated labels can approach the accuracy of human-annotated baselines, especially when prompt design and filtering are done carefully [96].

While LLMs have many strengths, empirical studies reveal that their performance is not uniformly strong across all settings. (i) Domain knowledge gaps. Even large models, pretrained on massive corpora, do not always have deep understanding of domain-specific legal or administrative or scientific content. For example, in [97], researchers found that fine-tuned models with domain-specific training outperform non-fine-tuned LLMs for legal semantic role labeling. Similarly, in the biomedical domain, standard prompting often outperforms more elaborate reasoning methods (chain-of-thought, retrieval augmentation) only when supplemented by external or domain-specific knowledge [98]. (ii) Prompt sensitivity and design matters. Small changes in wording, format, whether you include definitions, or how you frame examples can shift performance significantly. Prompt complexity (e.g. label definitions, formatting) can affect accuracy by more than 10% across tasks [68]. Also, in [99], authors show that different prompt styles (manual, optimized, etc.) yield varying performance depending on the task. (iii) Performance on long or noisy texts. Large language models often encounter significant limitations when dealing with long or noisy texts—that is, documents that contain repeated sections, unexplained legal references, or boilerplate language that dilutes the relevant information. For example, [100] shows that performance can remain nearly unchanged even when models process only the most relevant portions of a clinical document rather than its full content. Similarly, [101] demonstrates that many models struggle when document length varies significantly, leading to unstable accuracy and inconsistent results across different text segments. These findings confirm that, for lengthy or heterogeneous texts, identifying and isolating the informative sections can be as crucial as the model's intrinsic capacity. (iv) Reasoning and subtle distinctions. Some tasks require reasoning beyond topic detection: distinguishing between types of legal interpretation, causal relations, or normative reasoning. In [102], for instance, zero-shot methods showed limitations when asked to classify statements involving causal or conditional relations. Such issues are central when the classes are not just topics but require interpreting norms, actions, decisions, or legal consequences. (v) Cost, computational resources, and scalability. The limitations of LLMs in long-text classification are not merely conceptual, they also have practical consequences in terms of computational cost. Processing large inputs implies higher token counts, longer inference times, and substantially greater financial cost. In [103], the authors show that dynamically managing token budgets in reasoning prompts can reduce token usage by more than 60–70%, while keeping accuracy losses below 5%. Other studies, such as [104], point out that latency, token limits, and per-call expenses can make many prompting strategies impractical for large-scale applications. This is particularly relevant in public or institutional contexts—like those addressed in this thesis—where both time

and budget are limited. (vi) Class imbalance, rare categories, and ambiguity. Finally, another persistent challenge in LLM-based classification is class imbalance. When certain categories are rare or partially overlapping, models tend to overpredict the dominant classes, leading to low recall for minority categories. In [105], for example, the authors observed that macro-F1 scores were considerably lower than micro-F1, revealing the model's systematic difficulty in identifying rare categories. Likewise, [106] found that augmenting data with LLM-generated synthetic samples improved performance on minority classes, yet did not entirely remove the bias toward more frequent ones. These findings highlight the structural limits of zero-shot approaches in domains where the natural class distribution is heavily skewed or ambiguous.

2.5 The Role of Domain Knowledge and Reasoning

In many real-world classification tasks, the mere ability to parse and compare text is insufficient. What often makes the difference is domain-specific knowledge: the rules, terminology, context, and nuances that belong to a field such as law, public administration, or policy. In this section we explore how domain knowledge and reasoning interact with LLMs' capabilities, and why they are particularly relevant to settings like administrative acts or legal documents.

Domain knowledge as a boundary for LLMs

At their core, large language models are trained on very broad corpora covering many domains, but that does not guarantee depth in any particular one. Without explicit domain adaptation, their knowledge of legal statutes, regulatory practices, or administrative procedures may be partial or superficial. Some recent work has studied this gap. For instance, [107] presents various strategies for integrating domain expertise (e.g., static embeddings, retrieval augmentation) to enhance performance in specialized tasks. Another relevant study, [108], surveys techniques for adapting LLMs to specific domains, emphasizing that specialization is both a need and a challenge.

In a more applied sense, [109] shows that combining human-annotated domain features with LLM-derived representations can improve performance, compared to pure LLM outputs. This suggests that even powerful models benefit from explicit domain signals, especially in contexts where subtleties matter.

Reasoning over long texts and implicit structures

Beyond raw knowledge, classification in legal or administrative domains often requires reasoning: linking parts of a text, interpreting indirect references, applying normative logic, or dismissing irrelevant boilerplate. LLMs do reasonably well on simple classification tasks, but their reasoning over long documents or multi-step inference remains fragile. The survey [110] highlights that LLMs struggle with multi-step reasoning, especially when supervision is unavailable. Another strand of research shows how adding seemingly innocuous clauses can degrade model reasoning performance dramatically [111]. Also, analyses of real models suggest their "reasoning" often mimics pattern matching rather than symbolic logic, which becomes a limitation when the classification depends on applying norms or legal logic.

Furthermore, in domain-rich settings, texts can refer to statutes, case law, or administrative regulations implicitly, expecting the reader (or model) to "know" the reference. Without external grounding or context, the model may misinterpret or ignore such references. Some works propose hybrid strategies: for instance, [112] introduces a paradigm where LLMs and knowledge graphs mutually reinforce each other to strengthen domain reasoning.

When reasoning complements domain knowledge in classification

In domains such as law or public administration, the boundary between topic detection and reasoned classification is typically nuanced. A document may appear to belong to one class by virtue of its surface terminology, but to another when construed in terms of domain-specific conventions. The distinction hinges not only on pattern recognition but also on familiarity with contextual and procedural logic. This has been observed in several legal text classification experiments, which show that performance drops significantly when models are asked to make decisions involving implicit reasoning or cross-references between statutes [4]. In such cases, accurate classification involves more than lexical overlap—it requires an interpretation that is consistent with domain reasoning.

Chapter 3

Case Study: Administrative Acts and Limitations of Unsupervised Methods

Over the past few years, public administrations have started digitalizing their internal documents, producing a large volume of textual data in the form of reports, communication, administrative and legal acts. While transitioning to the digital format creates benefits, such as improved transparency and efficiency, it also creates a few challenges. The most pressing one being the structured classification of the documents, in order to grant meaningful retrievals. This classification has always been done by domain experts, but the large volume of data is making their job more and more challenging. Automatic document classification, the task of assigning a label to a piece of text based on its content, is a well-established research area in NLP, however, most of the existing works in this field focus on contexts where the categories are known, and fully labeled dataset are readily available.

In contrast, public administration documents, and administrative acts in particular, present a set of difficulties. Texts are often very long, and written in technical, legal or bureaucratic language that varies not only among institutions, but also based on the person who is writing the document within the same institution. Moreover, the categories used to classify them are not always standardized, which means that every public administration has it's own internal classification that may differ from the other ones, and that can evolve through time as new types of acts appear. Despite these complexities, the automatic classification of administrative acts is a crucial step to improve how public administrations manage their documents and how it can be accessed at a later time.

3.1 Research Objective

The aim of this research is to develop a method for automatically classifying administrative acts according to a new national taxonomy that has recently been introduced. This new classification system replaces a variety of regional schemes that were previously used by individual public administrations. While the introduction of a unified standard is a positive step toward harmonizing practices across institutions, it has also created a practical problem: at the moment, there are no existing datasets where documents have already been labeled according to this new taxonomy. The regional systems used in the past were too different to be mapped directly, and re-annotating existing archives would require a significant investment of time and expert knowledge.

Because of this, supervised machine learning methods are not a viable option—we simply don't have the labeled data needed to train them. The project therefore explores an alternative approach, based on unsupervised methods, specifically zero-shot classification using LLMs. The idea is to test whether these models, which have been trained on vast amounts of general text, can assign the correct category to a document using only the label names and their descriptions, without ever having seen examples. The broader goal is to understand whether this strategy can be used to support public administrations in implementing the new classification system, even in the absence of a labeled dataset.

3.2 Challenge: Lack of Labeled Data

One of the main challenges of this project is the complete absence of labeled data. As already mentioned, this is caused by the introduction of a national taxonomy, described in the *Titolario delle Giunte Regionali*[1], that substitutes the old regional classification and was never used before. Unfortunately, the old classes cannot be mapped into the new ones, so all the old documents cannot be reused as training material. Manually reclassifying the existing documents would require domain experts a significant amount of time, as they are still not familiar with the new taxonomy. Also, there are 19 different classes in the new Titolario, so the amount of documents needed to create a dataset would require an effort that is not feasible for the domain experts at Regione Emilia-Romagna.

The lack of annotated data rules out the traditional supervised machine learning methods, which rely on thousands of labeled examples for each class, and also the semi-supervised

[1] `https://www.agid.gov.it/sites/default/files/repository_files/documenti_indirizzo/titolario_giunte_regionali.pdf`

or weakly supervised methods, which still need a few examples to learn from. In short, we have a new and stable classification schema, but we have no training data to work with. This scenario, while not uncommon in institutional contexts, requires unsupervised techniques that can operate without relying on annotated corpora and only need a small set of annotated examples that allow to validate the results.

In order to obtain this small test set, we asked the domain experts to label a few documents. This process highlighted two other challenges: the structure of the documents can be very confusing due to the presence of numerous citations to laws and administrative decrees which are only referred to using their numerical identifier, and the distribution of documents in the different classes is highly imbalanced. After a long and time-consuming labeling process, we ended up with a total of 264 tagged documents, which are distributed as in Table 3.1.

Class Name	**N. of documents**
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	19
3 - Risorse Umane	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	15
5 - Sistema Informativo	9
6 - Programmazione, Coordinamento e Controllo	41
7 - Agricoltura e Allevamento	35
8 - Artigianato	3
9 - Commercio	5
10 - Industria e Attività Estrattive	6
11 - Turismo e Strutture Ricettive	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	6
13 - Infrastrutture e Trasporti	8
14 - Tutela dell'Ambiente	15
15 Sanità, Igiene e Veterinaria	28
16 - Politiche Sociali	7
17 Istruzione, Formazione e Lavoro	11
18 - Beni e Attività Culturali	15
19 - Attività Sportive e Ricreative	6

Table 3.1: Class distribution of tagged documents.

Upon a further analysis, data that belongs to some classes, such as class 8, appeared to

be almost impossible to find. This could be due to the wide variety of topics covered by the classes, some of which can be popular in some Italian regions and rare in others. Even after more attempts to balance the dataset, it became apparent that it is an almost impossible task, as we could not manage to obtain enough data for some of the classes. This means that it is not possible to use it to train a model, even using few-shot learning. It is nevertheless useful for validation.

3.3 Challenge: Long and Unstructured Text

As already mentioned, the structure of the administrative acts proved to be challenging even for domain experts. The documents are at least 6 pages long, with a mean of around 3000 words per document, and there are no specific guidelines to write them, so, based on the author of the document, the legal premises could either be at the beginning or in the middle and take up more or less space. This lack of a coherent and stable structure limits the applicability of segmentation techniques based on textual patterns or formal markers. The only fixed part is the presence of the keyword *"Determina"* leading up to the decision of the judge. Initially, we thought that that part of text would be the most informative one, but unfortunately that is not the case, as for the most part it states that money will or will not go to administrative bodies or private companies, but it does not state what the money is for. This makes it the only part of text that we can remove during the preprocessing step.

Another problem stems from the fact that the relevant part, i.e., the part before the keyword *"Determina"*, is written in bureaucratic language, which makes it even harder to comprehend for a non-expert and for a model. The text contains several citations to laws and legislative decrees with no explicit reference to their content. These citations, used to justify administrative decisions, take up relevant portions of text, but they do not contain relevant information for the correct classification of the document, resulting in the relevant information being hidden inside text that looks meaningless to anyone who is not familiar with this topic.

3.4 First Unsupervised Classification Approach: Clustering

Given the absence of labeled data, the first experiment conducted consisted of an unsupervised classification attempt through clustering. The objective of this experiment was to identify 19 clusters, corresponding to the 19 categories defined in the Titolario. This algorithm needs neither labeled data, nor the names of the classes, as it only groups the documents in a number of sets that is previously defined.

Since clustering algorithms cannot operate directly on raw text, the documents were transformed into numerical representations using the Term Frequency–Inverse Document Frequency (TF–IDF) vectorizer. This technique quantifies the importance of each word within a document relative to the entire corpus. To improve the quality of the representation, common stopwords were removed. As the scikit-learn implementation of TF–IDF provides only an English stopword list, an external Italian stopword list was retrieved from the Ranks website[2]

The clustering itself was performed using the k-means algorithm, which iteratively optimizes the position of a predefined number of centroids until convergence is reached. The outcomes of the clustering procedure were visualized through two complementary approaches. First, the high-dimensional vectors were projected into two dimensions using Principal Component Analysis (PCA) and plotted with the matplotlib library[3]

To check what the appropriate number of clusters would be, the Elbow method was applied. This technique evaluates the Within-Cluster Sum of Squares (WCSS) across a predefined range of possible cluster counts. When the WCSS values are plotted against the corresponding values of K, the resulting curve typically displays a characteristic "elbow" shape: the WCSS decreases steeply at first, then levels off and becomes almost parallel to the x-axis. The point at which this transition occurs indicates the optimal number of clusters, as adding further clusters beyond this point yields only marginal improvements in compactness. It was therefore of interest to examine whether the number of clusters suggested by the Elbow method could provide a more effective classification than the one defined by the Titolario. But, as visible in Figure 3.1, the optimal number of clusters is not really defined, as the WCSS keeps decreasing even with 30 possible clusters.

[2] `https://www.ranks.nl/stopwords/italian`

[3] `https://matplotlib.org/`

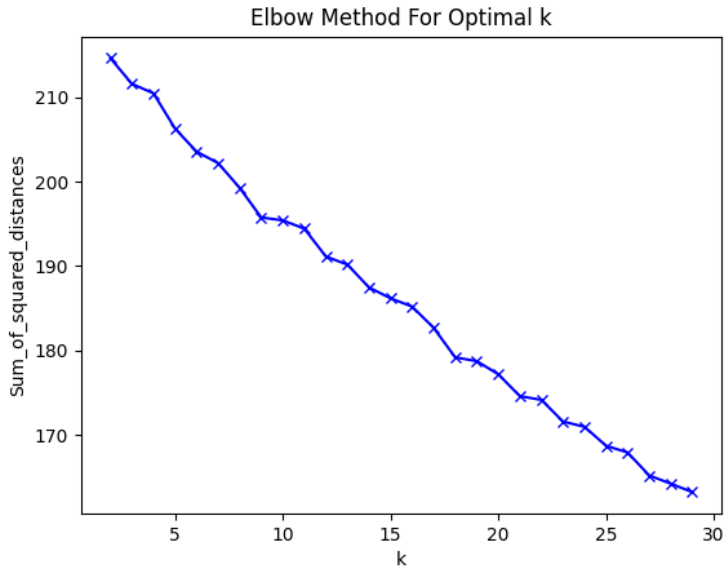


Figure 3.1: Results of the Elbow method.

This result already indicates that the clustering won't be successful, an hypothesis confirmed by the results of the k-means algorithm with 19 clusters visible in Figure 3.2. Not only the clusters are not well defined, but also, they contain documents that belong to many different classes, meaning that almost all of them are overlapping in these results.

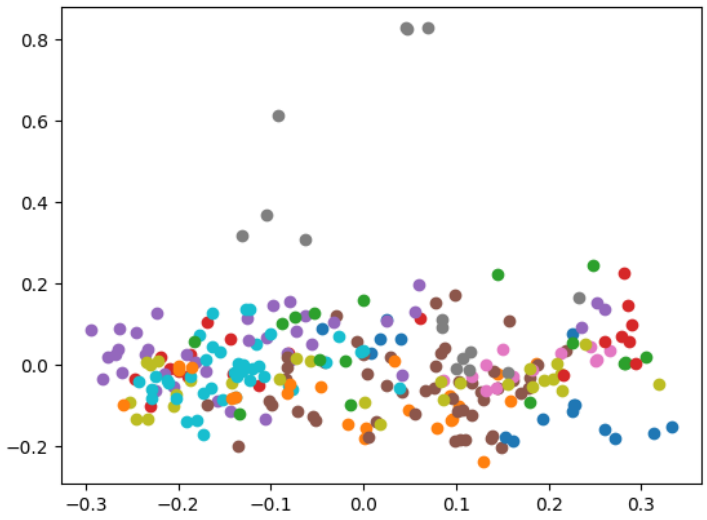


Figure 3.2: Clustering results using the k-means method.

3.4.1 Lbl2Vec

Lbl2Vec [60] is an algorithm for unsupervised document classification and retrieval that automatically generates jointly embedded label, document and word vectors, and returns documents of categories modeled by manually predefined keywords.

The first step of the algorithm is the definition of keywords that describe each class; as they need to be defined manually, the process requires a certain degree of domain knowledge. Ideally, the all the keywords of a class should be descriptive and similar to each other, but different form the ones that describe other classes.

Describing the classes can be quite beneficial for the classification experiment: firstly, it will be possible to label the clusters, whereas in the classical clustering methods there is no way of understanding which class the cluster refers to, as they only group data without being given any information; also, describing the classes can help identify the areas of interest for our classification, instead of classifying blindly.

Class Name	**Keywords**
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	Giunta, Ente locale, Istituzione, Contenzioso
2 - Organizzazione, Patrimonio e Risorse Strumentali	Organizzazione, Servizio, Patrimonio, Sicurezza
3 - Risorse Umane	Personale, Dipendente, Assunzione, Sindacato
4 - Risorse Finanziare, Gestione Contabile e Fiscale	Fiscale, Contabile, Finanza, Bilancio
5 - Sistema Informativo	Informatica, Informatizzazione, Statistica
6 - Programmazione, Coordinamento e Controllo	Programmazione, Coordinamento
7 - Agricoltura e Allevamento	Agricoltura, Allevamento, Zoorecnica, Foresta
8 - Artigianato	Artigianato, Arte, Disegno, Scultura
9 - Commercio	Commercio, Fiera, Distribuzione, Mercato
10 - Industria e Attività Estrattive	Industria, Estrazione, Miniera
11 - Turismo e Strutture Ricettive	Turismo, Viaggio, Albergo
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	Urbanistica, Edilizia, Cartografia, Appalto
13 - Infrastrutture e Trasporti	Trasporto, Mobilità, Ferrovia, Aeroporto
14 - Tutela dell'Ambiente	Ambiente, Natura, Inquinamento, Rifiuti
15 Sanità, Igiene e Veterinaria	Sanità, Igiene, Veterinario, Farmacia
16 - Politiche Sociali	Sociale, Bambino, Anziano, Disabile
17 Istruzione, Formazione e Lavoro	Istruzione, Lavoro, Formazione
18 - Beni e Attività Culturali	Cultura, Museo, Biblioteca, Spettacolo
19 - Attività Sportive e Ricreative	Sport, Caccia, Pesca, Ricreativo

Table 3.2: Examples of keywords for the classes in the Titolario.

After the definition of the keywords (the ones that we chose are visible in Table 3.2, the algorithm creates jointly embedded documents and word vectors. These vectors allow to represent the text in a multi-dimensional space, so that similar words or text documents will have similar vectors and will therefore be located close to each other and to the most distinguishing words in the vector space.

The algorithm then computes the cosine similarity between documents and the manually defined keywords for each category and uses the Local Outlier Factor algorithm to identify and remove the outliers.

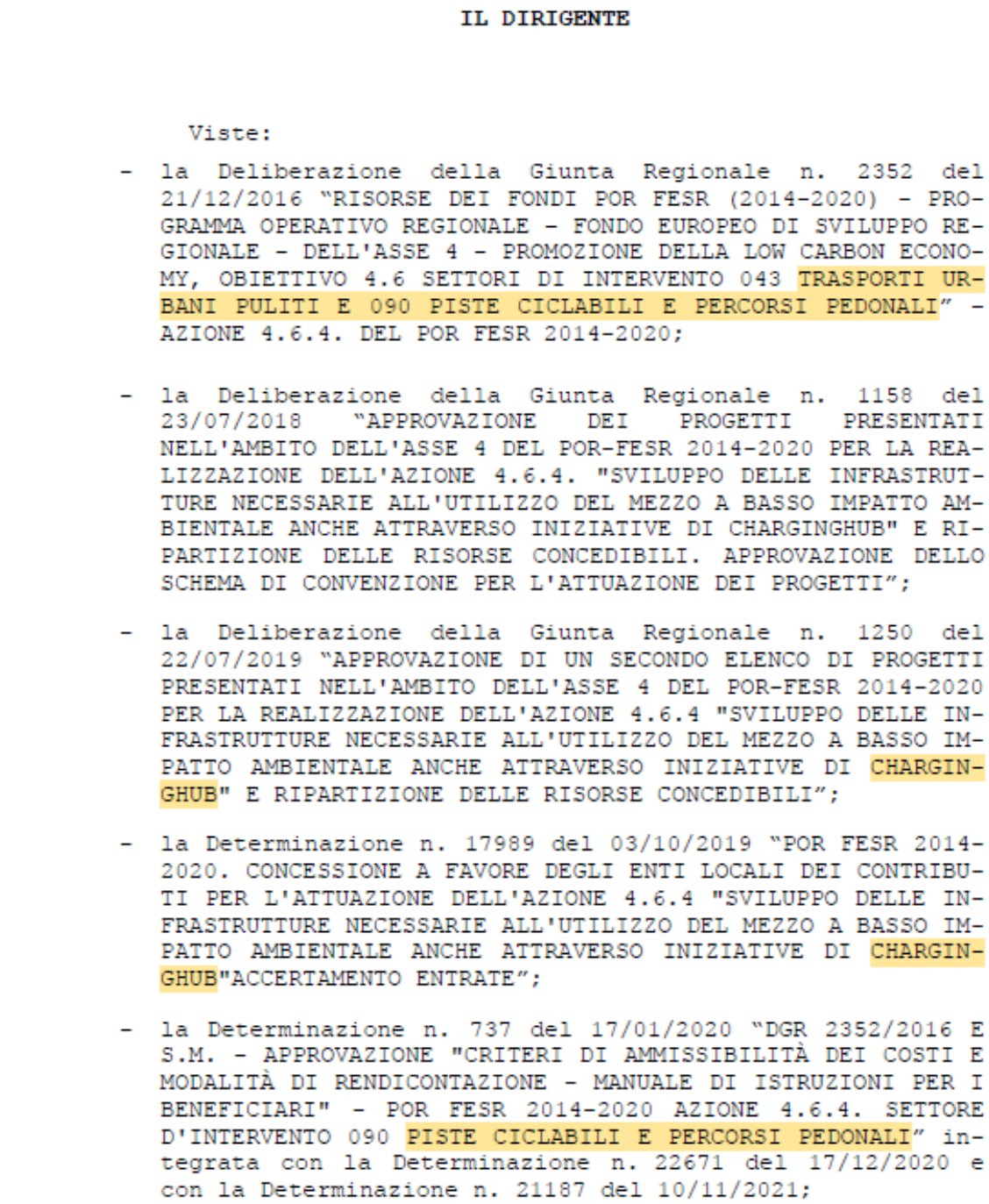
The final step of the algorithm is the computation of the centroid as the label vector of each category. It is computed as the document's centroid rather than the keyword's one because experiments showed that it is more difficult to classify documents based on their similarity to keywords only, even if they share the same vector space. The final classification of the document is the one with the highest label vector - document vector similarity.

Unfortunately, even with using this technique, the results are not good at all. Only 41 out of the 264 documents are correctly classified, meaning that these methods are not a viable solution for our problem.

3.5 Observed Limitations

The unsuccessful outcome of the clustering experiment can be explained by several factors. A first difficulty in the classification lies in the length and structure of the documents. With an average of nearly 3,000 words each, the texts are already complex to process. This challenge is compounded by the fact that many words recur frequently because the documents often cite laws and decrees. As a result, unsupervised methods, which rely on textual similarity, struggle to discriminate between them. For example, in the extract taken from one of the documents in Figure 3.3, it is evident that the bulk of the document consists of lengthy passages with little informative value, while the truly relevant content tends to be "hidden" within them, making it hard to identify. Furthermore, the irrelevant information is full of terms that are similar across paragraphs and among the documents belonging to every class.

Another complication arises form the overlap between categories. For example, class "7 - Agricoltura e Allevamento" (*7 - Agriculture and livestock breeding*) can create a conflict with class "19 - Attività Sportive e Ricreative" (*19 - Sports and Recreational Activities*). The most frequent case is the subject of hunting: a document could both be placed in class 7 because it might discuss the effects of wild animals on farmland, or in class 19 because



IL DIRIGENTE

Viste:

- la Deliberazione della Giunta Regionale n. 2352 del 21/12/2016 "RISORSE DEI FONDI POR FESR (2014-2020) - PROGRAMMA OPERATIVO REGIONALE - FONDO EUROPEO DI SVILUPPO REGIONALE - DELL'ASSE 4 - PROMOZIONE DELLA LOW CARBON ECONOMY, OBIETTIVO 4.6 SETTORI DI INTERVENTO 043 TRASPORTI URBANI PULITI E 090 PISTE CICLABILI E PERCORSI PEDONALI" - AZIONE 4.6.4. DEL POR FESR 2014-2020;

- la Deliberazione della Giunta Regionale n. 1158 del 23/07/2018 "APPROVAZIONE DEI PROGETTI PRESENTATI NELL'AMBITO DELL'ASSE 4 DEL POR-FESR 2014-2020 PER LA REALIZZAZIONE DELL'AZIONE 4.6.4. "SVILUPPO DELLE INFRASTRUTTURE NECESSARIE ALL'UTILIZZO DEL MEZZO A BASSO IMPATTO AMBIENTALE ANCHE ATTRAVERSO INIZIATIVE DI CHARGINGHUB" E RIPARTIZIONE DELLE RISORSE CONCEDIBILI. APPROVAZIONE DELLO SCHEMA DI CONVENZIONE PER L'ATTUAZIONE DEI PROGETTI";

- la Deliberazione della Giunta Regionale n. 1250 del 22/07/2019 "APPROVAZIONE DI UN SECONDO ELENCO DI PROGETTI PRESENTATI NELL'AMBITO DELL'ASSE 4 DEL POR-FESR 2014-2020 PER LA REALIZZAZIONE DELL'AZIONE 4.6.4 "SVILUPPO DELLE INFRASTRUTTURE NECESSARIE ALL'UTILIZZO DEL MEZZO A BASSO IMPATTO AMBIENTALE ANCHE ATTRAVERSO INIZIATIVE DI CHARGINGHUB" E RIPARTIZIONE DELLE RISORSE CONCEDIBILI";

- la Determinazione n. 17989 del 03/10/2019 "POR FESR 2014-2020. CONCESSIONE A FAVORE DEGLI ENTI LOCALI DEI CONTRIBUTI PER L'ATTUAZIONE DELL'AZIONE 4.6.4 "SVILUPPO DELLE INFRASTRUTTURE NECESSARIE ALL'UTILIZZO DEL MEZZO A BASSO IMPATTO AMBIENTALE ANCHE ATTRAVERSO INIZIATIVE DI CHARGINGHUB"ACCERTAMENTO ENTRATE";

- la Determinazione n. 737 del 17/01/2020 "DGR 2352/2016 E S.M. - APPROVAZIONE "CRITERI DI AMMISSIBILITÀ DEI COSTI E MODALITÀ DI RENDICONTAZIONE - MANUALE DI ISTRUZIONI PER I BENEFICIARI" - POR FESR 2014-2020 AZIONE 4.6.4. SETTORE D'INTERVENTO 090 PISTE CICLABILI E PERCORSI PEDONALI" integrata con la Determinazione n. 22671 del 17/12/2020 e con la Determinazione n. 21187 del 10/11/2021;

Figure 3.3: Extract of a real document. The information relevant for classification is highlighted in yellow.

hunting is considered a sport.

Other overlaps stem from the historical context; today, topics like the COVID emergency, sustainable mobility, gender violence and discrimination, and all the activities related to the Italian PNRR (Piano Nazionale Ripresa e Resilienza), would be explicitly mentioned in a standardized classification, but the one that we are working with, even though it has never been used before, dates back to 2010-2011 and is affected by the mindset of that time. For example, acts that mention the COVID emergency could either be labeled as "15 - Sanità, Igiene e Veterinaria" (*15 - Healthcare, Hygiene and Veterinary*) as COVID was a healthcare emergency, but also as "14 - Tutela dell'Ambiente" (*14 - Environmental Protection*), a class that has a specific point about Civil Protection.

These gray areas, combined with the effort that has to be made to find the relevant information inside the text, create uncertainties not only for classification algorithms, but often for human experts too, who, given that it is not possible to give multiple labels to a document, must reach an agreement on how to classify these borderline cases.

3.6 From Unsupervised Clustering to Zero-Shot Classification

The shortcomings observed in the clustering experiments point to a central problem: when faced with long and heterogeneous administrative texts, unsupervised methods only go so far. Even approaches such as lbl2vec, which already perform better than standard clustering techniques, remain limited because they cannot directly integrate domain knowledge or make use of natural language descriptions of the categories. Lbl2vec managed to classify a small number of documents correctly, but its representation space is still shaped by embeddings alone, without the possibility of drawing on the expressiveness of category definitions formulated in plain language.

This is precisely where zero-shot classification becomes relevant. Unlike clustering, zero-shot methods allow the researcher to supply a textual description of the classes, which the model then uses during classification. The idea is that the link between documents and categories does not have to be reduced to vector similarity alone: instead, it can be mediated by the model's prior exposure to large amounts of text and its ability to align unseen examples with meaningful linguistic labels. This seems especially valuable in domains such as administrative law, where categories are often discursive in nature and difficult to capture through keywords or predefined embeddings.

To explore this possibility, initial experiments were carried out with three transformer-based models: BART, DistilRoBERTa-base, and T5. When tested in zero-shot, the three models showed very limited capacity to perform the classification task. As

visible in Tables 3.3, 3.4 and 3.5, their overall accuracy was very low—0.08 for BART, 0.11 for T5, and 0.06 for RoBERTa—confirming that none of them could understand the linguistic or conceptual structure required by the task. Although varying in structure and training objectives, their difficulties were surprisingly comparable, suggesting that the underlying limitation is more in the data and the classification problem than in the models themselves.

BART performed slightly better than the others in some individual cases, with a recall of 0.30 for class 1, 0.25 for class 11, and 0.24 for class 6, but its precision and F1-scores also remained extremely low with a macro-average F1 of a mere 0.04. The majority of the categories, like classes 7, 14, and 15 that are well represented in the corpus, were not recognized at all. This would be in line with the hypothesis that BART might be able to catch some surface patterns—perhaps duplicating lexical forms, but was unable to build up an effective internal representation of the underlying categories. In other words, the model seemed to rely on fragments of textual evidence rather than on a genuine appreciation of the form and function of the documents.

T5 showed the same trend but with modestly higher scores. Its accuracy was 0.11 and its macro-average F1 0.09. The model at times was highly accurate but for a very few cases (classes 1, 3, 4, 7 and 17) where it was sure but missed most of the relevant documents. This skew indicates that T5 was at times capable of detecting strongly prototypical examples but still had narrow understanding relying heavily on surface features. The model responses are reminiscent of an algorithm that "guesses right" every so often but never produces a stable correspondence between the words and the conceptual classes.

RoBERTa, despite being the most experienced transformer encoder, performed even worse. It reported an accuracy of 0.06 and a macro F1 of 0.06 and was unable to produce coherent predictions for almost all classes. Some isolated improvements were observed, e.g., in classes 5, 7, 14, and 18, but these were exceptions rather than patterns. RoBERTa favored high-frequency or semantically general classes, as if trying to avoid making mistakes to the smallest extent possible by resorting to the most frequent tags. This outcome underlines the extent to which pre-trained models on open-domain datasets fail to transfer to the dense, highly formalized, and context-rich language of administrative and legal documents.

Collectively, these results reveal a common shortcoming of smaller zero-shot models. Although they are technically capable of interpreting label descriptions in natural language, their general-purpose pre-training does not equip them to handle long, structurally complex, and highly terminologically specialized documents. The administrative acts used within this test are approximately 3000 words long on average,

with frequent referencing to legislation and bureaucratic procedures that cannot be addressed by superficial text matching. Presented with this sort of material, the models appeared overwhelmed, not capable of disambiguating rival categories, nor the contextual cues that distinguish between one category of actions from another. In other words, the added expressiveness of natural language labels was not sufficient to compensate for the restricted size and training data of these models.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.05	0.30	0.09	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.33	0.11	0.16	19
3 - Risorse Umane	0.13	0.14	0.14	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0	0	0	15
5 - Sistema Informativo	0.05	0.11	0.07	9
6 - Programmazione, Coordinamento e Controllo	0.20	0.24	0.22	41
7 - Agricoltura e Allevamento	0	0	0	35
8 - Artigianato	0	0	0	3
9 - Commercio	0.03	0.20	0.05	5
10 - Industria e Attività Estrattive	0	0	0	6
11 - Turismo e Strutture Ricettive	0.08	0.25	0.12	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	0	0	0	6
13 - Infrastrutture e Trasporti	0	0	0	8
14 - Tutela dell'Ambiente	0	0	0	15
15 Sanità, Igiene e Veterinaria	0	0	0	28
16 - Politiche Sociali	0	0	0	7
17 Istruzione, Formazione e Lavoro	0	0	0	11
18 - Beni e Attività Culturali	0	0	0	15
19 - Attività Sportive e Ricreative	0	0	0	6
accuracy			0.08	264
macro avg	0.05	0.07	0.04	264
weighted avg	0.07	0.08	0.07	264

Table 3.3: Classification report of facebook/BART-large-mnli.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	1	0.10	0.18	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.05	0.37	0.09	19
3 - Risorse Umane	1	0.05	0.09	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	1	0.07	0.12	15
5 - Sistema Informativo	0.04	0.33	0.07	9
6 - Programmazione, Coordinamento e Controllo	0.22	0.22	0.22	41
7 - Agricoltura e Allevamento	1	0.03	0.06	35
8 - Artigianato	0	0	0	3
9 - Commercio	0.33	0.20	0.25	5
10 - Industria e Attività Estrattive	0	0	0	6
11 - Turismo e Strutture Ricettive	0	0	0	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	0	0	0	6
13 - Infrastrutture e Trasporti	0.33	0.12	0.18	8
14 - Tutela dell'Ambiente	0	0	0	15
15 Sanità, Igiene e Veterinaria	0	0	0	28
16 - Politiche Sociali	0	0	0	7
17 Istruzione, Formazione e Lavoro	1	0.09	0.17	11
18 - Beni e Attività Culturali	0.60	0.20	0.30	15
19 - Attività Sportive e Ricreative	0	0	0	6
accuracy			0.11	264
macro avg	0.35	0.09	0.09	264
weighted avg	0.44	0.11	0.11	264

Table 3.4: Classification report of sjrhuschlee/flan-T5-base-mnli.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.03	0.10	0.05	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.08	0.05	0.06	19
3 - Risorse Umane	0	0	0	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.04	0.07	0.05	15
5 - Sistema Informativo	0.17	0.11	0.13	9
6 - Programmazione, Coordinamento e Controllo	0	0	0	41
7 - Agricoltura e Allevamento	0.50	0.06	0.10	35
8 - Artigianato	0	0	0	3
9 - Commercio	0	0	0	5
10 - Industria e Attività Estrattive	0	0	0	6
11 - Turismo e Strutture Ricettive	0	0	0	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	0.07	0.17	0.10	6
13 - Infrastrutture e Trasporti	0	0	0	8
14 - Tutela dell'Ambiente	0.22	0.13	0.17	15
15 Sanità, Igiene e Veterinaria	0.12	0.11	0.12	28
16 - Politiche Sociali	0.08	0.14	0.10	7
17 Istruzione, Formazione e Lavoro	0	0	0	11
18 - Beni e Attività Culturali	0.17	0.20	0.18	15
19 - Attività Sportive e Ricreative	0	0	0	6
accuracy			0.06	264
macro avg	0.08	0.06	0.06	264
weighted avg	0.12	0.06	0.06	264

Table 3.5: Classification report of distilbert/distilroberta-base.

This does not mean, however, that the zero-shot approach should be abandoned. On the contrary, the results suggest that model scale and training coverage are critical factors. Larger models—such as GPT, Gemini or LLaMA—come with a far broader linguistic repertoire and a stronger ability to interpret prompts alongside lengthy input documents. It therefore makes sense to extend the experiments to these systems and to see whether the poor performance observed with smaller transformers was an intrinsic limitation of zero-shot classification, or simply a consequence of having used models that were not powerful enough, this analysis is explored in depth in Capter 6. The move from clustering to zero-shot classification, and then from smaller to larger models, is thus a logical progression in the research: it allows us to test not only a different method, but also to assess the impact of scale on the feasibility of automatic classification in this kind of domain.

Chapter 4

Preliminary Evaluation of Large Language Models

In the previous chapter, the challenges of applying unsupervised methods to administrative acts were highlighted. The lack of labeled data and the unstructured nature of the text hindered the effectiveness of traditional classification techniques. In such contexts, classical supervised or semi-supervised approaches are not feasible due to the absence of reliable training data and the difficulty of defining consistent features.

LLMs, however, offer an alternative through zero-shot classification, where predictions are generated directly from natural language prompts without the need for task-specific training. The use of LLMs in this setting is particularly promising for complex and heterogeneous documents, as models can leverage their internal knowledge base rather than relying solely on labeled datasets or handcrafted features.

To assess the real potential of zero-shot classification with LLMs, this chapter focuses on two experimental domains in which labeled data was available to us. Unlike the case of administrative acts presented in Chapter 3, the availability of annotated datasets allowed for a systematic evaluation of model performance. The two domains were selected based on their different levels of structure and standardization:

- Medical documents, a domain characterized by established terminology and guidelines, where we hypothesized that LLMs would perform reliably.
- Asylum requests in Italy, a socially and legally complex domain where the absence of unified public guidelines and high contextual variability was expected to challenge LLM predictions.

By comparing these two experiments, we aim to understand under which conditions LLM-based zero-shot classification can provide accurate results, and where its limitations become evident. The outcomes of this evaluation will serve as a foundation for the methodology proposed in the following chapter.

4.1 Experimental Setup

The experiments presented in this chapter were designed to evaluate the capability of LLMs to perform zero-shot reasoning in domains with varying levels of domain-specific knowledge and standardization. Rather than addressing a strict text classification task, the experiments focused on testing the models' ability to provide accurate decisions when presented with domain-related questions.

Models Under Evaluation

Three state-of-the-art LLMs were selected for testing: GPT-4o-mini, Gemini-pro, and Llama 3.1 (8B Instruct version). The comparison aimed to highlight not only overall performance but also potential differences in domain knowledge across models of different sizes and costs, and developed by different providers. Specifically, the tests performed on Llama aim to understand whether the tasks can be addressed using a model that is free and deployable on a personal computer. Where Llama fails, we test whether a model that can be used for free (i.e., Gemini) is able to compete with GPT, whose API has a per token cost that Public Administrations may not be able to sustain. Many other models could have been tested, but due to our limited resources we chose these ones as they are representative of the three different categories.

Tasks

In the medical domain, models were asked to identify the correct diagnosis given a short clinical case. This task reflects a well-defined decision problem, as medical diagnoses follow established guidelines strictly related to the described symptoms and rely on standardized terminology.

In the asylum domain, models were presented with simplified case descriptions of asylum requests in Italy, consisting of attributes such as country of origin, gender, and other potential risk factors for discrimination. The task was to decide whether asylum should be granted and, if so, to assign the appropriate protection status (refugee status,

subsidiary protection, or special protection). Unlike the medical domain, this task does not have a universally standardized framework, as decisions depend on complex and context-dependent legal and social considerations.

Data and Preprocessing

In both experiments, no significant preprocessing was required. The clinical cases were concise and free of misleading details, while asylum-related cases were provided as structured lists of attributes. This simplicity allowed the evaluation to focus on the models' intrinsic reasoning and domain knowledge rather than on input normalization.

Prompting Strategy

Prompts were written in plain natural language without engineered templates or additional instructions. The goal was to assess the baseline ability of LLMs to interpret and respond to the tasks without relying on extensive prompt optimization.

Evaluation of Correctness

Since both datasets contained human-annotated ground truth labels, the correctness of the model outputs could be directly assessed. This availability of labeled data distinguishes the current experiments from those discussed in Chapter 3, enabling a more systematic and reliable evaluation of performance.

4.2 Experiment 1: Medical Domain

This experiment was conducted during a research period abroad and builds upon a line of work that had already been initiated prior to this thesis. The original study focused on the use of a domain-specific medical ontology to support text classification tasks, with the goal of ensuring consistency and interpretability in the annotation process. The experiment presented here takes that work as a starting point and explores a related, yet distinct, question: whether large language models can serve as a reliable alternative to ontology-based approaches in the same domain. Rather than assuming ontological knowledge as an explicit, structured resource, the analysis investigates to what extent such knowledge can be implicitly recovered through prompting alone. This comparison provides a concrete setting in which to assess the strengths and limitations of LLMs when

confronted with tasks that traditionally rely on carefully curated domain representations, as described in the earlier ontology-driven study [113].

4.2.1 Context and Data

The first evaluation was carried out in the medical field [114], using cases drawn from the Antidote Casimedicos Dataset [115], a multilingual collection of short clinical scenarios accompanied by multiple-choice questions. Each case included a patient description and several possible answers, one of which had been validated by medical experts. Although the dataset covers a range of questions (e.g. diagnosis, treatment, additional examinations, clinical considerations), for this study we restricted the analysis to diagnostic tasks.

The medical domain was chosen because it offers a relatively structured environment: clinical practice is guided by internationally shared terminology and well-established diagnostic criteria. This makes it particularly suitable for assessing whether LLMs can apply general medical knowledge in a controlled experimental setting.

4.2.2 Task Formulation

In practical terms, the models were asked to select the correct diagnosis from a list of alternatives, based only on the short case description provided. The task therefore goes beyond surface-level text matching and requires a degree of medical reasoning. In a second phase of the experiment, the models were also invited to produce a short explanation of their choice, clarifying why the selected option was correct and why the others should be excluded. This additional step allowed us to test not only the accuracy of the responses but also the models' ability to generate reasoning that was coherent with the clinical information.

4.2.3 Results

The comparison between GPT, Gemini, and Llama, visible in Table 4.1, showed clear performance differences. GPT consistently obtained the best results, with the lowest number of incorrect diagnoses across all three languages tested (English, French, and Spanish). Gemini performed at an intermediate level, while Llama accumulated the highest error rates.

When the explanatory paragraph was included in the prompt, the models' behavior

LLM	Language	Number of wrong diagnoses
GPT	English	20
	French	22
	Spanish	24
Llama	English	71
	French	94
	Spanish	93
Gemini	English	57
	French	64
	Spanish	56

Table 4.1: Number of wrong diagnosis, out of 158, for every LLM in every language. These results refer to the maximum number of errors in a run.

diverged further. While GPT's performance remained consistent, Llama and Gemini exhibited a significant decline in accuracy. Llama failed in an average of 110 diagnoses, and Gemini in approximately 90, which was unexpected given that the prompt was identical, except for the addition of the explanatory paragraph. Furthermore, GPT's explanations were notably more structured. For example, this is an answer given by GPT, where it analyzes all the possible answers and relates them to the patient's symptoms:

Prompt:
This is a clinical case: A 65-year-old woman who underwent hematopoietic precursor transplantation 30 days ago and has severe neutropenia. She starts with cough and hemoptotic expectoration in addition to fever and moderate exertional dyspnea that does not improve despite five days of treatment with amoxicillin-clavulanic acid (875/125 mg/8 h) and levofloxacin (500 mg/12 h). Chest X-ray shows multiple ill-defined pulmonary nodules, some of them cavitated. Among the following, which is the most probable diagnosis of suspicion? These are the possible answers:
1: Septic embolisms.
2: Invasive pulmonary aspergillosis.
3: Pneumonia due to Staphylococcus aureus.
4: Systemic candidiasis.
Which one is the correct option? Generate a text that explains which is the correct option and why, and also why the other options are incorrect. The first line of the text must contain the number of the correct option.

GPT's Answer:
Invasive pulmonary aspergillosis is most likely the correct diagnosis given the patient's history and the unique combination of symptoms. Specifically, her recent transplantation and subsequent neutropenia place her at a higher risk of opportunistic fungal infections such as aspergillosis. The multiple cavitary pulmonary nodules on the chest X-ray could suggest the angiitis produced by this fungus. Furthermore, the poor response to current antibiotic treatment also suggests a fungal rather than a bacterial cause.
Option 1, Septic embolisms, is less likely because they most commonly associate with endocarditis or venous thrombosis, neither of which is mentioned in this case. Option 3, Pneumonia due to Staphylococcus aureus, is less likely as well. Although Staphylococcus aureus can induce pneumonia, particularly in patients with impaired immune systems, it typically responds to the antibiotic regimen that is already being administered (such as amoxicillin-clavulanic acid and levofloxacin) but the patient has not shown any improvement. Systemic candidiasis (option 4) typically presents with more systemically dispersed symptoms and not limited to respiratory issues, hence making it less likely in this case than invasive pulmonary aspergillosis.

Info:
Correct option: 2
GPT's answer: 2

In contrast, Gemini's explanations did not always effectively relate the incorrect diagnoses

to the clinical case, merely listing symptoms associated with the alternative diagnoses:

Prompt:
This is a clinical case: A 65-year-old woman who underwent hematopoietic precursor transplantation 30 days ago and has severe neutropenia. She starts with cough and hemoptotic expectoration in addition to fever and moderate exertional dyspnea that does not improve despite five days of treatment with amoxicillin-clavulanic acid (875/125 mg/8 h) and levofloxacin (500 mg/12 h). Chest X-ray shows multiple ill-defined pulmonary nodules, some of them cavitated. Among the following, which is the most probable diagnosis of suspicion? These are the possible answers:
1: Septic embolisms.
2: Invasive pulmonary aspergillosis.
3: Pneumonia due to Staphylococcus aureus.
4: Systemic candidiasis.
Which one is the correct option? Generate a text that explains which is the correct option and why, and also why the other options are incorrect. The first line of the text must contain the number of the correct option.

Gemini's Answer:
The most probable diagnosis of suspicion is septic embolisms because the patient is neutropenic and has multiple pulmonary nodules, some of them cavitated. This is a classic presentation of septic embolisms, which are caused by the hematogenous spread of bacteria to the lungs.
Option 2 is incorrect because invasive pulmonary aspergillosis is more common in patients with prolonged neutropenia and is typically characterized by the presence of hyphae on microscopy.
Option 3 is incorrect because pneumonia due to Staphylococcus aureus is more likely to cause lobar consolidation on chest X-ray.
Option 4 is incorrect because systemic candidiasis is more likely to cause disseminated infection with involvement of multiple organs.

Info:
Correct option: 2
Gemini's answer: 1

Llama provided more contextual information in its explanations, but, like Gemini, demonstrated inconsistencies in medical knowledge, and sometimes could not keep track of the numbers of the possible choices, resulting in a wrong answer even when the reasoning was on point:

Prompt:
This is a clinical case: A 65-year-old woman who underwent hematopoietic precursor transplantation 30 days ago and has severe neutropenia. She starts with cough and hemoptotic expectoration in addition to fever and moderate exertional dyspnea that does not improve despite five days of treatment with amoxicillin-clavulanic acid (875/125 mg/8 h) and levofloxacin (500 mg/12 h). Chest X-ray shows multiple ill-defined pulmonary nodules, some of them cavitated. Among the following, which is the most probable diagnosis of suspicion? These are the possible answers:
1: Septic embolisms.
2: Invasive pulmonary aspergillosis.
3: Pneumonia due to Staphylococcus aureus.
4: Systemic candidiasis.
Which one is the correct option? Generate a text that explains which is the correct option and why, and also why the other options are incorrect. The first line of the text must contain the number of the correct option.

Llama's Answer:
This patient's symptoms and presentation are most consistent with invasive pulmonary aspergillosis. The patient's recent history of hematopoietic precursor transplantation and severe neutropenia puts her at high risk for invasive fungal infections, particularly aspergillosis. The symptoms of cough, hemoptotic expectoration, fever, and dyspnea that do not improve with antibiotic therapy are classic for invasive pulmonary aspergillosis. The chest X-ray findings of multiple ill-defined pulmonary nodules, some of them cavitated, are also consistent with this diagnosis. Invasive pulmonary aspergillosis is a life-threatening infection that requires prompt treatment with antifungal therapy, such as voriconazole. The other options are less likely: septic embolisms are not typically associated with cavitated pulmonary nodules, pneumonia due to Staphylococcus aureus would likely respond to the antibiotics she has been taking, and systemic candidiasis is less likely given the specific symptoms and imaging findings.

Info:
Correct option: 2
Llama's answer: **1**

GPT was the most successful in contextualizing its responses, consistently referring to the patient's symptoms when justifying both correct and incorrect diagnoses. However, GPT also occasionally exhibited gaps in medical knowledge, such as incorrectly stating that a specific maneuver should result in an increase in a patient's vital sign when it should

decrease. Additionally, GPT sometimes went beyond identifying the correct diagnosis, opting for the underlying illness when the diagnosis was a complication of that condition, if the former was among the options.

4.2.4 Discussion

Overall, the experiment confirms that bigger LLMs are capable of dealing with diagnostic questions when the domain is well-defined and strongly represented in their training data. GPT, in particular, demonstrated a capacity to combine accuracy with plausible reasoning, which suggests that its internal knowledge base covers a substantial portion of standard medical practice. On the other hand, Llama's behavior confirms that, even in a domain where the knowledge is public such as the medical one, a smaller model is not fully reliable; even when it possesses the correct knowledge it can make mistakes such as mixing up the answers.

Furthermore, the study points to important limitations. Even the best-performing model occasionally confused related conditions, for instance preferring an underlying disease to the specific complication asked in the question. The performance drop observed in Gemini and Llama when explanatory output was required also illustrates how difficult it still is to reconcile accuracy and explainability in medical applications.

4.3 Experiment 2: Asylum Requests in Italy

This experiment is part of a broader research aimed at supporting organizations that assist migrants in the asylum application process [116]. Specifically, with the help of domain experts, we developed a chatbot, whose interface is visible in Figure 4.1, that asks a series of questions and, based on the person's answers, returns an indication of what kind of protection they are more likely to get and why. This experiment has a qualitative purpose: it aims to explore what is the level of knowledge that LLMs possess in this domain and whether they can assist in distinguishing different levels of protection.

4.3.1 Context and Data

The second experiment was conducted on asylum requests in the Italian context. Each case included a set of attributes describing the applicant's background—such as gender, country of origin, experiences of violence, or vulnerability—and the task was to determine whether protection should be granted and, if so, which type of protection applied. The

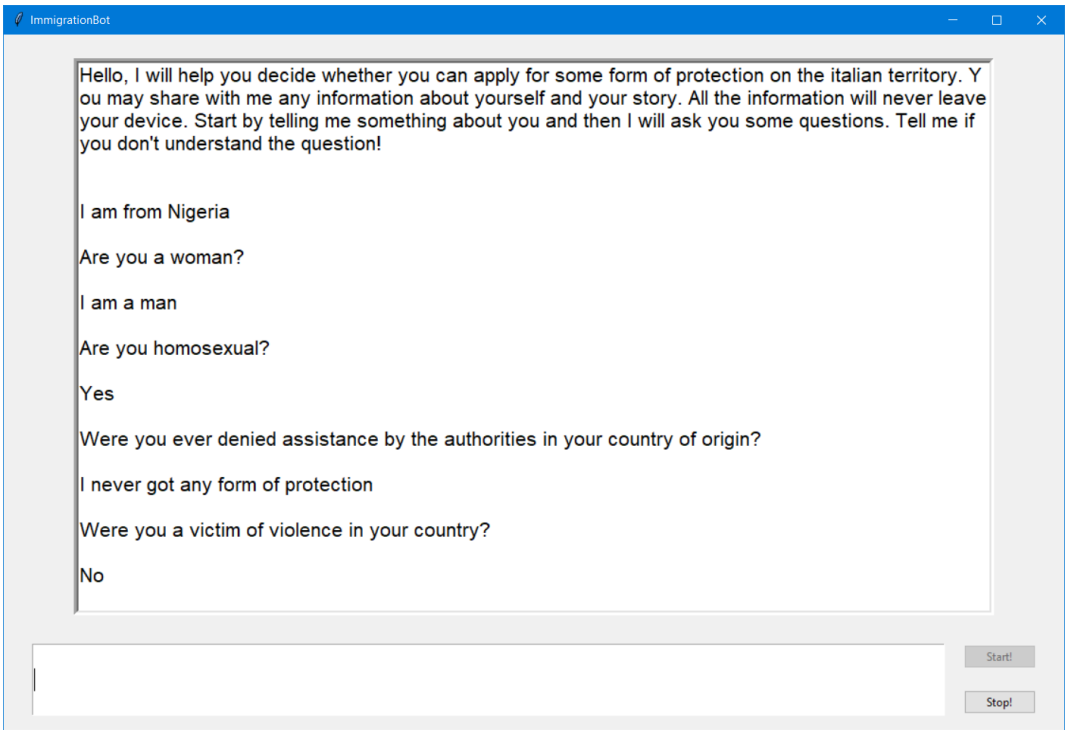


Figure 4.1: User interface of the chatbot for asylum-seeking migrants in Europe [116].

human expert's decisions served as the reference point, reflecting the real practice in which international protection is rarely guaranteed and most applications end with a negative outcome.

This scenario differs significantly from the medical case study. While medicine relies on widely shared guidelines and codified knowledge, asylum law in Italy is characterized by legal complexity and limited availability of uniform public criteria. As a result, judgments are often context-dependent and require a nuanced evaluation that is difficult to reproduce in an automated way.

4.3.2 Task Formulation

In this experiment, the task assigned to the models was to determine the appropriate legal outcome for each asylum request. Four possible categories were considered: no protection, special protection, subsidiary protection, and refugee status. These categories reflect the main forms of international protection recognized in the Italian legal framework, with "no protection" representing the most frequent real-world outcome and the other three representing different levels of safeguards available under specific conditions.

The input to the models consisted of a simplified description of the applicant, represented as a set of attributes such as country of origin, gender, history of violence or trafficking, sexual orientation, and other elements that might increase the individual's vulnerability. No additional context or background knowledge was supplied, the prompts were written in plain natural language and directly listed the applicant's characteristics, followed by the instruction to decide whether protection should be granted and, if so, which category applied.

This design was intended to be minimalist. It was not the intent to mimic the full sophistication of legal argument deployed by commissions or judges, but to see whether LLMs could reflect human experts' judgments where provided with the essential facts. The approach thus eschewed fine-tuning and particular legal guidance and relied on models' internal knowledge and their own understanding of humanitarian and legal principles.

By constructing the task in this way, it was then possible to evaluate not just the ability of the models to provide a decision, but also how far their reasoning departed from, or matched, the stricter, and usually more restrictive, criteria used in real asylum adjudication.

4.3.3 Results

The distribution of results revealed a marked difference between human experts and the models. While experts denied protection in the majority of cases, the models tended to grant some form of protection almost systematically, as visible in Table 4.2, where we reported 7 different cases.

Gemini and Llama displayed a uniform pattern, the former assigning special protection in nearly all cases, with only one instance of refugee status; the latter showing an extreme bias toward subsidiary protection, which it assigned consistently across all cases.

GPT produced more varied outcomes, granting refugee status in several cases and special protection in others, while rarely opting for denial of protection.

Overall, the LLMs proved to be more "generous" than the expert benchmark, often opting for higher levels of protection even in cases where the human decision was negative.

4.3.4 Discussion

This experiment illustrates the difficulty LLMs face when confronted with domains that lack clear and universally recognized criteria. In contrast with the medical task, where

Question	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Are you a woman?	yes	no	yes	no	yes	no	yes
Were you a victim of the Voodoo Juju ritual?	no	no	yes	yes	no	no	no
Do you have a low education level?	no	no	no	no	yes	no	yes
Are you in a precarious economic condition?	no	no	no	no	no	yes	no
Did you have issues during the journey?	yes	no	yes	no	yes	no	no
Did you receive threats?	no	yes	no	no	no	yes	no
Are you from Nigeria?	no	no	yes	yes	yes	yes	no
Are you a victim of violence?	no	yes	no	no	no	yes	no
Are you a victim of human trafficking?	yes	no	no	no	yes	no	no
Are you homosexual?	no	no	yes	no	yes	no	no
Were you denied protection in your country?	yes	no	no	no	no	no	yes
Do you have a job in this country?	yes	yes	yes	yes	yes	yes	no
Are you a vulnerable subject?	yes	no	yes	no	no	no	yes
Answers							
Gold	Special	No prot.	No prot.	No prot.	No prot.	Special	No prot.
Gemini	Special or Subsidiary	Special	Refugee status	Special	Special	Special	Special
GPT	Special	Special	Refugee status	No prot.	Refugee status	Subsidiary	Special
Llama	Subsidiary	Subsidiary	Subsidiary	Subsidiary	Subsidiary	Subsidiary	Subsidiary

Table 4.2: LLMs answers for the protection assignment task.

structured knowledge supported accurate predictions, the asylum context exposed the tendency of LLMs to rely on general humanitarian reasoning rather than on the restrictive interpretation applied in practice.

The three models behaved differently, but they all shared a strong inclination to grant protection. Gemini displayed consistency but little differentiation, GPT was more flexible but often overly optimistic, while Llama fell into the opposite extreme of systematically selecting the same outcome, a result that might be caused by the smaller size of the model when compared to Gemini and GPT. These patterns suggest that, without explicit legal guidance, LLMs are unable to reproduce the nuanced and often restrictive approach followed by experts.

The results therefore underline a crucial limitation: zero-shot reasoning with LLMs is not sufficient for decision-making in complex legal and social domains, where outcomes depend less on general knowledge and more on detailed contextual interpretation.

4.4 Cross-Experiment Comparison

The two experiments make clear that the effectiveness of LLMs is strongly shaped by the type of domain in which they are applied. In the medical context, the models could rely on well-defined knowledge, clear diagnostic categories, and guidelines that are largely shared across countries. This framework helped especially GPT to reach results close to those of the expert, and in many cases the explanations it provided were coherent

with the symptoms described. When mistakes occurred, they usually involved fine distinctions—such as confusing an underlying disease with its complication—rather than a complete misinterpretation of the case.

The asylum setting told a very different story. Here the absence of universally applied rules and the variability of real decisions highlighted the limits of zero-shot reasoning. All three models were noticeably more inclined than the experts to grant some form of protection, often in situations where the actual decision would have been negative. This suggests that, in the face of vague or inconsistent legal criteria, the models fall back on general humanitarian reasoning rather than mirroring the more restrictive approach typically adopted in practice.

Looking at the models individually, their behavior across the two domains also differed. GPT was the most convincing in the medical task and, in the asylum task, showed more variation than the others, even if it tended to lean towards overly positive outcomes. Gemini was less flexible: it performed reasonably in the medical cases but, when applied to asylum requests, mostly repeated the same answer, usually special protection. Llama, being a smaller model, struggled more than the others, with lower accuracy in the clinical cases and a strong bias toward always assigning subsidiary protection in the asylum task.

Taken together, these findings suggest a broader conclusion: LLMs can be useful tools when the field is structured, standardized, and well covered in the training data, but they become unreliable in areas where decisions are context-dependent and shaped by interpretation, such as asylum law.

Chapter 5

A Pipeline for LLMs in a Data Scarce Regime

From this chapter on, we shift our focus on classifications tasks where the annotated data was very limited or, as in this case, initially unavailable. In many real-world scenarios, including the classification of scientific articles according to the United Nations Sustainable Development Goals (SDGs), the cost of manual annotation is prohibitively high and the required expertise is scarce. Traditional supervised approaches cannot be applied effectively without a sufficiently large and balanced training set. This motivates the development of alternative strategies that minimize the burden on human experts while enabling the creation of reliable datasets for model training.

In this chapter, we introduce a classification pipeline specifically designed for domains where labeled data is scarce or unavailable. The pipeline is based on three main steps:

1. Gold dataset creation, where a small collection of documents is annotated by domain experts following carefully designed guidelines;
2. Silver dataset creation, where LLMs are prompted with the same guidelines to automatically annotate a larger corpus of unlabeled documents, producing a scalable silver resource;
3. Training lightweight classifiers, where machine learning models are trained on the silver dataset and evaluated against the gold benchmark to assess their effectiveness.

The methodology was validated in the context of scientific publications, where the task was to assign one or more SDG labels to a paper based on its title, abstract, and keywords [117]. In this setting, we created the AlmaSDG-gold dataset, a high-quality but small

benchmark annotated by sustainability experts, and the AlmaSDG-silver dataset, a larger automatically labeled collection obtained through knowledge distillation from LLMs. By comparing LLM-generated labels with expert annotations, we identified the most reliable model and used it to construct the silver dataset.

The proposed approach provides a scalable compromise between the reliability of human annotations and the efficiency of automatic labeling. It enables the training of specialized, lightweight, proprietary classifiers—such as Logistic Regression, Support Vector Classifiers, and Transformer-based models—that can be deployed in contexts where the direct use of large language models is either impractical or undesirable. This chapter details the pipeline, the construction of both datasets, and the comparative evaluation of different models trained under this methodology.

5.1 Pipeline Overview

The proposed pipeline addresses the lack of large, high-quality labeled datasets by combining minimal human supervision with automatic annotation. It is designed to be flexible and domain-agnostic, making it suitable for classification tasks where manual labeling is costly or impractical. The pipeline unfolds through three main stages:

1. **Gold Dataset Creation:** A small set of documents is annotated by domain experts following well-defined guidelines.

   This gold dataset serves two purposes: (i) it provides a high-quality benchmark for evaluation, and (ii) it enables the calibration and validation of LLM-based automatic labeling.

2. **Silver Dataset Creation:** LLMs are prompted using the same annotation guidelines applied by experts.

   Their predictions are compared against the gold annotations to identify the most reliable model.

   The selected model is then applied to a larger pool of unlabeled documents, producing a silver dataset: a scalable resource that preserves the annotation consistency of the guidelines while drastically reducing the human effort required.

3. **Model Training and Evaluation:** Lightweight classifiers are trained on the silver dataset and evaluated on the gold dataset to measure their effectiveness.

   The evaluation includes comparisons with direct LLM predictions and, if present,

with models trained on alternative datasets, highlighting the trade-offs between scalability, efficiency, and accuracy.

This pipeline achieves a pragmatic balance: the gold dataset guarantees reliability, the silver dataset ensures scalability, and the lightweight classifiers provide efficiency and adaptability. Crucially, the approach is iterative: improvements in guidelines or in LLM capabilities can be seamlessly integrated to refine the silver dataset and retrain the classifiers.

To demonstrate the effectiveness of this methodology, we instantiated the pipeline in the domain of scientific publications with the task of classifying articles according to the UN Sustainable Development Goals (SDGs). The first step in this process was the construction of a high-quality gold dataset, annotated by domain experts following carefully designed guidelines. The next section provides a detailed account of the gold dataset creation process.

5.2 Gold Dataset Creation

The first step of the proposed pipeline is the creation of a gold dataset, a small but reliable collection of expert-annotated documents. This dataset is crucial as it provides both a reliable benchmark for evaluating the performance of the LLMs and the foundation for calibrating LLM-based annotation, acting as a reference set for selecting the most aligned LLM, which will then be used to annotate a much larger collection of documents in the creation of the silver dataset. In our case study, the task was to identify the contributions of scientific articles to the SDGs.

When building a gold dataset in low-resource settings, methodological rigor becomes a crucial point, as the limited size of the dataset amplifies the impact of every annotation decision. First, label definitions must be made as explicit and practical as possible. In domains where formal guidelines are not present, categories often remain conceptually clear but practically ambiguous, leading to overlaps and inconsistent interpretations. It is therefore essential to develop concrete annotation criteria, supported by examples and counterexamples. Second, consistency is more important than size: a smaller but internally coherent dataset is preferable to a larger noisier one. Iterative refinement of the annotation guidelines through multiple annotation sessions followed by discussions with domain experts, can significantly reduce ambiguity and edge cases. Third, even if perfect balance is unattainable, awareness of potential skewness helps interpret evaluation results and prevents overestimating model performance on dominant classes. Finally,

documentation of the annotation rationale is crucial. In low-resource scenarios, the gold dataset does not merely serve as ground truth, but also as a knowledge artifact: it encodes expert reasoning patterns that will later guide LLM prompting and silver data generation. For this reason, transparency, traceability, and clear justification of labeling choices become as important as the labels themselves. In the following subsections, the creation of a gold dataset is described using the SDGs detection case study as a practical example for the description of the pipeline workflow.

5.2.1 Data Source and Selection

We considered as candidate documents the titles, abstracts, and keywords of scientific publications from a comprehensive institutional repository[1]. Full texts were not included, as metadata alone was deemed sufficient to capture the core thematic contributions while keeping the annotation workload manageable. Given the broad scope of the 17 SDGs and their uneven distribution across disciplines, we applied keyword-based filtering—derived from the annotation guidelines—to select a set of potentially relevant articles. From this pool, 139 documents were manually curated to ensure coverage across all SDGs.

5.2.2 Annotation Guidelines

One of the main challenges in SDG classification is the ambiguity and interconnectedness of the goals. To mitigate subjectivity, we developed a set of annotation guidelines in collaboration with sustainability experts. These guidelines included:

- Clear definitions of what constitutes a direct or indirect contribution to an SDG;
- A list of 165 thematic contribution areas linked to the 17 SDGs;
- Positive and negative examples to illustrate borderline cases.

The guidelines were designed not only to harmonize expert judgments but also to be reusable for prompting LLMs, ensuring consistency between human and machine annotation.

5.2.3 Annotation Process

Annotation was performed by a multi-disciplinary team of experts and trained annotators. Each article was independently reviewed by at least two annotators, who assigned binary

[1] `https://www-scopus-com.ezproxy.unibo.it/pages/home?display=basic#basic`

labels for each of the 17 SDGs. Inter-annotator agreement (IAA), shown in Table 5.1, was measured using Cohen's Kappa, yielding an overall average of 0.54 (moderate agreement [118]), with some SDGs such as gender equality (SDG 5) reaching high consistency (0.86), while others, such as decent work and economic growth (SDG 8), proved more challenging (0.23). Disagreements were resolved through discussion and, when necessary, arbitration by additional experts.

SDG	IAA (g)	# Positive (g)	% Positive (g)	# Positive (s)	% Positive (s)
1	0.44	5	3.6	31	1.1
2	0.39	12	8.6	47	1.6
3	0.60	25	18.0	363	12.6
4	0.57	10	7.2	33	1.1
5	0.86	13	9.4	29	1.0
6	0.46	8	5.8	26	0.9
7	0.76	10	7.2	72	2.5
8	0.23	15	10.8	48	1.7
9	0.27	13	9.4	117	4.1
10	0.59	32	23.0	75	2.6
11	0.53	23	16.5	77	2.7
12	0.65	25	18.0	102	3.5
13	0.62	17	12.2	95	3.3
14	0.76	7	5.0	35	1.2
15	0.74	9	6.5	38	1.3
16	0.24	11	7.9	51	1.8
17	0.49	7	5.0	3	0.1
Avg	0.54	14.24	10.24	73.06	2.54

Table 5.1: AlmaSDG corpus statistics. (g) stands for AlmaSDG-gold, (s) for AlmaSDG-silver. We report IAA for (g) and statistics on positive labels.

5.2.4 Dataset Characteristics

The resulting dataset, which we refer to as AlmaSDG-gold, comprises 2,363 annotations across 139 documents, with each document potentially linked to multiple SDGs. As visible in Table 5.2, only 13 documents were found to contribute to none of the goals, while nearly half were associated with multiple goals, reflecting the multi-label and interdisciplinary nature of the task. The distribution of positive labels, visible in Table 5.1, is uneven, with SDG 10 (reduced inequality) exhibiting the highest number of positive instances (32),

followed by SDG 12 (responsible consumption and production) and 3 (good health and well-being). On the other hand, SDG 1 (no poverty), 14 (life below water), 17 (partnerships for the goals), 6 (clean water and sanitation), and 15 (life on land) have less than 10 positive instances, which reflects the difficulty in retrieving an even amount of articles.

# of labels per document	**0**	**1**	**2**	**3**	**4+**
# of documents (AlmaSDG-gold)	13	57	39	19	11
# of documents (AlmaSDG-silver)	1985	646	170	56	21

Table 5.2: Number of labels associated with each document.

5.3 Silver Dataset Creation

While the gold dataset provides a reliable benchmark, its small size limits its use as training data for machine learning models. To address this, we proceed by constructing a silver dataset, a larger automatically annotated collection obtained through knowledge distillation from LLMs. This step enables the scalability of the pipeline while maintaining alignment with expert guidelines.

Manual annotation by experts for this case study has, in fact, proven to be costly and time-consuming, making it infeasible to build large-scale resources covering all 17 SDGs. However, recent evidence suggests that LLMs can perform annotation tasks with a level of accuracy comparable to humans, especially when guided by clear instructions. Leveraging this observation, we designed a method to expand our labeled data by using LLMs as automatic annotators. This step will lead to the creation of a silver dataset, annotated by the LLM that demonstrates a higher agreement with the human annotators, measured using the gold dataset.

This silver dataset will provide the necessary scale for training machine learning models while remaining anchored to human-defined annotation principles. Together with the gold dataset, it will establish a two-tiered resource: gold for evaluation and calibration, silver for training. This combination will allow us to balance annotation quality with data availability, ultimately enabling the development of specialized, lightweight classifiers for classification. The following subsections describe the creation of the silver dataset for the SDGs detection case study.

5.3.1 LLM Selection

Before producing the silver dataset, we evaluated several LLMs—specifically GPT, Gemini, and Llama[2]—on the gold dataset under a zero-shot classification setting. Each model was prompted using the same annotation guidelines developed for human annotators, ensuring consistency across human and machine labels. We then measured the agreement between LLM predictions and expert annotations.

Table 5.3 reports all the agreement scores, computed as the average IAA between the LLMs and each human annotator; GPT showed the highest alignment with human labels, with an average agreement of 0.57. Gemini followed closely (0.51), while Llama showed lower consistency (0.35).

While the substantial agreement among the larger LLMs indicates a robustness of our silver data creation method across possible backbones, the inferior agreement between LLMs and human experts might indicate that determining a scientific article's contribution to an SDG requires a level of understanding not yet achieved by state-of-the-art LLMs.

Given these results, we selected GPT as the annotator for the silver dataset.

5.3.2 Automatic Annotation Procedure

Using GPT, we annotated an additional set of 2,878 publications randomly sampled from the same institutional repository used for the gold dataset. For each article, the model was provided with the title, abstract, and keywords and was asked to assign zero or more SDG labels according to the established guidelines. To ensure reproducibility and reduce randomness, inference was performed with a temperature set to 0.

5.3.3 Dataset Characteristics

The resulting resource, which we call AlmaSDG-silver, contains multi-label annotations for 2,878 documents. Unlike AlmaSDG-gold, which was curated to maximize SDG coverage, the silver dataset reflects the natural distribution of contributions in the corpus. As a result:

- A large proportion of documents received no SDG label, confirming that many scientific articles do not directly contribute to sustainable development goals.

[2]We used GPT-4o-mini, Gemini-1.5-flash-001, and Llama3-8B-Intruct.

SDG	Human-Llama	Human-GPT	Human-Gemini	GPT-Gemini	GPT-Llama	Gemini-Llama
1	0.20	0.37	0.37	0.66	0.35	0.34
2	0.56	0.60	0.63	0.73	0.55	0.70
3	0.32	0.69	0.41	0.52	0.44	0.43
4	0.34	0.60	0.56	0.76	0.33	0.38
5	0.75	0.67	0.89	0.57	0.49	0.75
6	0.54	0.66	0.63	0.71	0.64	0.63
7	0.40	0.75	0.66	0.85	0.56	0.70
8	0.16	0.34	0.30	0.43	0.07	0.14
9	0.22	0.37	0.38	0.64	0.36	0.60
10	0.39	0.44	0.46	0.69	0.37	0.44
11	0.44	0.52	0.48	0.67	0.48	0.60
12	0.55	0.80	0.64	0.68	0.57	0.80
13	0.34	0.68	0.54	0.77	0.48	0.60
14	0.42	0.72	0.71	0.81	0.64	0.64
15	0.17	0.77	0.35	0.40	0.17	0.47
16	0.11	0.31	0.24	0.46	0.09	0.18
17	0.08	0.46	0.39	0.26	0.03	0.13
Avg	0.35 ± 0.18	0.57 ± 0.16	0.51 ± 0.16	0.62 ± 0.16	0.39 ± 0.19	0.50 ± 0.21

Table 5.3: Agreement between human annotators and LLMs.

- SDG 3 (Good Health and Well-Being) emerged as the most frequent category, with 363 positive instances, likely due to the strong representation of medical publications in the source database.
- Other goals, such as SDG 17 (Partnerships for the Goals), were rare, with only 3 positive instances.

Tables 5.1 and 5.2 report all of this data.

5.4 Classification Model and Training Setup

Having established both gold and silver datasets, the next step of the pipeline is the training of classification models on the silver dataset and their evaluation on the gold benchmark. The aim is to determine whether lightweight supervised models, when trained on automatically annotated data, can achieve performance comparable to or better than LLMs used directly in zero-shot classification. The following subsections describe how light weight models were trained and tested for the SDGs detection case study.

5.4.1 Models Considered

We evaluated four representative models, spanning both traditional machine learning approaches and transformer-based architectures:

- Logistic Regression (LR) with TF–IDF features
- Support Vector Classifier (SVC) with TF–IDF features
- BERT (frozen encoder with a fine-tuned classification head)
- RoBERTa (frozen encoder with a fine-tuned classification head)

The inclusion of LR and SVC reflects their robustness in text classification with limited resources, while BERT and RoBERTa serve as state-of-the-art deep learning baselines.

5.4.2 Training Procedure

Input Representation: For LR and SVC, we extracted TF–IDF features from titles, abstracts, and keywords. Parameters included a maximum document frequency of 0.95, a maximum feature size of 15k, and an n-gram range of (1,2).

Hyperparameters: LR and SVC were trained with balanced class weights to mitigate label imbalance, with regularization set to C = 0.15 for LR and C = 0.05 for SVC.

Transformer Models: For BERT and RoBERTa, we froze the encoder layers and fine-tuned only the classification head. Training was performed for 15 epochs with a batch size of 16, using the Adam optimizer (lr = 1e-3, weight decay = 1e-5) and a dropout of 0.20 before the output layer. Each experiment was repeated with 5 different random seeds to ensure robustness.

5.4.3 Evaluation Protocol

Training Data: All models were trained exclusively on AlmaSDG-silver, our LLM-annotated dataset.

Test Data: Evaluation was conducted on AlmaSDG-gold, ensuring that test documents were never seen during training.

Metrics: Given the multi-label nature of the task, we reported macro-averaged precision, recall, and F1-score, as well as per-class results for the 17 SDGs.

Baselines: Results were compared against:

- Direct LLM predictions (GPT, Gemini, Llama in zero-shot classification)
- Existing external tools (Aurora [119], Elsevier [120], OSDG [121] classifiers)
- Models trained on alternative datasets (OSDG [122], SKH [123]), to contextualize performance differences.

This experimental setup allows us to assess the added value of silver data in enabling supervised training, while also positioning our approach relative to both human-expert gold standards and existing state-of-the-art resources.

5.5 Experimental Results on Scientific Texts

This section presents the results of our experiments, comparing the performance of direct LLM classification with that of supervised models trained on the AlmaSDG-silver dataset. The evaluation was carried out on AlmaSDG-gold, ensuring a rigorous benchmark based on human expert annotations. All the details are shown in Tables 5.4, 5.5, 5.6, 5.7and 5.8.

5.5.1 LLM Baselines

We first assessed the zero-shot classification ability of three large language models—GPT, Gemini, and Llama—using prompts derived from the annotation guidelines. The results indicate that GPT achieved the highest overall performance with a macro F1-score of 0.69, outperforming both existing SDG classifiers and other LLMs. Gemini performed competitively, with a macro F1 of 0.63, showing particular strength on SDG 5 (Gender Equality), where it reached 0.96 F1. Llama underperformed relative to the bigger LLMs, with a macro F1 of 0.48, highlighting its weaker alignment with human annotations, again a result that is probably due to the smaller size of the model. These results confirm that modern LLMs can act as strong baselines for SDG classification, but also reveal variability across goals and between models.

5.5.2 Models Trained on AlmaSDG-silver

Supervised models trained on the silver dataset demonstrated competitive performance despite the inherent noise of automatic annotation: Logistic Regression (LR) and Support Vector Classifier (SVC) were the best-performing models among the ones trained on AlmaSDG-silver, each achieving a macro F1-score of 0.57. Their balance between precision and recall made them robust across multiple SDGs. BERT and RoBERTa,

Method	Train/ Tuning	Macro Precision	Macro Recall	Macro F1
Gemini	No	0.54	0.81	0.63
GPT	No	0.78	0.66	**0.69**
Llama	No	0.36	0.87	0.48
Aurora	No	0.63	0.28	0.35
Elsevier S	No	0.60	0.40	0.44
Elsevier M	No	0.54	0.66	0.54
OSDG	No	0.60	0.46	0.49
LR	AlmaSDG-silver	0.58	0.60	0.57
SVC	AlmaSDG-silver	0.61	0.58	0.57
BERT	AlmaSDG-silver	0.37	0.78	0.48
RoBERTa	AlmaSDG-silver	0.28	0.85	0.41
LR	OSDG	0.67	0.53	0.57
SVC	OSDG	0.70	0.55	0.59
BERT	OSDG	0.55	0.53	0.52
RoBERTa	OSDG	0.50	0.51	0.48
LR	SKH	0.64	0.53	0.56
SVC	SKH	0.69	0.53	0.58
BERT	SKH	0.46	0.68	0.51
RoBERTa	SKH	0.48	0.68	0.50

Table 5.4: Experimental evaluation as macro-averaged precision, recall, and F1 score.

while benefiting from transformer-based representations, suffered from lower precision, resulting in macro F1-scores of 0.48 and 0.41, respectively. Notably, LR and SVC showed high performance on SDG 5 and SDG 7, achieving F1-scores above 0.80, while RoBERTa significantly outperformed BERT on SDG 17, where traditional models failed to detect positive samples.

5.5.3 Comparison with Alternative Training Datasets

To contextualize the performance of silver-trained models, we also trained the same classifiers on OSDG and SKH datasets. Models trained on OSDG achieved macro F1-scores up to 0.59 (SVC), slightly surpassing the silver-trained counterparts. Models trained on SKH reached 0.58 with SVC, which is, again, close to AlmaSDG-silver results.

Despite being smaller in size (3k documents compared to OSDG's 14k and SKH's 9k),

SDG	Gemini	GPT	Llama	Aurora	Elsevier S	Elsevier M	OSDG
1	0.43	0.50	0.28	0.50	0.36	0.21	0.32
2	0.67	0.67	0.59	0.27	0.40	0.64	0.44
3	0.64	**0.83**	0.56	0.28	0.28	0.33	0.65
4	0.71	0.75	0.44	0.43	0.67	0.78	0.75
5	**0.96**	0.63	0.80	0.67	0.60	0.93	0.87
6	0.73	**0.88**	0.67	0.67	0.71	0.67	0.70
7	0.74	0.78	0.50	0.40	0.40	0.47	0.59
8	**0.51**	0.42	0.29	0.30	0.46	0.41	0.29
9	0.55	**0.61**	0.37	0.00	0.00	0.41	0.12
10	**0.63**	0.58	0.56	0.36	0.27	0.54	0.31
11	0.69	**0.71**	0.64	0.20	0.47	0.59	0.42
12	0.78	**0.84**	0.73	0.21	0.51	0.56	0.40
13	0.58	0.74	0.45	0.26	0.50	0.60	**0.75**
14	0.82	**0.86**	0.56	0.33	0.50	0.67	0.46
15	0.46	0.80	0.26	0.62	**0.82**	0.89	0.67
16	0.38	0.44	0.25	0.40	0.48	0.55	0.55
17	0.44	**0.60**	0.16	0.00	0.00	0.00	0.00

Table 5.5: Experimental evaluation (F1) – Zero-shot models, no tuning.

AlmaSDG-silver delivered comparable performance, demonstrating the effectiveness of our pipeline for generating usable training data.

5.5.4 Key Insights

Looking at the results, a few observations stand out. To begin with, large language models, and in particular GPT, confirmed to be a very strong baseline. Its performance on the gold dataset was consistently higher than that of the other approaches, showing that zero-shot classification with state-of-the-art LLMs can already be quite effective.

At the same time, the models we trained on AlmaSDG-silver proved surprisingly competitive. Even though the silver dataset is much smaller than other human-annotated resources such as OSDG or SKH, classifiers like Logistic Regression and SVC achieved results that were similar. This suggests that automatically annotated data, if produced carefully and grounded in expert-designed guidelines, can provide real value for downstream training.

A closer look at the results on the single SDGs also helps to explain some of the

SDG	LR	SVC	BERT	RoBERTa
1	0.35	0.38	0.23	0.16
2	0.69	0.69	0.50	0.40
3	0.76	0.77	0.77	0.76
4	0.63	0.63	0.53	0.34
5	0.82	0.82	0.62	0.54
6	0.71	0.71	0.36	0.30
7	0.83	0.83	0.49	0.40
8	0.37	0.40	0.27	0.27
9	0.41	0.39	0.50	0.44
10	0.62	0.55	0.63	0.57
11	0.43	0.48	0.66	0.47
12	0.72	0.72	0.64	0.57
13	0.57	0.61	0.46	0.38
14	0.67	0.55	0.53	0.31
15	0.60	0.60	0.27	0.26
16	0.57	0.59	0.52	0.38
17	0.00	0.00	0.19	0.41

Table 5.6: Experimental evaluation (F1) – Training/Tuning on **AlmaSDG-silver**.

strengths and weaknesses of the models. For goals such as SDG 5 (Gender Equality), the performance was consistently high across methods, this is most likely due to the fact that the textual signals are more explicit and the boundaries of the category are clearer. In contrast, more complex goals like SDG 8 (Decent Work and Economic Growth) and SDG 16 (Peace, Justice and Strong Institutions) remained challenging, with both humans and models struggling to agree.

Taken together, these findings support the central idea of our pipeline. Even if silver data is inevitably noisy, it can be used to train lightweight classifiers that perform at a level close to much more resource-intensive approaches. This makes the pipeline not only cost-effective but also adaptable to other domains where labeled data is scarce.

5.6 Discussion

The results presented in the previous section give us a clearer idea of what this pipeline can and cannot do. On the positive side, the experiments show that, for known domains such as the SDGs, it is possible to move beyond the traditional barrier of data scarcity. With only

SDG	LR	SVC	BERT	RoBERTa
1	0.55	**0.67**	0.39	0.35
2	0.70	0.75	0.61	0.50
3	0.65	0.65	0.79	0.80
4	**0.78**	**0.78**	0.71	0.70
5	0.87	0.87	0.88	0.85
6	0.59	0.57	0.50	0.26
7	0.82	0.82	0.70	0.70
8	0.38	0.42	0.22	0.07
9	0.00	0.11	0.16	0.25
10	0.33	0.34	0.34	0.37
11	0.67	0.70	0.50	0.28
12	0.61	0.61	0.66	0.72
13	0.67	0.71	0.58	0.55
14	0.67	0.62	0.72	0.77
15	0.70	0.70	0.48	0.49
16	0.76	**0.78**	0.53	0.53
17	0.00	0.00	0.00	0.00

Table 5.7: Experimental evaluation (F1) – Taining/Tuning on **OSDG**.

a limited investment of expert time, we managed to build a gold dataset that, while small, was sufficient to test and calibrate the system. From there, the silver dataset provided the necessary scale to train classifiers, and the performance of those models was far from negligible. In fact, in several cases they came close to, or even matched, systems trained on much larger human-annotated corpora. This is an encouraging result: it suggests that even imperfect data, if produced carefully, can be a valuable resource.

Of course, there are also limitations that cannot be ignored. Automatically generated labels inevitably introduce noise, and the quality of the silver dataset depends heavily on the alignment between human experts and the chosen LLM. In our experiments GPT proved to be the most reliable annotator, but this may not hold in different contexts or with future models. For this reason, the pipeline should be thought of as flexible and iterative rather than fixed. Guidelines can be refined, the choice of LLM revisited, and the silver dataset expanded or rebalanced over time.

Another point that clearly emerges from the experiments is that not all SDGs are equally easy to detect. Some of them, such as gender equality or health, tend to be more straightforward: the vocabulary is recognisable, and the link between the text and the

SDG	LR	SVC	BERT	RoBERTa
1	0.25	0.44	0.22	0.24
2	0.67	0.67	0.67	0.67
3	0.59	0.54	0.66	0.76
4	0.50	0.50	0.62	0.54
5	0.83	0.87	0.64	0.74
6	0.71	0.71	0.36	0.75
7	**0.86**	**0.86**	0.65	0.54
8	0.27	0.27	0.26	0.39
9	0.38	0.48	0.40	0.48
10	0.42	0.33	0.45	0.55
11	0.40	0.41	0.60	0.27
12	0.74	0.70	0.65	0.69
13	0.65	0.71	0.56	0.62
14	0.57	0.67	0.72	0.36
15	0.64	0.64	0.37	0.36
16	**0.78**	0.74	0.67	0.37
17	0.27	0.27	0.20	0.14

Table 5.8: Experimental evaluation (F1) – Taining/Tuning on **SKH**.

goal is usually explicit. Others, like decent work or strong institutions, are much harder to capture. Their scope is broader, the language more nuanced, and in many cases the contribution of an article is only indirect or implicit. This uneven difficulty was already visible during the gold annotation process, where experts themselves sometimes struggled to agree, and it is no surprise that the models reflect the same challenge. What this tells us is that the problem is not only technical but also conceptual: the way the goals are defined makes some of them inherently more ambiguous than others.

The comparison with direct LLM classification is also informative. GPT, used in a zero-shot setting, delivered excellent results and in many cases outperformed the models trained on silver data. Nevertheless, there are good reasons not to rely on LLMs alone. Running them at scale is expensive, and they remain dependent on external providers. By contrast, the smaller models trained on silver data are lighter, more cost-effective to deploy, and can be fine-tuned to the specific needs of an institution. In practice, the two approaches should be seen as complementary. LLMs are invaluable for producing large quantities of labeled data quickly, while traditional classifiers offer a sustainable way of turning that data into models that are transparent, efficient, and easier to control.

5.7 Summary and Reflections

In this chapter we outlined a methodology for tackling text classification tasks in situations where labeled data is scarce. The pipeline builds on a simple idea: combine a small, carefully constructed gold dataset with a much larger silver dataset produced automatically through large language models. The gold annotations ensure reliability and provide a benchmark, while the silver annotations offer the scale needed to train supervised models.

Applied to the classification of scientific publications according to the UN Sustainable Development Goals, the approach showed that even with limited human effort it is possible to construct resources that support meaningful experimentation. The gold dataset, though modest in size, was sufficient to evaluate both LLMs and traditional classifiers. The silver dataset, generated by prompting GPT with expert guidelines, supplied the volume required for training. Together, these resources allowed us to compare direct zero-shot classification with the performance of smaller, domain-specific models trained on silver data.

The results confirmed that zero-shot LLMs are very strong baselines, but they also demonstrated the value of the silver-based approach. Logistic Regression and SVC, trained only on automatically labeled data, achieved results close to those of models trained on much larger human-annotated datasets. This suggests that the proposed pipeline can lower the barrier to developing effective classifiers in domains where large-scale manual labeling is not realistic.

To conclude, the methodology is not meant as a definitive solution but as a practical compromise. It balances the reliability of expert input with the scalability of machine annotation, and it offers a framework that can be adapted, extended, and improved in future work. The next chapter will build on this foundation by validating the pipeline on long and unstructured legal text, assessing how well it performs outside the controlled setting of our experiments.

Chapter 6

Classification of Legal Texts

The previous chapter introduced a pipeline designed to overcome the lack of labeled data in text classification. The methodology combined a small gold dataset, annotated by experts, with a larger silver dataset produced through large language models, and then used these resources to train lightweight classifiers. While the experiments in Chapter 5 focused on scientific publications, the real test for the pipeline lies in its application to more complex and less controlled scenarios.

This chapter presents two such validation cases. The first is the classification of administrative acts of Regione Emilia-Romagna. These documents are long, highly unstructured, and full of legal-administrative language. At the start of this project, they were also completely unlabeled, which made them an ideal but very challenging ground to test whether the proposed methodology could really help where traditional supervised approaches are not feasible.

Alongside this central case, a second, complementary dataset was considered: Bulgarian family case law. Unlike the administrative acts, these documents are shorter and somewhat more structured, but they also present some of the same difficulties, such as the need to correctly interpret references to laws and legal articles. In this case we were able to rely on a "dictionary" of legal references, which was integrated into the prompts given to LLMs in order to support disambiguation.

A further point of interest in both experiments is language. Whereas the experiments described in Chapter 5 were carried out entirely on English texts, here the material is in Italian and Bulgarian. This makes it possible to assess not only the robustness of the pipeline, but also the ability of LLMs to work effectively with languages other than English, a question that had already emerged in Chapter 4, where we tested LLMs on multilingual

medical data.

6.1 Case Study 1: Emilia-Romagna Administrative Acts

The Emilia-Romagna administrative acts, already introduced in Chapter 3, represent the core case study of this thesis. These documents pose several challenges that make them a particularly demanding benchmark for testing the pipeline. On average, each act is almost three thousand words long, which means that any automatic classification method must handle not only the volume of text but also its unstructured nature. Moreover, the acts are often filled with references to laws and legal articles that are mentioned without explanation. For a model, this creates considerable "noise": the truly relevant information for classification is scattered among passages that are formulaic or legally mandated but not discriminative for the task.

A further complication comes from the taxonomy of classes. Although the classification problem is single-label, some categories may overlap, and a document might touch on themes that could plausibly place it in more than one class. This structural ambiguity adds another layer of difficulty, as the model must commit to a single label even when the boundaries between categories are blurred.

The gold dataset used for evaluation consists of 264 documents. As noted earlier, the distribution of classes is imbalanced. In particular, certain classes such as Class 8 are under-represented, and despite repeated searches it proved impossible to identify additional examples. This means that the gold set, while sufficient for evaluation, also reflects the real-world constraints of working with institutional data: some categories are simply rare, and the classifier must be able to deal with this imbalance.

With this background in mind, the focus of the present section is on the experiments carried out with large language models. The goal was to test different LLMs and prompting strategies against the gold dataset in order to identify the combination that offered the best balance of accuracy and consistency. Once the most reliable setup was found, it was applied to the larger unlabeled collection to create the silver dataset, which then served as training material for lightweight classifiers.

6.1.1 Experiments with LLMs

The first step in applying the pipeline to the Emilia-Romagna administrative acts was to test how different prompting strategies would perform on the gold dataset. Given the

length and complexity of the documents, this stage was particularly important: it allowed us to measure how sensitive the models were to different formulations of the task, and it provided the basis for selecting the approach to be used in the construction of the silver dataset.

Four prompting strategies were evaluated using GPT 4o, they are summarized below and better shown in the Appendix 8.1, were we also report all the confusion matrices of the experiments.

1. *One Call*: a single prompt containing the description of all classes, together with one example of input and one of output. The model was asked to assign the document to one of the classes.

2. *Multiple Calls*: instead of giving the full taxonomy, the prompt described a single class, and the model was asked whether the document did or did not belong to that class through a correlation score of the document to the class. This meant having 19 different API calls for every document, and then comparing the correlation scores to define which label to assign.

3. *Two Steps*: designed to deal with cases of overlap; in the first call, the model was shown all class descriptions and asked to select those that seemed most relevant. Then, in a second call, only the selected classes were described again, and the model had to choose exactly one among them. The idea was to reduce the number of candidate classes so that the model could focus more effectively on ambiguous cases.

4. *Second Chance*: also uses a two-step process but with a different emphasis: the model was asked to classify the document as in the first strategy, and then in a second call it was reminded of its initial choice and explicitly asked whether it was confident about it or wanted to change its answer.

The results of these experiments are summarized in Table 6.1 and detailed in Tables 6.2, 6.3, 6.4, 6.5, and 6.6. All strategies were tested using GPT-4o. For strategy 1, two variants were tested: one with the full text of the document and one where the portion of text after the keyword "DETERMINA" was excluded. As mentioned in Chapter 3, the text after this keyword contains the decision of the judge, but does not mention the context of this decision, so the exclusion was motivated by the observation that this part of the acts often contains lengthy references to legislative articles and boilerplate formulations, which tend to obscure rather than clarify the relevant information. For the other strategies, only

the variant without the text after the keyword was tested, given the initial evidence that removing this section improved results.

The results show a number of consistent patterns across prompts. With One Call strategy with the whole text of the document (Table 6.2), performance was limited, with the model often distracted by the large amount of legal references and failing to capture the essential part of the acts. Certain classes, such as 4 and 7, were handled reasonably well, but many minority classes were either missed or confused with larger categories.

The variant of the first prompting strategy without the text after the keyword "DETERMINA" produced a clear improvement (Table 6.3). By removing the repetitive and noisy tail of the documents, the model was able to focus more effectively on the sections containing the substantive decisions. Here, classes such as 1, 4, 7, and 15 were classified with good precision and recall, and even some of the less frequent categories, such as 9 and 19, showed reasonable performance. Nevertheless, the model still struggled with very underrepresented classes, notably class 8, which only had a handful of examples. Excluding the text after "DETERMINA" proved consistently beneficial, and for this reason all subsequent experiments were carried out with this variant as the baseline.

The Multiple Calls strategy achieved mixed results (Table 6.4). While in theory breaking the problem into 19 binary decisions could reduce confusion, in practice the approach amplified the problem of false positives. The model tended to over-associate documents with multiple classes, and the final selection based on correlation scores did not always yield coherent results. Some classes, such as 2 and 3, benefited slightly from the focused attention, but overall the gains were not sufficient to compensate for the complexity of the procedure.

The Two Steps strategy was the most accurate in absolute terms, correctly classifying 173 documents out of 264 in the gold set (Table 6.5). It was particularly effective in resolving overlaps, since the narrowing down in the first call forced the model to concentrate on the most plausible candidates. For example, it performed well on classes that often overlap, such as 14 and 15, where the second step helped disambiguate borderline cases. However, this gain came at a very high cost: each document required two calls, with an average of more than 6,000 tokens per call. Given the size of the dataset, this strategy proved expensive both in terms of time and money, making it impractical for large-scale annotation.

The Second Chance strategy, by contrast, did not produce significant improvements over the basic One Call strategy (Table 6.6). Asking the model to confirm or revise its initial choice rarely led to a different answer, and when it did, the revision was not systematically

better. The performance therefore remained close to that of One Call without the text after the keyword "DETERMINA", but with the overhead of a second call.

Prompting Strategy	**Correctly Labeled**	**Macro F1-score**
1: Single Call	171	0.67
2: Multiple Calls	145	0.53
3: Two Steps	173	0.63
4: Second Chance	163	0.60

Table 6.1: Classification results of different prompting strategies on the gold dataset using GPT 4o. The dataset contains 264 documents.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.35	0.70	0.47	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.44	0.42	0.43	19
3 - Risorse Umane	0.65	0.52	0.58	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.43	0.87	0.58	15
5 - Sistema Informativo	0.17	0.11	0.13	9
6 - Programmazione, Coordinamento e Controllo	0.67	0.44	0.53	41
7 - Agricoltura e Allevamento	0.67	0.89	0.77	35
8 - Artigianato	1	0.67	0.80	3
9 - Commercio	1	0.80	0.89	5
10 - Industria e Attività Estrattive	0.8	0.67	0.73	6
11 - Turismo e Strutture Ricettive	1	0.25	0.40	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	1	0.67	0.80	6
13 - Infrastrutture e Trasporti	0.67	0.75	0.71	8
14 - Tutela dell'Ambiente	0.57	0.80	0.67	15
15 Sanità, Igiene e Veterinaria	0.76	0.79	0.77	28
16 - Politiche Sociali	0.38	0.43	0.40	7
17 Istruzione, Formazione e Lavoro	0.71	0.45	0.56	11
18 - Beni e Attività Culturali	0.90	0.60	0.72	15
19 - Attività Sportive e Ricreative	0	0	0	6
accuracy			0.61	264
macro avg	0.64	0.57	0.57	264
weighted avg	0.63	0.61	0.60	264

Table 6.2: Classification report of prompting technique 1: Single Call using GPT 4o, with all the text of the documents.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.41	0.70	0.52	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	1	0.37	0.54	19
3 - Risorse Umane	0.60	0.43	0.50	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.34	1	0.51	15
5 - Sistema Informativo	0.29	0.22	0.25	9
6 - Programmazione, Coordinamento e Controllo	0.66	0.51	0.58	41
7 - Agricoltura e Allevamento	0.90	0.74	0.81	35
8 - Artigianato	1	0.67	0.80	3
9 - Commercio	1	1	1	5
10 - Industria e Attività Estrattive	0.75	0.50	0.60	6
11 - Turismo e Strutture Ricettive	1	0.50	0.67	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	1	0.83	0.91	6
13 - Infrastrutture e Trasporti	0.89	1	0.94	8
14 - Tutela dell'Ambiente	0.75	0.80	0.77	15
15 Sanità, Igiene e Veterinaria	0.75	0.86	0.80	28
16 - Politiche Sociali	0.38	0.43	0.40	7
17 Istruzione, Formazione e Lavoro	0.38	0.45	0.42	11
18 - Beni e Attività Culturali	0.82	0.60	0.69	15
19 - Attività Sportive e Ricreative	1	1	1	6
accuracy			0.65	264
macro avg	0.73	0.66	0.67	264
weighted avg	0.71	0.65	0.65	264

Table 6.3: Classification report of prompting technique 1: Single Call using GPT 4o, without the text after the keyword "DETERMINA".

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.26	0.70	0.38	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.62	0.26	0.37	19
3 - Risorse Umane	0.56	0.43	0.49	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.30	0.87	0.44	15
5 - Sistema Informativo	0.50	0.22	0.31	9
6 - Programmazione, Coordinamento e Controllo	0.65	0.37	0.47	41
7 - Agricoltura e Allevamento	0.67	0.83	0.74	35
8 - Artigianato	1	1	1	3
9 - Commercio	0.60	0.60	0.60	5
10 - Industria e Attività Estrattive	0.33	0.17	0.22	6
11 - Turismo e Strutture Ricettive	1	0.50	0.67	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	1	0.50	0.67	6
13 - Infrastrutture e Trasporti	0.67	0.75	0.71	8
14 - Tutela dell'Ambiente	0.65	0.73	0.69	15
15 Sanità, Igiene e Veterinaria	0.71	0.61	0.65	28
16 - Politiche Sociali	0.38	0.43	0.40	7
17 Istruzione, Formazione e Lavoro	0.47	0.64	0.54	11
18 - Beni e Attività Culturali	0.90	0.60	0.72	15
19 - Attività Sportive e Ricreative	0	0	0	6
accuracy			0.55	264
macro avg	0.59	0.54	0.53	264
weighted avg	0.60	0.55	0.54	264

Table 6.4: Classification report of prompting technique 2: Multiple Calls using GPT 4o.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.39	0.70	0.50	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.80	0.42	0.55	19
3 - Risorse Umane	0.65	0.62	0.63	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.43	1	0.60	15
5 - Sistema Informativo	0.33	0.33	0.33	9
6 - Programmazione, Coordinamento e Controllo	0.81	0.41	0.55	41
7 - Agricoltura e Allevamento	0.73	0.91	0.81	35
8 - Artigianato	1	1	1	3
9 - Commercio	1	0.40	0.57	5
10 - Industria e Attività Estrattive	0.80	0.67	0.73	6
11 - Turismo e Strutture Ricettive	1	0.50	0.67	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	0.83	0.83	0.83	6
13 - Infrastrutture e Trasporti	0.70	0.88	0.78	8
14 - Tutela dell'Ambiente	0.63	0.80	0.71	15
15 Sanità, Igiene e Veterinaria	0.81	0.89	0.85	28
16 - Politiche Sociali	0.38	0.43	0.40	7
17 Istruzione, Formazione e Lavoro	0.56	0.45	0.50	11
18 - Beni e Attività Culturali	0.82	0.60	0.69	15
19 - Attività Sportive e Ricreative	1	0.17	0.29	6
accuracy			0.66	264
macro avg	0.72	0.63	0.63	264
weighted avg	0.71	0.66	0.65	264

Table 6.5: Classification report of prompting technique 3: Two Steps using GPT 4o.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.37	0.70	0.48	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.50	0.32	0.39	19
3 - Risorse Umane	0.62	0.48	0.54	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.42	0.93	0.58	15
5 - Sistema Informativo	0.33	0.22	0.27	9
6 - Programmazione, Coordinamento e Controllo	0.58	0.46	0.51	41
7 - Agricoltura e Allevamento	0.70	0.94	0.80	35
8 - Artigianato	1	1	1	3
9 - Commercio	1	0.80	0.89	5
10 - Industria e Attività Estrattive	1	0.67	0.80	6
11 - Turismo e Strutture Ricettive	1	0.5	0.67	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	0.62	0.83	0.71	6
13 - Infrastrutture e Trasporti	0.67	0.55	0.71	8
14 - Tutela dell'Ambiente	0.67	0.67	0.67	15
15 Sanità, Igiene e Veterinaria	0.77	0.82	0.79	28
16 - Politiche Sociali	0.29	0.29	0.29	7
17 Istruzione, Formazione e Lavoro	0.71	0.45	0.56	11
18 - Beni e Attività Culturali	0.89	0.53	0.67	15
19 - Attività Sportive e Ricreative	0	0	0	6
accuracy			0.62	264
macro avg	0.64	0.60	0.60	264
weighted avg	0.63	0.62	0.60	264

Table 6.6: Classification report of prompting technique 4: Second Chance using GPT 4o.

Taken together, these findings suggest that the first strategy strikes the best balance between accuracy and efficiency. Although the "Two Step" method technically gave the highest number of correct classifications, the marginal advantage of two extra correct predictions over the gold set does not justify doubling the computational cost when scaling up to thousands of documents. Given that the ultimate goal was to annotate over 1,400 administrative acts for the silver dataset, this trade-off was decisive.

Once the best-performing prompting strategy had been identified with GPT, we tested the same setup on two other models: Gemini and Llama[1]. These models could be used consistently by public administrations because they are both free, and Llama would be even better as, being smaller than the other two, it can run locally on a standard personal computer. In both cases, though, the results were noticeably worse, as visible in Tables 6.7 and 6.8 respectively. The results were significantly weaker than those obtained with GPT-4o. Both models struggled with the longer documents and showed lower precision

[1]The versions of Gemini and Llama were "gemini-1.5-flash-latest" and "Meta-Llama-3-8B-Instruct" respectively.

and recall across most classes. Minority classes were almost entirely missed, and even larger categories such as 7 and 15, which GPT-4o handled well, suffered from reduced accuracy. In terms of macro F1-score, Gemini performed better than Llama, which is a considerably smaller model, but neither approached the baseline established by GPT-4o. For this reason, we did not test the other prompting strategies on these two models.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.22	0.60	0.32	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.67	0.21	0.32	19
3 - Risorse Umane	0.54	0.62	0.58	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.14	0.87	0.24	15
5 - Sistema Informativo	0.33	0.11	0.17	9
6 - Programmazione, Coordinamento e Controllo	0.39	0.17	0.24	41
7 - Agricoltura e Allevamento	0.75	0.51	0.61	35
8 - Artigianato	0	0	0	3
9 - Commercio	1	0.60	0.75	5
10 - Industria e Attività Estrattive	1	0.50	0.67	6
11 - Turismo e Strutture Ricettive	1	0.50	0.67	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	1	0.33	0.50	6
13 - Infrastrutture e Trasporti	1	0.62	0.77	8
14 - Tutela dell'Ambiente	0.67	0.80	0.73	15
15 Sanità, Igiene e Veterinaria	0.73	0.39	0.51	28
16 - Politiche Sociali	0.17	0.14	0.15	7
17 Istruzione, Formazione e Lavoro	0.80	0.36	0.50	11
18 - Beni e Attività Culturali	1	0.53	0.70	15
19 - Attività Sportive e Ricreative	0	0	0	6
accuracy			0.43	264
macro avg	0.60	0.41	0.44	264
weighted avg	0.60	0.43	0.45	264

Table 6.7: Classification report of prompting technique 1: One call using gemini-1.5-flash-latest.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0	0	0	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.22	0.26	0.24	19
3 - Risorse Umane	0.22	0.10	0.13	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.16	0.53	0.24	15
5 - Sistema Informativo	0	0	0	9
6 - Programmazione, Coordinamento e Controllo	0.36	0.12	0.18	41
7 - Agricoltura e Allevamento	0.19	0.14	0.16	35
8 - Artigianato	0	0	0	3
9 - Commercio	0.05	0.60	0.10	5
10 - Industria e Attività Estrattive	0	0	0	6
11 - Turismo e Strutture Ricettive	0	0	0	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	0.11	0.17	0.13	6
13 - Infrastrutture e Trasporti	0.75	0.75	0.75	8
14 - Tutela dell'Ambiente	0.45	0.60	0.51	15
15 Sanità, Igiene e Veterinaria	0.75	0.11	0.19	28
16 - Politiche Sociali	0	0	0	7
17 Istruzione, Formazione e Lavoro	0	0	0	11
18 - Beni e Attività Culturali	0.67	0.13	0.22	15
19 - Attività Sportive e Ricreative	0	0	0	6
accuracy			0.19	264
macro avg	0.21	0.18	0.15	264
weighted avg	0.29	0.19	0.18	264

Table 6.8: Classification report of prompting technique 1: One call using Meta-Llama-3-8B-Instruct.

In summary, the experiments with LLMs confirmed both the potential and the limitations of the approach. Prompt design had a clear impact on the quality of the results, with relatively simple adjustments—such as removing irrelevant boilerplate text—leading to tangible improvements. At the same time, strategies that required multiple calls, while theoretically appealing, were not sustainable in practice for large-scale annotation. GPT, with the "1: One Call" prompt, therefore emerged as the most reliable and efficient option for the creation of the silver dataset.

6.1.2 Silver Dataset Creation

After identifying the most effective prompting strategy in the previous experiments, the next step was to generate the silver dataset. For this purpose, we used GPT-4o with the "1: One Call" approach, which had shown the best balance between accuracy and efficiency. This choice was motivated by the fact that, while slightly more complex

strategies could in principle recover a few additional correct classifications, their cost in time and computational resources made them impractical for annotating a large corpus.

The silver dataset was created from 1,405 administrative acts drawn from the same institutional source as the gold set. The documents share the same characteristics as those used in evaluation: they are long, averaging nearly 3,000 words each, and often contain boilerplate legal references that make them noisy and difficult to process. Importantly, there is no overlap between the documents in the gold and silver datasets, ensuring that evaluation remains independent of training data.

The resulting dataset reflects the uneven distribution of classes already observed in the gold set. Some categories are well represented, while others are rare. For example, Class 7 dominates with 422 documents, followed by Class 17 with 218, whereas Class 8 has only 4 examples despite repeated attempts to identify more. A full overview of the class distribution is provided in Table 6.9. This imbalance mirrors the realities of administrative practice: certain types of acts are produced frequently, while others are much less common.

In summary, the silver dataset extends the coverage of the gold set from a few hundred to over a thousand annotated documents, providing the scale necessary for supervised training. At the same time, it preserves the real-world complexity of the material, including long text length, noisy structure, and class imbalance. Together with the gold dataset, it forms the basis for testing whether lightweight classifiers can be trained to match or surpass the performance of direct LLM classification.

6.1.3 Training Lightweight Models

Once the silver dataset was available, the next step was to train lightweight classifiers on it and test their performance against the gold set. For this purpose, we experimented with two models that had already shown solid results in the experiments on scientific publications: Logistic Regression and Support Vector Classifier (SVC). Both models were trained on the 1,405 silver-labeled acts and then evaluated on the 264 gold documents, with performance measured in terms of macro F1-score.

The results are summarised in Table 6.10, and detailed in Tables 6.11 and 6.12. Logistic Regression reached a macro F1 of 0.38, while SVC achieved 0.55. As expected, both models performed below GPT-4o, which had been used directly in zero-shot classification. However, the comparison with other LLMs tells a different story. SVC in particular not only outperformed Logistic Regression, but also surpassed Gemini, and both lightweight models did better than Llama. This is significant, because it shows that even relatively

Class Name	N. of documents
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	47
2 - Organizzazione, Patrimonio e Risorse Strumentali	110
3 - Risorse Umane	26
4 - Risorse Finanziare, Gestione Contabile e Fiscale	45
5 - Sistema Informativo	28
6 - Programmazione, Coordinamento e Controllo	86
7 - Agricoltura e Allevamento	422
8 - Artigianato	4
9 - Commercio	19
10 - Industria e Attività Estrattive	29
11 - Turismo e Strutture Ricettive	21
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	37
13 - Infrastrutture e Trasporti	32
14 - Tutela dell'Ambiente	82
15 Sanità, Igiene e Veterinaria	45
16 - Politiche Sociali	50
17 Istruzione, Formazione e Lavoro	218
18 - Beni e Attività Culturali	53
19 - Attività Sportive e Ricreative	51

Table 6.9: Class distribution of automatically labeled documents in the silver dataset.

simple classifiers trained on silver data can offer an effective alternative to free or open-source LLMs.

Model	Precision	Recall	Macro F-1 Score
LR	0.42	0.40	0.38
SVC	0.64	0.55	0.55

Table 6.10: Macro avg performance of lightweight models trained on the silver dataset and tested on the gold dataset.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.35	0.70	0.47	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.22	0.21	0.22	19
3 - Risorse Umane	0.45	0.24	0.31	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.58	0.93	0.72	15
5 - Sistema Informativo	0.29	0.22	0.25	9
6 - Programmazione, Coordinamento e Controllo	0.50	0.07	0.13	41
7 - Agricoltura e Allevamento	0.68	0.77	0.72	35
8 - Artigianato	0	0	0	3
9 - Commercio	1	0.40	0.57	5
10 - Industria e Attività Estrattive	0	0	0	6
11 - Turismo e Strutture Ricettive	0	0	0	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	0.43	0.50	0.46	6
13 - Infrastrutture e Trasporti	0.45	0.62	0.53	8
14 - Tutela dell'Ambiente	0.39	0.47	0.42	15
15 Sanità, Igiene e Veterinaria	0.68	0.75	0.71	28
16 - Politiche Sociali	0	0	0	7
17 Istruzione, Formazione e Lavoro	0.27	0.55	0.36	11
18 - Beni e Attività Culturali	0.62	0.33	0.43	15
19 - Attività Sportive e Ricreative	1	0.83	0.91	6
accuracy			0.44	264
macro avg	0.42	0.40	0.38	264
weighted avg	0.48	0.44	0.42	264

Table 6.11: Classification report of Logistic Regression trained on the silver dataset and tested on the gold dataset.

	Precision	Recall	F1-Score	Support
1 - Organi di Governo, Affari Giuridico-Istituzionali e Comunicazione	0.50	0.60	0.55	10
2 - Organizzazione, Patrimonio e Risorse Strumentali	0.33	0.21	0.26	19
3 - Risorse Umane	0.61	0.52	0.56	21
4 - Risorse Finanziare, Gestione Contabile e Fiscale	0.31	0.93	0.47	15
5 - Sistema Informativo	0.29	0.22	0.25	9
6 - Programmazione, Coordinamento e Controllo	0.75	0.15	0.24	41
7 - Agricoltura e Allevamento	0.60	0.89	0.71	35
8 - Artigianato	1	0.33	0.50	3
9 - Commercio	1	0.80	0.89	5
10 - Industria e Attività Estrattive	0.50	0.50	0.50	6
11 - Turismo e Strutture Ricettive	1	0.25	0.40	4
12 - Pianificazione Territoriale, Urbanistica ed Edilizia	0.80	0.67	0.73	6
13 - Infrastrutture e Trasporti	0.55	0.75	0.63	8
14 - Tutela dell'Ambiente	0.71	0.80	0.75	15
15 Sanità, Igiene e Veterinaria	0.71	0.89	0.79	28
16 - Politiche Sociali	0	0	0	7
17 Istruzione, Formazione e Lavoro	0.75	0.55	0.63	11
18 - Beni e Attività Culturali	0.73	0.53	0.62	15
19 - Attività Sportive e Ricreative	1	0.83	0.91	6
accuracy			0.56	264
macro avg	0.64	0.55	0.55	264
weighted avg	0.62	0.56	0.54	264

Table 6.12: Classification report of SVC trained on the silver dataset and tested on the gold dataset.

Looking more closely at the per-class scores (the confusion matrices are reported in the appendix 8.1), the same patterns emerge that we had already seen during the gold annotation and in the LLM experiments. Some classes, like 7 and 15, are relatively easy to recognise, and both models picked them up consistently. Others are trickier. Logistic Regression in particular struggled with the minority classes, often missing them completely, whereas SVC was a little more robust and managed to capture some of the rarer cases. This difference largely explains why SVC's macro F1 is considerably higher.

Another factor that really matters in practice is time. Running Llama on the gold dataset took an entire day on a standard personal computer, while Gemini, despite using an API, needed several hours. By contrast, training and testing SVC or Logistic Regression is a matter of minutes and can be done on the same standard hardware. For an institution that wants to update or retrain models regularly, this efficiency makes a big difference: once the silver dataset exists, producing a new classifier is almost effortless.

In the end, GPT-4o still gives the best results if one can afford to use it directly, but

SVC shows that the pipeline delivers on its promise. It produces models that are lighter, more cost-effective, and surprisingly competitive, especially when compared with freely available LLMs. This makes the whole approach not just an academic exercise, but something that could actually work in day-to-day settings where resources are limited.

6.1.4 Results and Discussion

The experiments with the Emilia-Romagna acts highlight both the potential and the limits of the proposed pipeline when applied to long and unstructured administrative texts. Among the large language models, GPT-4o was clearly the most reliable. With the "One Call" prompting strategy (excluding the boilerplate text after DETERMINA), it classified 171 out of 264 gold documents correctly. The alternative two-step prompt achieved a slightly higher score, with 173 correct classifications, but the marginal gain of two additional correct documents came at the cost of doubling the number of calls. Considering that the annotation of over 1,400 documents was required for the silver set, this trade-off was decisive: accuracy had to be balanced against both financial cost and processing time.

The comparison with Gemini and Llama reinforces this conclusion. Both were tested with the same "One Call" prompt, yet their results were noticeably worse. Gemini performed somewhat better than Llama, but neither came close to GPT-4o. This suggests that the particular characteristics of these acts—length, legal jargon, embedded references, and structural noise—demand the kind of robustness that GPT-4o currently offers more than its competitors.

Once the silver dataset was created, training lightweight classifiers showed that the approach can indeed deliver usable results. SVC, with a macro F1 of 0.55, did not reach GPT-4o's level but it did outperform Gemini and Llama. Logistic Regression was less effective overall, with a macro F1 of 0.38, yet even this model performed better than Llama. These findings are important because they demonstrate that the silver dataset is rich enough to support supervised training, and that relatively simple models can match or surpass free or open-source LLMs when evaluated on the gold set.

Efficiency also matters in practice. Running Llama required almost a full day of computation, and Gemini took several hours. By contrast, SVC can be trained and evaluated in a fraction of the time, even on modest hardware. This efficiency makes the pipeline more appealing for institutions that need to process large volumes of documents but lack the resources to run advanced LLMs at scale.

Overall, the Emilia-Romagna case shows that the pipeline is both feasible and effective.

GPT-4o remains the strongest option when accuracy is the only priority, but lightweight models trained on silver data offer a viable and much more efficient alternative. The results confirm that, even in a setting as difficult as long administrative acts, the methodology provides a practical way to move from unlabeled text to working classifiers.

6.2 Case Study 2: Bulgarian Supreme Court of Cassation Acts

The second case study is dedicated to Bulgarian family case law, a subject that raises similar challenges to those of Emilia-Romagna administrative acts but within another language and juridical environment. The collection includes judgments issued by the Bulgarian Supreme Court of Cassation, drafted wholly in Bulgarian. The documents are about 1,840 words in length on average, thus the texts are relatively shorter compared to the administrative acts studied earlier but complex enough to require proper interpretation.

As in the previous case, there are numerous cross-references to law provisions and law articles in the documents. However, in this dataset, most of these references could be disambiguated with the help of a dictionary given to us by domain experts, which mapped the referred articles to the corresponding legal texts. This extra resource assisted in reducing ambiguity and ensuring that the classification models could access the contextual information relevant to them without being confused by undefined cross-references.

The documents in this experiment span across 14 classes, each of which corresponds to a specific area or topic under family law. Unlike in the previous experiments, where class definitions were given in terms of text descriptions, in this case, the classes were defined in terms of keywords rather than full sentences. This made the task slightly different: the model was required to infer the connection of the material of a decision and a set of rather broad terms, rather than compare the text with detailed definitions.

Overall, the case study provided the opportunity to experiment with the resilience of the proposed pipeline in a new setting—one that is still in the legal domain, but in a more difficult language, and is qualitatively distinct from the Emilia-Romagna regional administrative acts. It also provided us with the opportunity to investigate the effects of the availability of a disambiguation dictionary and the unavailability of narrative class descriptions on LLMs' and lightweight classifiers' behavior.

6.2.1 Gold Dataset Creation

The creation of the gold dataset for the Bulgarian family case law followed the same general approach used in the previous experiments, but several issues appeared early in the process. The goal was to assemble a small, reliable benchmark that could later be used to test both the large language models and the lighter classifiers trained on automatically labeled data.

Finding representative examples for each class turned out to be more difficult than expected. The dataset was divided into 14 categories, but in practice some of these categories were hard to populate. For classes 2 and 8, despite repeated searches and cross-checks with domain experts, we could not identify any documents that clearly fit their definitions. The material simply did not provide examples that could be labeled with enough confidence.

Even among the other classes, the number of documents was often small. After several iterations, the final gold dataset included 123 documents, distributed as shown in Table 6.13.

Class	N. of documents
1 - Getting married	5
2 - Destruction of a marriage	0
3 - Divorce by mutual consent	15
4 - Divorce due to marital breakdown	15
5 - Personal relations between spouses	6
6 - Parent-child relationships	15
7 - Property relations between spouses – legal regime of community of property	15
8 - Property relations between spouses – legal separation regime	0
9 - Property relations between spouses – negotiated regime	2
10 - Origin	15
11 - Adoption	15
12 - International adoption	1
13 - Guardianship and custody	4
14 - Maintenance	15

Table 6.13: Class distribution of tagged documents.

The limited number of documents in some categories was not due to a lack of effort in collection, but rather to restrictions on data accessibility. Unlike the administrative acts of Regione Emilia-Romagna, publicly available for inspection, most of the family law issues addressed by the Bulgarian Supreme Court are concerned with personal data. Thus, only a portion of the material is available for the purposes of research, and even such texts are often anonymized or partly blacked out. This naturally constrained the amount of usable data and made it impossible to balance the classes more evenly. Still, the resulting gold dataset provided a sufficiently diverse sample to evaluate model behavior and to guide the creation of the silver dataset in the next phase of the study.

6.2.2 Experiments with LLMs

During the LLM experimentation phase on the Bulgarian family case law, we were no longer able to continue with GPT, due to the high cost of the API. Although the very first, small pilot with GPT had been encouraging (30 documents correctly classified out of 32),

logistical constraints during a short research period abroad meant that the silver dataset had to be produced with Gemini[2]. Llama was not tested in this phase: earlier experiments showed it to be weaker than the other bigger LLMs on similar tasks, and the time available did not allow re-running a full comparison.

The performance of Gemini on the gold dataset (N = 123, after dropping classes 2 and 8 which had no examples) is summarized in the classification report provided in Table 6.14. Overall accuracy is 0.66, with a macro-average F1 of 0.56 and a weighted F1 of 0.63. These numbers tell a mixed but informative story: Gemini handles many of the frequent and well-defined classes reasonably well, yet it fails systematically on a few categories — in particular, class 5 (six examples) and class 12 (one example), both of which receive F1 = 0.00 because the model never predicts them.

	Precision	Recall	F1-Score	Support
1 - Getting married	1	0.40	0.57	5
3 - Divorce by mutual consent	0.83	0.33	0.48	15
4 - Divorce due to marital breakdown	0.54	0.87	0.67	15
5 - Personal relations between spouses	0	0	0	6
6 - Parent-child relationships	0.56	0.93	0.70	15
7 - Property relations between spouses – legal regime of community of property	0.61	0.73	0.67	15
9 - Property relations between spouses – negotiated regime	0.50	0.50	0.50	2
10 - Origin	0.80	0.80	0.80	15
11 - Adoption	0.86	0.80	0.83	15
12 - International adoption	0	0	0	1
13 - Guardianship and custody	1	1	1	4
14 - Maintenance	0.54	0.47	0.50	15
accuracy			0.66	123
macro avg	0.60	0.57	0.56	123
weighted avg	0.66	0.66	0.63	123

Table 6.14: Classification report of Gemini on the gold dataset of Bulgarian case law.

[2]Gemini-2.0-flash

Looking at the per-class results gives more insights. Some classes show strong performance: class 13 is an interesting case, indicating that when the class is expressed with a distinctive signal Gemini recognizes it reliably. Classes 10 and 11 also perform well (F1 = 0.80 and 0.83 respectively), suggesting that for categories with clear keyword coverage or recurring patterns the model generalizes adequately. Classes 4, 6, 7 show acceptable to good F1s (around 0.67–0.70), with recall generally higher than precision in some of these cases, which implies the model tends to over-predict those labels (i.e., some false positives).

At the other end of the spectrum are classes where Gemini struggles. Class 5 is particularly striking: despite having six gold examples, precision and recall are both zero. Class 12 (support 1) is similarly missed. Intermediate cases include class 1 (precision 1.00, recall 0.40) and class 3 (precision 0.83, recall 0.33): here the model is conservative — when it predicts the label it is almost always correct (high precision), but it fails to find many of the true instances (low recall). Such a pattern often points to thresholds or decision rules in the model's internal scoring: only the clearest examples cross the decision boundary.

Several plausible factors explain these patterns:

- Data scarcity and class definition. Some categories are represented by very few examples (e.g., class 12: 1 example; class 9: 2 examples). With such small support it is unrealistic to expect robust zero-shot performance across the board. Moreover, in this dataset classes were described by keywords rather than full narrative definitions; where keywords are similar across categories, the model may hesitate or default to the more frequent label.

- Domain and language mismatch. The texts are in Bulgarian and contain legal references; while Gemini is multilingual, domain-specific phrasing and idiomatic legal expressions may not be captured well by a single prompt. Even with the disambiguation dictionary available for many legal references, some class signals are subtle and require interpretive steps that the model did not consistently perform.

- Ambiguity and overlap between classes. Where labels are conceptually close, Gemini sometimes favors a dominant category. This can explain why recall is higher for some classes (the model lumps ambiguous cases into those classes) and why minority classes like 5 are never predicted — the model may systematically map borderline instances to a safer, more frequent label.

- Prompt design. In this case, the description of the classes was made by keywords only, without any textual description or example. This might have effected the

recognition of some classes where the information is expressed through legal reasoning.

Nevertheless, Gemini delivered a usable baseline (accuracy 66%, macro F1 0.56 and weighted F1 0.63), with a result that is better than the one that it obtained for the classification of administrative acts from Regione Emilia-Romagna, and enabled the generation of the silver dataset under the project constraints. However, the per-class analysis reveals systematic blind spots that must be addressed if the silver annotations are to be relied upon for training production classifiers.

6.2.3 Silver Dataset Creation

Following the initial experiments described in the previous section, a larger dataset was created to extend the analysis beyond the gold sample. A total of 216 additional documents were downloaded from the Bulgarian Supreme Court of Cassation and automatically annotated using the Gemini model, which had shown the most reliable performance among the models that were available during the research period.

The ensuing dataset—hereafter referred to as the silver dataset—exhibited a highly skewed distribution across classes. Table 6.15 shows the number of documents per category, from 0 for some classes (e.g., 1, 2, 3, and 12) to a high of 84 for class 6.

Class	N. of documents
1 - Getting married	0
2 - Destruction of a marriage	0
3 - Divorce by mutual consent	0
4 - Divorce due to marital breakdown	20
5 - Personal relations between spouses	1
6 - Parent-child relationships	84
7 - Property relations between spouses – legal regime of community of property	75
8 - Property relations between spouses – legal separation regime	1
9 - Property relations between spouses – negotiated regime	1
10 - Origin	7
11 - Adoption	6
12 - International adoption	0
13 - Guardianship and custody	3
14 - Maintenance	18

Table 6.15: Class distribution of documents in the silver dataset.

As may be observed, the majority of documents are concentrated in a few classes—especially 6 and 7—which also happen to be among the classes that are frequent in the corpus and easier for Gemini to identify. Several other classes, on the other hand, are altogether absent or contain only a single document or two.

This bias is partly structural. Unlike the administrative act used in the previous case study, the family case law decisions available from the Bulgarian Supreme Court of Cassation are fewer, and in a high proportion of instances, not publicly available because they contain sensitive personal data. This makes it very hard to find additional material to rebalance the dataset or increase coverage of rare categories.

Under these limitations, and in light of the distribution present in the gold dataset, see Table 6.13, the experiments were restricted to the six most populous classes: 4, 6, 7, 10, 11, and 14. These classes represent the majority of available examples and provide a sturdy basis for evaluating the classification pipeline without reading too much into performance on extremely sparse categories.

Although the resulting silver dataset remains unbalanced, its shape still reflects realistic working conditions for most administrative and legal domains, where data availability is biased and tends to be skewed towards some subset of procedural types. Experiments on this subset therefore not only serve the purpose of assessing the model's performance but also the one of exploring the feasibility of applying the proposed pipeline under realistic data sparsity; a frequent challenge in domain-specific text classification.

6.2.4 Training Lightweight Models

As with the earlier case study on administrative acts from the Emilia-Romagna Region, the analysis relied on two relatively simple supervised learning algorithms: Logistic Regression and the Support Vector Classifier (SVC). Both were trained using the silver dataset and tested on the gold dataset. The evaluation was restricted to the six most frequent categories (4, 6, 7, 10, 11, and 14) to enable a fair comparison between models trained on automatically annotated data and the zero-shot capabilities of LLMs working within the same reduced label space.

The Logistic Regression model achieved an overall accuracy of 0.57 and a macro-averaged F1-score of 0.54. Its performance, however, was uneven across categories. As reported in Table 6.16, the model performed best on classes 4, 6, and 7, where F1-scores ranged between 0.57 and 0.77. Classes 4 and 7 show an encouraging balance between precision and recall, suggesting that the classifier managed to capture their most distinctive textual characteristics. Class 6 has a very high recall value (0.93), indicating that it frequently assigned this label—sometimes excessively—at the expense of precision.

	Precision	Recall	F1-Score	Support
4 - Divorce due to marital breakdown	0.91	0.67	0.77	15
6 - Parent-child relationships	0.41	0.93	0.57	15
7 - Property relations between spouses – legal regime of community of property	0.52	0.87	0.65	15
10 - Origin	0.70	0.47	0.56	15
11 - Adoption	0.67	0.13	0.22	15
14 - Maintenance	0.71	0.33	0.45	15
accuracy			0.57	90
macro avg	0.65	0.57	0.54	90
weighted avg	0.65	0.57	0.54	90

Table 6.16: Classification report of Logistic Regression trained on the silver dataset and tested on the gold dataset of Bulgarian case law.

On the other hand, classes 10, 11, and 14 yielded less convincing results. Class 11 was particularly problematic, with an F1-score of only 0.22, caused mainly by a very low recall despite reasonably high precision. This means the model rarely identified examples from this class, though the few it did identify were generally correct. Such disparities likely come from the uneven distribution of data within the silver dataset—an expected issue given the imbalance observed in the automatically annotated data.

As shown in Table 6.17, the Support Vector Classifier achieved a global accuracy of 0.54, close to the one of the Logistic Regression model, and its macro-average and weighted F1-scores were both 0.51, slightly lower overall, yet the classifier exhibited sharper contrasts across categories. In particular, SVC achieved perfect precision (1.00) on classes 4 and 10, but this was associated with notably low recall (0.47 and 0.20, respectively), indicating that it only predicted these classes when highly confident. The opposite trend emerged in class 7, where recall reached 1.00 but precision dropped to 0.47, suggesting a tendency to overgeneralize that label in uncertain cases.

	Precision	Recall	F1-Score	Support
4 - Divorce due to marital breakdown	1	0.47	0.64	15
6 - Parent-child relationships	0.39	0.93	0.55	15
7 - Property relations between spouses – legal regime of community of property	0.47	1	0.64	15
10 - Origin	1	0.20	0.33	15
11 - Adoption	0.88	0.47	0.61	15
14 - Maintenance	0.75	0.20	0.32	15
accuracy			0.54	90
macro avg	0.75	0.54	0.51	90
weighted avg	0.75	0.54	0.51	90

Table 6.17: Classification report of SVC trained on the silver dataset and tested on the gold dataset of Bulgarian case law.

Interestingly, class 11, which was one of the weakest categories for Logistic Regression, showed a great improvement under SVC, reaching an F1-score of 0.61 (compared to 0.22 of LR). This difference implies that the non-linear decision boundaries of SVC can better recognize subtle textual distinctions that the linear model cannot capture. Nevertheless, class 14 continued to pose difficulties.

Overall, both models confirm that even this silver dataset can provide a viable training signal. Although performance remains variable, the achieved results are not trivial considering the constraints posed by limited data size and unbalanced categories.

To contextualize these outcomes, Gemini itself was re-evaluated in a zero-shot configuration limited to the same six categories. As visible in Table 6.18, removing all the other labels from the prompt, the model reached an accuracy of 0.82 and a macro-average F1-score of 0.82, outperforming both supervised models by a wide margin. In most categories, Gemini scored above 0.85, particularly in classes 4, 7, and 10, where it displayed remarkably consistent behavior.

	Precision	Recall	F1-Score	Support
4 - Divorce due to marital breakdown	0.88	0.93	0.90	15
6 - Parent-child relationships	0.61	0.93	0.74	15
7 - Property relations between spouses – legal regime of community of property	0.93	0.93	0.93	15
10 - Origin	0.87	0.87	0.87	15
11 - Adoption	0.92	0.80	0.86	15
14 - Maintenance	0.88	0.47	0.61	15
accuracy			0.82	90
macro avg	0.85	0.82	0.82	90
weighted avg	0.85	0.82	0.82	90

Table 6.18: Classification report of Gemini on the gold dataset of Bulgarian case law with only classes 4, 6, 7, 10, 11, and 14.

This impressive result should be interpreted carefully as the observed improvement largely derives from the simplified prompt design: with only six possible classes, the model's decision space was drastically reduced, making the task substantially easier. In practical applications, the silver dataset generated through Gemini needs to cover the complete range of original categories, consequently, the zero-shot evaluation represents a best-case scenario under ideal conditions rather than an accurate reflection of the real task complexity.

6.2.5 Results and Discussion

The experiments conducted on the Bulgarian family case law confirm several of the patterns already observed in the previous case study, while also highlighting new challenges that emerge when moving to a different language and legal system. Overall, Gemini's zero-shot classification achieved the highest performance among the tested models, reaching an accuracy of 0.82 and a macro-average F1-score of 0.82 when restricted to the six most populated classes. In contrast, the lightweight models trained on the Gemini-generated silver dataset performed substantially worse, with Logistic Regression obtaining an F1 score of 0.54 and SVC reaching 0.51.

Although the gap in performance may seem large, it must be interpreted in context. The silver dataset was built automatically from real judicial decisions, with little to no manual

correction, and was highly unbalanced. Under such conditions, achieving accuracy levels close to 60% is already a non-trivial result. Both lightweight models demonstrate that a silver dataset can provide enough signal for a classifier to generalize, even across complex and domain-specific text. Yet, their limitations are equally clear: rare or ambiguous classes tend to collapse into more frequent categories, and small differences in the structure of legal arguments can mislead the models.

The strength of Gemini in the zero-shot setting appears to come from its ability to retain a richer representation of the text. When the number of classes was reduced and the prompt simplified, the model could focus on the semantic distinctions that truly matter for classification, achieving consistently high scores across all six classes. However, this result should be seen as an upper bound. In practice, when Gemini was used to annotate the complete dataset containing documents from all original classes, its precision and recall dropped markedly. The improvement observed in the reduced setup thus reflects a narrower decision space rather than a genuine increase in understanding.

A closer look at the per-class behavior also reveals some interesting asymmetries. Both lightweight models tend to overpredict classes 6 and 7, which are the most represented in the training data and also likely among the easiest to identify based on lexical cues. Conversely, classes such as 10 and 11, despite having clear conceptual boundaries, show lower F1-scores because of their more heterogeneous language. Gemini's zero-shot predictions exhibit similar tendencies, though its contextual understanding allows it to maintain better precision across these same categories. The consistent difficulty with class 14 across all models suggests that this category may involve subtler semantic features that are not easily captured by keyword-based cues or standard embedding spaces.

The comparison between Gemini and the trained classifiers also raises an important methodological point. While Gemini is superior in absolute terms, the light models are much more efficient: once trained, they will classify new documents in milliseconds and execute entirely offline, without access to proprietary APIs. In use cases such as judicial administration or public-sector data management, where budgets are constrained and data privacy is critical, this trade-off becomes valuable. A model trained on silver data, albeit one that is less accurate, may be the only viable solution for big-data or long-term deployment.

Finally, these results underscore the overall implication of this work: domain-specific classification cannot only rely on the shared knowledge encoded in LLMs. While such models are an excellent foundation for generating silver-labeled data, the combination of LLM-driven annotation and light-weight, task-specific retraining remains a feasible compromise. The Bulgarian case study demonstrates that even in a foreign language

and with very restrictive data limitations, the suggested pipeline remains consistent and expandable. The outcome is flawed but realistic—a methodology that remains faithful to the working realities of public institutions, where cost control, sustainability, and interpretability are as important as sheer accuracy.

6.3 Summary and Reflections

The two case studies presented in this chapter offer complementary perspectives on the practical application of the proposed classification pipeline. Taken together, they illustrate both the potential and the current limitations of using large language models and silver datasets in complex, domain-specific contexts.

In the Emilia-Romagna case study, the primary challenge was the unstructured and highly verbose nature of administrative acts. The documents were long, often around 3000 words, and filled with legal references that obscured the information relevant to classification. Despite these difficulties, the pipeline proved effective: once the optimal prompting strategy was identified, a consistent and reasonably accurate silver dataset could be produced, which in turn enabled the training of lightweight classifiers capable of approximating LLM performance while dramatically reducing computational costs.

The Bulgarian case law experiment extended the evaluation to a multilingual and more sensitive domain. Here, data scarcity and class imbalance were even more pronounced, and privacy constraints limited the amount of publicly available material. Still, the results showed that the methodology could adapt to a new setting without fundamental redesign. The zero-shot performance of Gemini, while impressive in controlled conditions, also revealed its fragility when faced with the full complexity of real legal texts.

A key insight emerging from both studies is that the true strength of the proposed pipeline lies not in outperforming large language models, but in making their capabilities accessible and sustainable. By shifting part of the annotation burden from human to LLM and subsequently compressing that knowledge into smaller, task-oriented models, the process acquires a balance between accuracy, explainability, and practicability. Institutional and research environments demand such a balance, as computational resources, budget, and confidentiality of data impose very tight limits on the direct use of proprietary models.

Equally important is the observation that domain knowledge remains a decisive factor. Both case studies confirm that where reasoning or contextual interpretation is required—be it the implicit logic of administrative decisions or the legal reasoning

embedded in court judgments—models trained only on surface patterns tend to fall short. The silver datasets bridge part of this gap, but they cannot fully replace domain expertise. Instead, they provide a reproducible and scalable platform for iterative improvement, where human judgment and machine learning work together without competing.

In summary, the chapter demonstrates that the above approach is not only technologically feasible but also practically meaningful. It can facilitate real-world classification problems in conditions of sparse supervision and imbalance in data and at the same time provide transparency and flexibility. The lessons learned from these two applications inform the broader reflections that follow in the next chapter, where the implications of this work—both methodological and ethical—are discussed in light of its potential for future research and institutional deployment.

Chapter 7

Discussion

This chapter revisits the research questions that guided the study, drawing on the empirical results presented in the previous chapters. Each question is discussed in light of the different experiments conducted — including those on scientific publications related to medical questions, asylum requests in Italy, the Sustainable Development Goals (SDGs), administrative acts from Regione Emilia-Romagna, and family case law from the Bulgarian Supreme Court of Cassation. Together, these domains provide a varied testing ground, ranging from well-structured and well-documented data to unstructured and knowledge-intensive texts.

7.1 The Role of Domain Knowledge

The first research question concerned the role of domain knowledge and the extent to which large language models can handle tasks that require specialized expertise. The experiments in the medical domain and with asylum requests in Italy were particularly informative in this regard.

In the medical domain, the models performed reasonably well in zero-shot classification. The task — finding the correct diagnosis for a clinical case — relied on terminology and conceptual structures that are well represented in the public datasets on which most LLMs are trained. This result suggests that when the target knowledge is widely available, documented, and linguistically standardized, LLMs can leverage their pretraining to achieve high levels of accuracy without additional fine-tuning.

The asylum application test, however, offered the flip side of the coin. Here, the information was more structured - a few questions followed by a simple yes or no

answer - but the accurate prediction of the protection assigned by the judge required an understanding of procedural thinking — identifying, say, if the case was being made on economic grounds, political persecution, or humanitarian grounds. Another difficulty lied in the ability of LLMs to identify of contradictory information, that meant that the person was lying and therefore the protection was to be denied. In this case, all the models were almost always granting some form of protection, whereas in reality that was really rare, as most of the decisions in our data refused it.

These discrepancies verify the answer to the first research question: LLMs perform best when the domain knowledge required for classification is explicit, static, and inherent in their training material, but are extensively impaired when the task involves implicit thinking or contextual understanding on domains where the information is private. Domain knowledge is both an enabler and a constraint: it enables high performance in situations where the model already "knows" the domain, but constrains performance in situations where understanding depends on specialized, proprietary, or procedural knowledge.

7.2 LLMs Behavior in Different Scenarios

The second question asked under what conditions LLMs appear to perform effectively, and what their behavior across different scenarios reveals about their strengths and weaknesses.

Across all experiments, there is a uniform trend. When tasks are very structured, short in duration, and paired with explicit linguistic indications, like in the medical domain and in the case of SDG classification task, LLMs produce very good results with minimal prompting. The experiment of SDG showed that the models could well map keywords and short textual descriptions to individual goals at high accuracy. The short document length (titles, abstracts, and keywords) also helped with this result, lessening noise and leaving the semantic content of interest that comfortably fit within the context windows of the models.

Conversely, on the longer, unstructured, or multi-layered document tasks like the Bulgarian family case law and the Emilia-Romagna administrative acts, performance declined. Both data sets presented extremely similar challenges: documents were long (around 3,000 words for the administrative acts from Regione Emilia-Romagna and 2,000 words for the Bulgarian family case law documents) with many citations to laws and articles that were not always clearly defined. In these settings, the models often failed to

separate essential information from contextual noise. The experiments demonstrated that even GPT-4o, the most advanced among the models tested, required significant prompt engineering and content filtering to perform adequately.

Another pattern is the sensitivity of LLMs to class overlap and ambiguity. In the administrative acts, certain categories shared semantic space — a document could plausibly belong to more than one class, though the task required a single label. The models overfitted most frequent classes or were shallowly linguistically motivated, which suggested the absence of real semantic meaning. A similar effect was seen, though to a lesser extent, in the Bulgarian dataset, where Gemini repeatedly ignored rare classes and overpredicted common ones such as class 6 and 7.

Overall, LLMs perform best in environments where text is very short, categories are clear-cut and well-defined, and domain understanding is part of their pretraining. Their weaknesses emerge as the data become longer, noisier, or more dependent on implicit reasoning. This conflict underlines the need for hybrid pipelines — where the interpretational strength of LLMs is combined with the form and specialism of lighter, domain-specialist models.

7.3 Feasibility and Value of Silver Datasets

The third research question addressed one of the most practical aspects of this work: whether LLMs can serve as substitutes for human annotators, at least to generate silver datasets large enough to train smaller, domain-specific models — and how such models compare with both the LLMs themselves and traditional supervised systems.

The experiments clearly show that while gold datasets remain essential for validation, their creation is costly and time-consuming. In both legal case studies, assembling even a small gold sample required substantial effort from domain experts, who had to read, interpret, and label complex documents. This limitation is not unique to law; similar constraints appear in specialized fields such as healthcare and policy research, where annotation requires advanced knowledge and confidentiality must be respected.

The silver datasets generated through LLM annotation, by contrast, proved to be a viable and efficient alternative. Despite being automatically produced, they retained sufficient quality to train lightweight models capable of generalizing reasonably well. In some cases, such as the classification of Emilia-Romagna administrative acts, models trained on the silver data even outperformed both free-to-use LLMs, i.e., Gemini and Llama. This result is significant: it demonstrates that knowledge distilled from a large model can be transferred

into smaller, more efficient classifiers that operate offline and at a fraction of the cost.

Of course, this approach has limits. Silver datasets inherit the biases and systematic errors of the LLMs that generate them — as seen with underrepresented or ambiguous classes. However, these shortcomings can be minimized through selective manual editing, careful design of prompts, and the application of light-weight models that are better-suited for every-day use and extended maintenance. Silver datasets in this sense augment human insight rather than replace it, allowing experts to focus their efforts on validation and revision rather than annotation.

Chapter 8

Conclusion and Future Work

This thesis was motivated by the objective to innovate how public administrations manage, classify, and make sense of their institutional documents. The work originated within a doctoral project co-funded by the Emilia-Romagna Region, and its first concrete challenge was therefore the automatic classification of the Region's administrative acts. From this initial case study, the research set out to explore how far LLMs of different dimensions can be trusted to perform text classification tasks in domains where annotated data is scarce and where expertise and reasoning are essential. Across several experiments and case studies—ranging from scientific publications and medical diagnostics to asylum requests, administrative acts, and family case law—the results have provided a multifaceted picture of both the potential and the limits of these models.

The research began with a fundamental question: to what extent can LLMs handle domain-specific classification problems when their training data may not cover the necessary knowledge? The answer, as the results have shown, is nuanced. In well-documented fields such as medicine or scientific research, large models like GPT-4o and Gemini achieved high accuracy, leveraging widely available knowledge and consistent terminologies. In contrast, domains like asylum requests or administrative law exposed their weaknesses. Here, tasks required reasoning over long, unstructured, and context-dependent narratives—something current LLMs still struggle to manage without hallucinations, oversimplifications, or bias toward surface cues.

A second key insight concerned the conditions under which LLMs perform well. When knowledge was explicit and standardized, as in the case of the SDG experiment, LLMs got results comparable to human experts, reaffirming their ability to generalize. When knowledge was implicit or intricately linked with institutional practice, as in the case of the Bulgarian or Emilia-Romagna case studies, their limitations were exposed. These findings

suggest that model size is not necessarily synonymous with reliability; domain alignment, prompt design, and task framing are still crucial to succeed.

The third research question focused on the feasibility of using LLMs not only as classifiers but also as data generators, capable of creating silver datasets for training lightweight models. This idea proved to be among the most promising outcomes of the thesis. Despite the difficulty of obtaining gold data—especially when expert annotation is required—the experiments demonstrated that silver data generated through carefully designed prompting can be remarkably effective. In the Emilia-Romagna case, for instance, light-weight models trained on silver data outperformed Gemini and Llama in zero-shot classification, offering a suitable and economic solution for public administrations. These results point to a broader methodological contribution: rather than replacing expert knowledge, LLMs are able to complement it, enabling to scale up domain-specific models in low-resource environments.

The cooperation with Regione Emilia-Romagna demonstrated how automatic classification can support administrative act standardization between regional governments, a step toward national unification. The asylum request task, complemented by the development of an argumentation-based chatbot [116], showed an example of how this technology can assist NGOs and social workers in guiding applicants through the justice system. In the legal domain, the experiments on Bulgarian case law pointed toward tools that could assist judges or legal clerks by offering preliminary document classification or case retrieval suggestions. Finally, the SDG experiments inspired a prototype integrated into the University of Bologna's information system[1], which received an institutional award[2] for innovation in research data management.

8.1 Future Work

While the findings of this thesis are promising, several issues remain to be addressed to improve both performance and applicability.

Data balancing remains a high-priority task. Both the gold and silver datasets often suffer from extreme class imbalances that affect model performance, particularly for sparse or overlapping classes. Future work should aim to increase the number of examples in underrepresented classes, through two parallel strategies: collecting new real documents

[1] https://www.forumpachallenge.it/premio_pa_colori/alma-gaie-alma-sustainable-development-goals-artificial-intelligence-enhanced/

[2] https://magazine.unibo.it/archivio/2024/06/07/artificial-intelligence-for-sustainable-pa-alma-gaie-wins-first-place

in collaboration with domain experts, and synthetically generating data using LLMs. For example, [124] shows that augmenting training data in the low-data regime via synthetic examples (focused especially on the cases the model misclassifies) improves performance significantly across classes. Producing entire synthetic texts modeled on real documents but tailored to specific classes can potentially reduce bias while preserving linguistic and contextual validity.

Knowledge injection is another promising direction. The idea is to supply models with domain knowledge such as legal statutes, taxonomies, and structured ontologies, or embed them into the prompt, adapter modules, or external retrieval mechanisms. For instance, [125] proposes a framework (KgDG) that uses legal knowledge to guide synthetic data generation and verifies the quality of that data, producing improvements in legal reasoning tasks. In the legal domain, robustness to knowledge perturbations has also been studied: [126] investigates whether LLMs are making decisions by reasoning logic or merely surface cues. Their results suggest that pure prompt-based approaches are insufficient without stronger domain signals. Integrating knowledge injection—through retrieval-augmented prompts, legal dictionaries, or structured legal graphs—could therefore strengthen performance, especially for rare or overlapping classes. A particularly promising and more structured form of knowledge injection is represented by Retrieval-Augmented Generation (RAG) architectures [127]. In this regard, one or more pertinent documents, for example legislative laws, administrative guidelines, or already classified acts, are retrieved, and the model's prediction is conditioned on such additional context. This is particularly interesting for the applications presented in this thesis, concerning classifications whose definition is given within institutional frameworks that are not necessarily explicated in the text under analysis. For example, concerning the dataset of administrative acts for Emilia-Romagna, the knowledge injection may consist in the retrieval of a formal regional territory taxonomy or of already correctly classified examples of similar acts. Concerning the experiments on asylum or Bulgarian case law, the knowledge injection may include the retrieval of appropriate laws or precedent cases. For the experiments concerning the domain of medicine, the knowledge injection may also consist of the retrieval of pertinent ontology or of standard diagnostical criteria. Future work should therefore evaluate whether retrieval-augmented classification pipelines outperform pure prompt-based zero-shot approaches, particularly in low-resource and high-specialization contexts.

Iterative re-annotation is similarly valuable. As we had limited resources, we only tested three models, but with increasingly advanced LLMs (for example, GPT-5 and successors or DeepSeek), existing silver datasets could be re-annotated to improve label

consistency, reduce noise, and better align with evolving domain definitions. Research like LLM2LLM [124] supports this kind of iterative process: models correct errors made in early synthetic labels, generate new synthetic examples for challenging cases, and retrain to improve signal. Combining better models with knowledge-injection and synthetic augmentation could push performance closer to human-annotated gold benchmarks. Another improvement given by the re-annotation could come from the addition of new information in the prompt, for example, for the classification of administrative acts from Regione Emilia-Romagna, explicitly stating where to classify all the documents that are about the COVID emergency, sustainable mobility, gender violence and discrimination, could solve some of the gray areas that we encountered, achieving better accuracy in the classification task.

Multilingual and cross-domain generalization should also be explored more deeply. In the recent years, new models have been developed to try to take AI beyond the boundaries of the English language. For example, in the medical field, Medical T5 [128] has been trained on a corpus of four languages, namely English, French, Italian and Spanish. This model could close the performance gap among languages, evident in Llama's performance on the medical domain, see Table 4.1. Also, many legal or administrative tasks operate across languages and jurisdictions, for example, the detection of unfair clauses in terms of service is a theme common to all languages and a classifier has recently been developed for their automatic detection in English, Italian, German and Polish [129]; methods that adapt to such variations can broaden impact. This includes evaluating how well knowledge injection and dataset augmentation strategies work in non-English domains, under different legal systems and document styles.

In brief, this thesis has shown that the combination of LLMs and lightweight classifiers is a viable, environmentally conscious solution to text classification in data-scarce, domain-specific settings. The results are not theoretical abstractions: they point toward a new model for uniting automation and expertise—one where artificial intelligence enhances, but does not replace, human judgment.

Bibliography

[1] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[2] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*, pp. 137–142, Springer, 1998.

[3] C. D. Manning, *Introduction to information retrieval*. Syngress Publishing,, 2008.

[4] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-bert: The muppets straight out of law school," *arXiv preprint arXiv:2010.02559*, 2020.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[6] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[7] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[8] B. Settles, "Active learning literature survey," 2009.

[9] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[11] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "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, 2021.

[12] R. Bommasani, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[14] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," 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. 3816–3830, 2021.

[15] N. Kassner and H. Schütze, "Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7811–7818, 2020.

[16] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.

[17] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[18] M. Jiang, Y. Liang, X. Feng, X. Fan, Z. Pei, Y. Xue, and R. Guan, "Text classification based on deep belief network and softmax regression," *Neural Computing and Applications*, vol. 29, pp. 61–70, 2018.

[19] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "Hdltex: Hierarchical deep learning for text classification," in *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pp. 364–371, IEEE, 2017.

[20] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Madison, WI, 1998.

[21] P. S. Jacobs, *Text-based intelligent systems: Current research and practice in information extraction and retrieval.* Psychology Press, 2014.

[22] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*, vol. 520. Addison-Wesley Reading, 2010.

[23] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?," in *Proceedings of the 26th annual computer security applications conference*, pp. 21–30, 2010.

[24] R. S. Gordon Jr, "An operational classification of disease prevention," *Public health reports*, vol. 98, no. 2, p. 107, 1983.

[25] A. L. Nobles, J. J. Glenn, K. Kowsari, B. A. Teachman, and L. E. Barnes, "Identification of imminent suicide risk among young adults using text messages," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–11, 2018.

[26] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[29] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[30] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[31] J. J. Rocchio Jr, "Relevance feedback in information retrieval," *The SMART retrieval system: experiments in automatic document processing*, 1971.

[32] R. E. Schapire, "The strength of weak learnability," *Machine learning*, vol. 5, pp. 197–227, 1990.

[33] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

[34] D. R. Cox, *Analysis of binary data.* Routledge, 2018.

[35] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[36] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved k-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012.

[37] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

[38] J. N. Morgan and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," *Journal of the American statistical association*, vol. 58, no. 302, pp. 415–434, 1963.

[39] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282, IEEE, 1995.

[40] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach," 2019.

[41] F. Graliński, T. Stanisławek, A. Wróblewska, D. Lipiński, A. Kaliska, P. Rosalska, B. Topolski, and P. Biecek, "Kleister: A novel task for information extraction involving long documents with complex layout," *arXiv preprint arXiv:2003.02356*, 2020.

[42] Z. Dong, T. Tang, L. Li, and W. X. Zhao, "A survey on long text modeling with transformers," *arXiv preprint arXiv:2302.14502*, 2023.

[43] D. Tsirmpas, I. Gkionis, G. T. Papadopoulos, and I. Mademlis, "Neural natural language processing for long texts: A survey on classification and summarization," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108231, July 2024.

[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

[45] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[46] I. Chalkidis, X. Dai, M. Fergadiotis, P. Malakasiotis, and D. Elliott, "An exploration of hierarchical attention transformers for efficient long document classification," *arXiv preprint arXiv:2210.05529*, 2022.

[47] Y. Liu, A. Ni, L. Nan, B. Deb, C. Zhu, A. H. Awadallah, and D. Radev, "Leveraging locality in abstractive text summarization," *arXiv preprint arXiv:2205.12476*, 2022.

[48] J. Zhang, P. Lertvittayakumjorn, and Y. Guo, "Integrating semantic knowledge to tackle zero-shot text classification," *arXiv preprint arXiv:1903.12626*, 2019.

[49] Y. Meng, Y. Zhang, J. Huang, C. Xiong, H. Ji, C. Zhang, and J. Han, "Text classification using label names only: A language model self-training approach," *arXiv preprint arXiv:2010.07245*, 2020.

[50] P. K. Pushp and M. M. Srivastava, "Train once, test anywhere: Zero-shot learning for text classification," 2017.

[51] T. Schopf, D. Braun, and F. Matthes, "Evaluating unsupervised text classification: zero-shot and similarity-based approaches," in *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, pp. 6–15, 2022.

[52] D. Braun, O. Klymenko, T. Schopf, Y. KAAN AKAN, and F. Matthes, "The language of engineering: training a domain-specific word embedding model for engineering," in *Proceedings of the 2021 3rd International Conference on Management Science and Industrial Engineering*, pp. 8–12, 2021.

[53] T. Schopf, S. Klimek, and F. Matthes, "Patternrank: Leveraging pretrained language models and part of speech for unsupervised keyphrase extraction," *arXiv preprint arXiv:2210.05245*, 2022.

[54] P. Schneider, T. Schopf, J. Vladika, M. Galkin, E. Simperl, and F. Matthes, "A decade of knowledge graphs in natural language processing: A survey," *arXiv preprint arXiv:2210.00105*, 2022.

[55] S. P. Veeranna, J. Nam, E. L. Mencıa, and J. Fürnkranz, "Using semantic similarity for multi-label zero-shot classification of text documents," in *Proceeding of european symposium on artificial neural networks, computational intelligence and machine learning. bruges, belgium: Elsevier*, pp. 423–428, 2016.

[56] Z. Haj-Yahia, A. Sieg, and L. A. Deleris, "Towards unsupervised text classification leveraging experts and word embeddings," in *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pp. 371–379, 2019.

[57] M.-W. Chang, L.-A. Ratinov, D. Roth, and V. Srikumar, "Importance of semantic representation: Dataless classification.," in *Aaai*, vol. 2, pp. 830–835, 2008.

[58] E. Gabrilovich, S. Markovitch, *et al.*, "Computing semantic relatedness using wikipedia-based explicit semantic analysis.," in *IJcAI*, vol. 7, pp. 1606–1611, 2007.

[59] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[60] T. Schopf, D. Braun, and F. Matthes, "Lbl2vec: An embedding-based approach for unsupervised document retrieval on predefined topics," *arXiv preprint arXiv:2210.06023*, 2022.

[61] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, pp. 1188–1196, PMLR, 2014.

[62] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," *arXiv preprint arXiv:2104.08821*, 2021.

[63] N. Reimers, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[64] P. Lepagnol, T. Gerald, S. Ghannay, C. Servan, and S. Rosset, "Small language models are good too: An empirical study of zero-shot classification," 2024.

[65] K. Halder, A. Akbik, J. Krapac, and R. Vollgraf, "Task-aware representation of sentences for generic text classification," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3202–3213, 2020.

[66] Y. Fei, P. Nie, Z. Meng, R. Wattenhofer, and M. Sachan, "Beyond prompting: Making pre-trained language models better zero-shot learners by clustering representations," *arXiv preprint arXiv:2210.16637*, 2022.

[67] J. Lu, D. Zhu, W. Han, R. Zhao, B. Mac Namee, and F. Tan, "What makes pre-trained language models better zero-shot learners?," *arXiv preprint arXiv:2209.15206*, 2022.

[68] Y. Mu, B. P. Wu, W. Thorne, A. Robinson, N. Aletras, C. Scarton, K. Bontcheva, and X. Song, "Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science," *arXiv preprint arXiv:2305.14310*, 2023.

[69] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[70] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[71] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[72] A. Ashok, N. Rhinehart, F. Beainy, and K. M. Kitani, "N2n learning: Network to network compression via policy gradient reinforcement learning," *arXiv preprint arXiv:1709.06030*, 2017.

[73] K.-H. Lai, D. Zha, Y. Li, and X. Hu, "Dual policy distillation," *arXiv preprint arXiv:2006.04061*, 2020.

[74] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

[75] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 126–136, 2018.

[76] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[77] J. Ba and R. Caruana, "Do deep nets really need to be deep?," *Advances in neural information processing systems*, vol. 27, 2014.

[78] G. Urban, K. J. Geras, S. E. Kahou, O. Aslan, S. Wang, R. Caruana, A. Mohamed, M. Philipose, and M. Richardson, "Do deep convolutional nets really need to be deep and convolutional?," *arXiv preprint arXiv:1603.05691*, 2016.

[79] X. Cheng, Z. Rao, Y. Chen, and Q. Zhang, "Explaining knowledge distillation by quantifying the knowledge," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12925–12935, 2020.

[80] M. Phuong and C. Lampert, "Towards understanding knowledge distillation," in *International conference on machine learning*, pp. 5142–5151, PMLR, 2019.

[81] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

[82] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, "Data-free learning of student networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3514–3522, 2019.

[83] R. G. Lopes, S. Fenu, and T. Starner, "Data-free knowledge distillation for deep neural networks," *arXiv preprint arXiv:1710.07535*, 2017.

[84] G. K. Nayak, K. R. Mopuri, V. Shaj, V. B. Radhakrishnan, and A. Chakraborty, "Zero-shot knowledge distillation in deep networks," in *International conference on machine learning*, pp. 4743–4751, PMLR, 2019.

[85] J. Ye, Y. Ji, X. Wang, X. Gao, and M. Song, "Data-free knowledge amalgamation via group-stack dual-gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12516–12525, 2020.

[86] P. Micaelli and A. J. Storkey, "Zero-shot knowledge transfer via adversarial belief matching," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[87] J. Yoo, M. Cho, T. Kim, and U. Kang, "Knowledge extraction with no observable data," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[88] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8715–8724, 2020.

[89] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4119–4128, 2018.

[90] P. Liu, I. King, M. R. Lyu, and J. Xu, "Ddflow: Learning optical flow with unlabeled data distillation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 8770–8777, 2019.

[91] W. Zhang, X. Miao, Y. Shao, J. Jiang, L. Chen, O. Ruas, and B. Cui, "Reliable data distillation on graph convolutional network," in *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, pp. 1399–1414, 2020.

[92] Z. Wang, Y. Pang, and Y. Lin, "Large language models are zero-shot text classifiers," *arXiv preprint arXiv:2312.01044*, 2023.

[93] A. Kermani, V. Perez-Rosas, and V. Metsis, "A systematic evaluation of llm strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. rag," *arXiv preprint arXiv:2503.24307*, 2025.

[94] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, and S. Joty, "Data augmentation using large language models: Data perspectives, learning paradigms and challenges," *arXiv preprint arXiv:2403.02990*, 2024.

[95] R. Li, X. Wang, and H. Yu, "Two directions for clinical data generation with large language models: data-to-label and label-to-data," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2023, p. 7129, 2023.

[96] N. Pangakis and S. Wolken, "Knowledge distillation in automated annotation: Supervised text classification with llm-generated training labels," *arXiv preprint arXiv:2406.17633*, 2024.

[97] R. M. Bakker, A. J. Schoevers, R. A. van Drie, M. P. Schraagen, and M. H. de Boer, "Semantic role extraction in law texts: a comparative analysis of language models for legal information extraction," *Artificial Intelligence and Law*, pp. 1–35, 2025.

[98] A. Nagar, V. Schlegel, T.-T. Nguyen, H. Li, Y. Wu, K. Binici, and S. Winkler, "Llms are not zero-shot reasoners for biomedical information extraction," *arXiv preprint arXiv:2408.12249*, 2024.

[99] Y. Li, "A practical survey on zero-shot prompt design for in-context learning," *arXiv preprint arXiv:2309.13205*, 2023.

[100] S. N. Cheetirala, G. Raut, D. Patel, F. Sanatana, R. Freeman, M. A. Levin, G. N. Nadkarni, O. Dawkins, R. Miller, R. M. Steinhagen, *et al.*, "Less context, same performance: A rag framework for resource-efficient llm-based clinical nlp," *arXiv preprint arXiv:2505.20320*, 2025.

[101] G. Han, J. Tsao, and X. Huang, "Length-aware multi-kernel transformer for long document classification," *arXiv preprint arXiv:2405.07052*, 2024.

[102] V. Kanjirangat, A. Antonucci, and M. Zaffalon, "On the limitations of zero-shot classification of causal relations by llms," in *Text2Story@ ECIR*, 2024.

[103] T. Han, Z. Wang, C. Fang, S. Zhao, S. Ma, and Z. Chen, "Token-budget-aware llm reasoning," *arXiv preprint arXiv:2412.18547*, 2024.

[104] S. Shekhar, T. Dubey, K. Mukherjee, A. Saxena, A. Tyagi, and N. Kotla, "Towards optimizing the costs of llm usage," *arXiv preprint arXiv:2402.01742*, 2024.

[105] H. Qi, G. Fu, J. Li, C. Song, W. Zhai, D. Luo, S. Liu, Y. Yu, B. Yang, and Q. Zhao, "Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media," *Bioengineering*, vol. 12, no. 8, p. 882, 2025.

[106] E. Le Priol, J. Potier, and A. Burgun, "Can generative llms help classify imbalanced real-world data? exploring rare diseases on social media," *Studies in health technology and informatics*, vol. 329, pp. 683–687, 2025.

[107] Z. Song, B. Yan, Y. Liu, M. Fang, M. Li, R. Yan, and X. Chen, "Injecting domain-specific knowledge into large language models: a comprehensive survey," *arXiv preprint arXiv:2502.10708*, 2025.

[108] C. Ling, X. Zhao, J. Lu, C. Deng, C. Zheng, J. Wang, T. Chowdhury, Y. Li, H. Cui, X. Zhang, T. Zhao, A. Panalkar, D. Mehta, S. Pasquali, W. Cheng, H. Wang, Y. Liu, Z. Chen, H. Chen, C. White, Q. Gu, J. Pei, C. Yang, and L. Zhao, "Domain specialization as the key to make large language models disruptive: A comprehensive survey," 2024.

[109] Y. Lee, D. Goldwasser, and L. S. Reese, "Towards understanding counseling conversations: Domain knowledge and large language models," in *Findings of the Association for Computational Linguistics: EACL 2024* (Y. Graham and M. Purver, eds.), (St. Julian's, Malta), pp. 2032–2047, Association for Computational Linguistics, Mar. 2024.

[110] M. A. Ferrag, N. Tihanyi, and M. Debbah, "Reasoning beyond limits: Advances and open problems for llms," 2025.

[111] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models," 2025.

[112] Y. Zhang, L. Chen, S. Li, N. Cao, Y. Shi, J. Ding, Z. Qu, P. Zhou, and Y. Bai, "Way to specialist: Closing loop between specialized llm and evolving domain knowledge graph," 2024.

[113] B. Molinet, S. Marro, E. Cabrio, and S. Villata, "Explanatory argumentation in natural language for correct and incorrect medical diagnoses," *Journal of Biomedical Semantics*, vol. 15, no. 1, p. 8, 2024.

[114] E. Palmieri, "Technical report: Clinical diagnosis through llms: A comparative evaluation of accuracy and explanation," 2024.

[115] E. Sviridova, A. Yeginbergen, A. Estarrona, E. Cabrio, S. Villata, and R. Agerri, "Casimedicos-arg: A medical question answering dataset annotated with explanatory argumentative structures," 2024.

[116] B. Fazzinga, E. Palmieri, M. Vestoso, L. Bolognini, A. Galassi, F. Furfaro, and P. Torroni, "A chatbot for asylum-seeking migrants in europe," in *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 702–707, IEEE, 2024.

[117] L. Bolognini, E. Palmieri, N. Donati, G. Grundler, G. Pappacoda, F. Ruggeri, A. Galassi, and P. Torroni, "Almasdg: A dataset of scientific articles' contributions to the un sustainable development goals," *Springer Nature*, 2025. Under Review.

[118] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[119] M. Vanderfeesten, R. Jaworek, and L. Keßler, "Ai for mapping multi-lingual academic papers to the united nations' sustainable development goals (sdgs)," *Zenodo*, 2022.

[120] R. Jaworek, "Sdgs multiclass clasiffier," Sept. 2022.

[121] L. Pukelis, N. Bautista-Puig, G. Statulevičiūtė, V. Stančiauskas, G. Dikmener, and D. Akylbekova, "Osdg 2.0: a multilingual tool for classifying text data by un sustainable development goals (sdgs)," 2022.

[122] OSDG, UNDP IICPSD SDG AI Lab, and PPMI, "OSDG community dataset (OSDG-CD)," Jan. 2023.

[123] D. U. Wulff and D. S. Meier, "SDG Knowledge Hub Dataset of SDG-labeled News Articles," Jan. 2023.

[124] N. Lee, T. Wattanawong, S. Kim, K. Mangalam, S. Shen, G. Anumanchipalli, M. W. Mahoney, K. Keutzer, and A. Gholami, "Llm2llm: Boosting llms with novel iterative data enhancement," *arXiv preprint arXiv:2403.15042*, 2024.

[125] Z. Zhou, K.-Y. Yu, S.-Y. Tian, X.-W. Yang, J.-X. Shi, P. Song, Y.-X. Jin, L.-Z. Guo, and Y.-F. Li, "Lawgpt: Knowledge-guided data generation and its application to legal llm," *arXiv preprint arXiv:2502.06572*, 2025.

[126] Y. Hu, H. Liu, Q. Chen, N. Zheng, C. Wang, Y. Liu, C. L. Clarke, and W. Shen, "J&h: evaluating the robustness of large language models under knowledge-injection attacks in legal domain," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 28106–28115, 2025.

[127] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[128] I. García-Ferrero, R. Agerri, A. A. Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, *et al.*, "Medical mt5: an open-source multilingual text-to-text llm for the medical domain," *arXiv preprint arXiv:2404.07613*, 2024.

[129] A. Galassi, F. Lagioia, A. Jabłonowska, and M. Lippi, "Unfair clause detection in terms of service across multiple languages," *Artificial Intelligence and Law*, pp. 1–49, 2024.

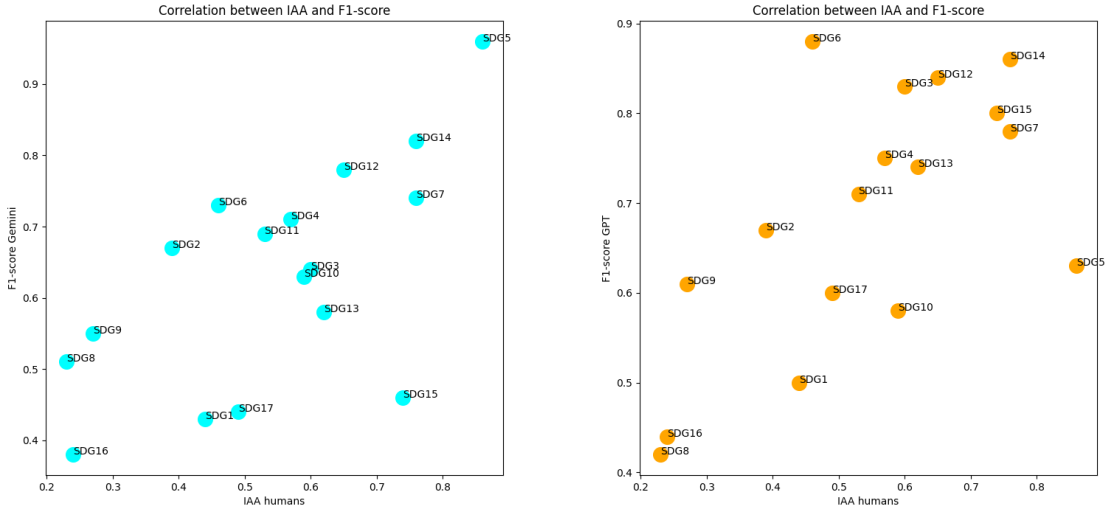
Appendix

Inter Annotator Agreement and F1 Score correlation of the Classification of SDGs

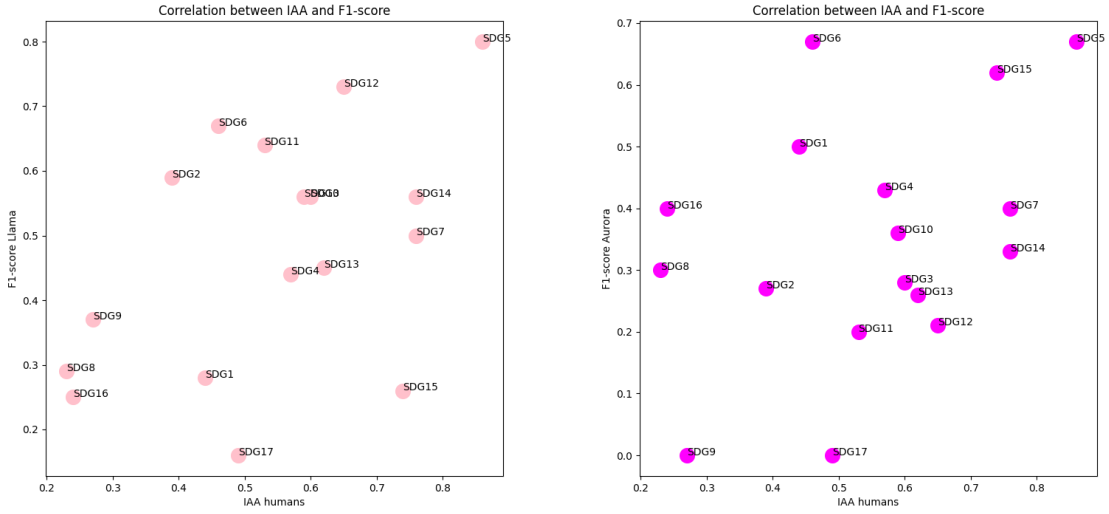
The following sections show the graphs representing the correlation between the IAA among human annotators and the F1-Scores achieved by the different models with and without fine-tuning, that are described in Chapeter 5.

Zero shot models without tuning

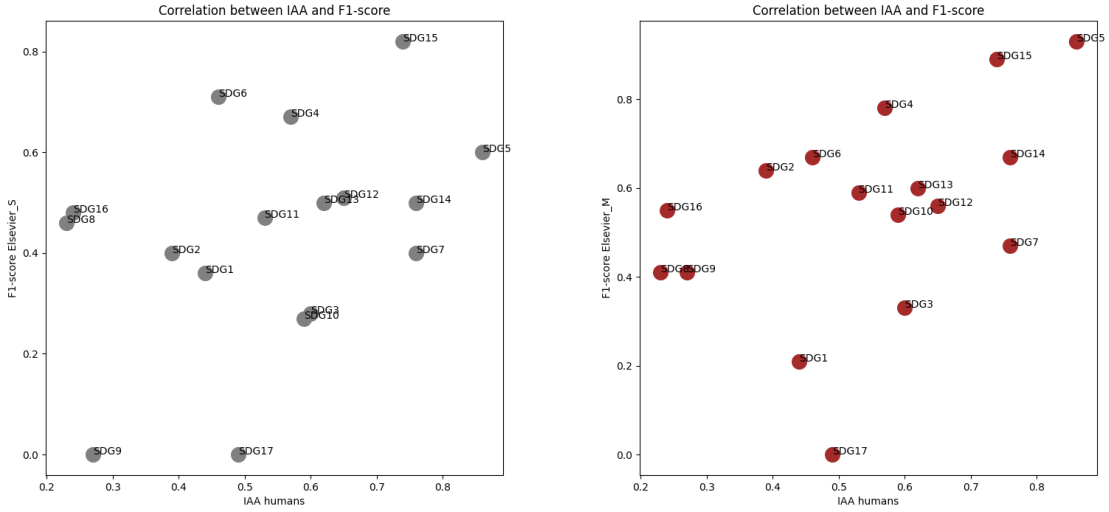
Figures 8.1a, 8.1b, 8.2a, 8.2b, 8.3a, 8.3b, and 8.4 refer to the correlation between the IAA among humans in Table 5.1 and the results obtained by the models without fine-tuning, visible in Table 5.5.



(a) Correlation between the IAA among humans and the F1-score obtained by Gemini without fine-tuning. (b) Correlation between the IAA among humans and the F1-score obtained by GPT without fine-tuning.



(a) Correlation between the IAA among humans and the F1-score obtained by Llama without fine-tuning. (b) Correlation between the IAA among humans and the F1-score obtained by Aurora without fine-tuning.



(a) Correlation between the IAA among humans and the F1-score obtained by Elsevier_S without fine-tuning.

(b) Correlation between the IAA among humans and the F1-score obtained by Elsevier_M without fine-tuning.

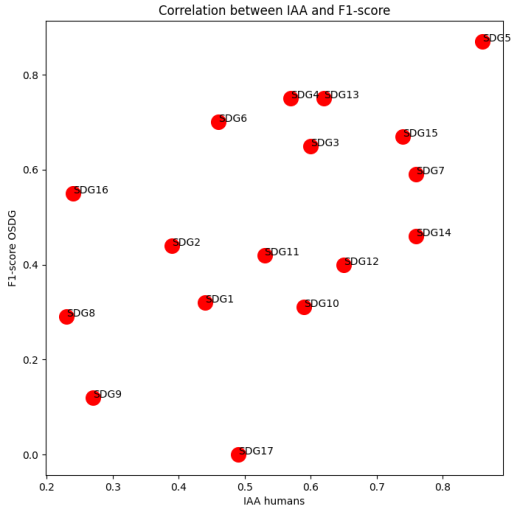
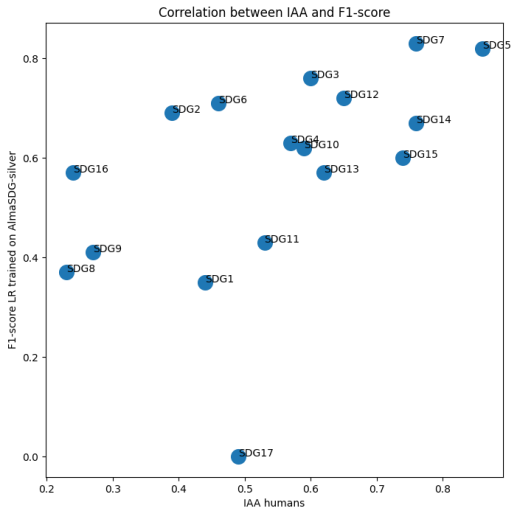


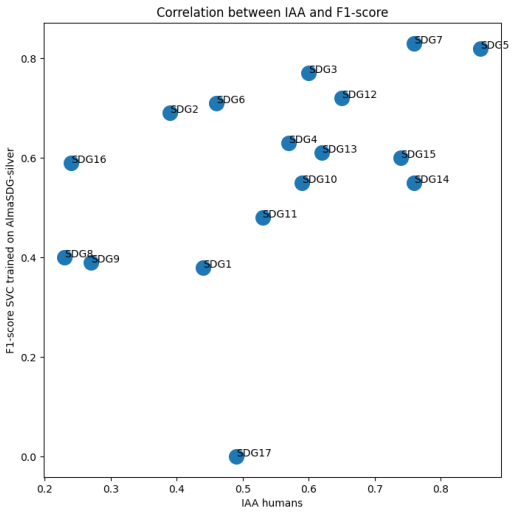
Figure 8.4: Correlation between the IAA among humans and the F1-score obtained by OSDG without fine-tuning.

Models trained on AlmaSDG-silver

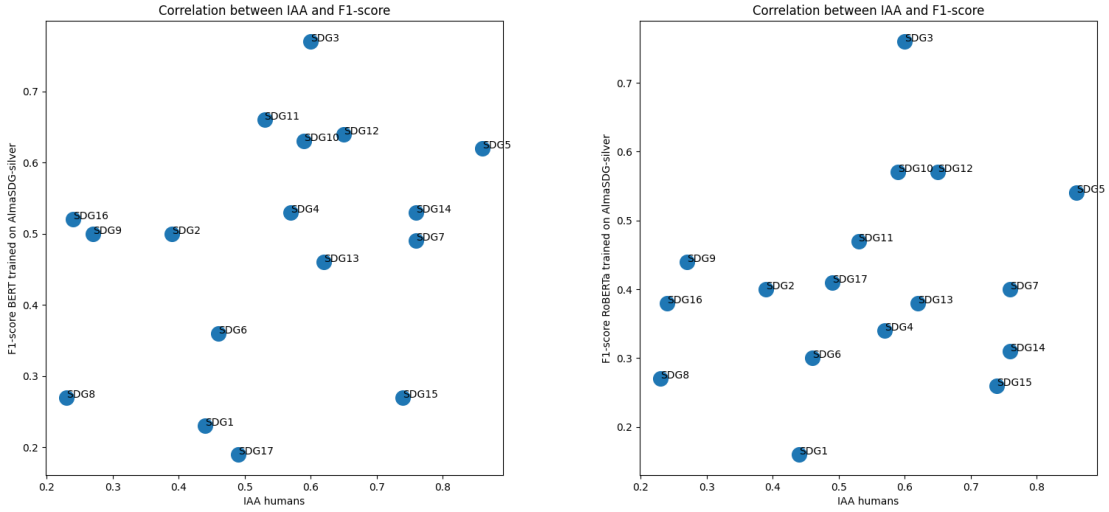
Figures 8.5a, 8.5b, 8.6a, and 8.6b refer to the correlation between the IAA among humans in Table 5.1 and the results obtained by the models fine-tuned on AlmsSDG-silver, visible in Table 5.6.



(a) Correlation between the IAA among humans and the F1-score obtained by LR fine-tuned on AlmaSDG-silver.



(b) Correlation between the IAA among humans and the F1-score obtained by SVC fine-tuned on AlmaSDG-silver.

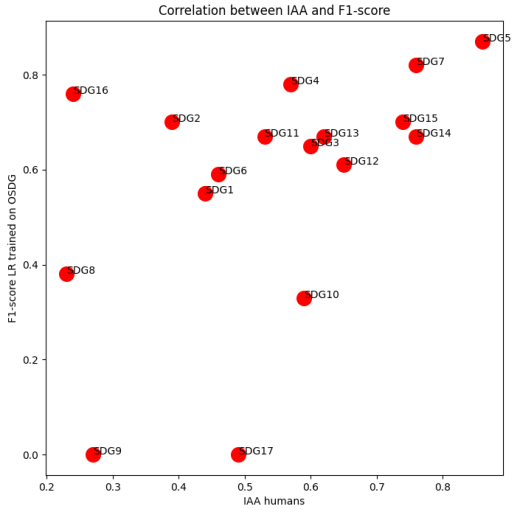


(a) Correlation between the IAA among humans and the F1-score obtained by BERT fine-tuned on AlmaSDG-silver.

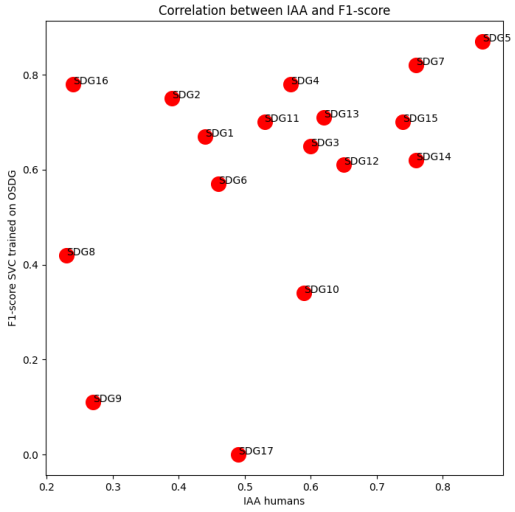
(b) Correlation between the IAA among humans and the F1-score obtained by RoBERTa fine-tuned on AlmaSDG-silver.

Models trained on OSDG

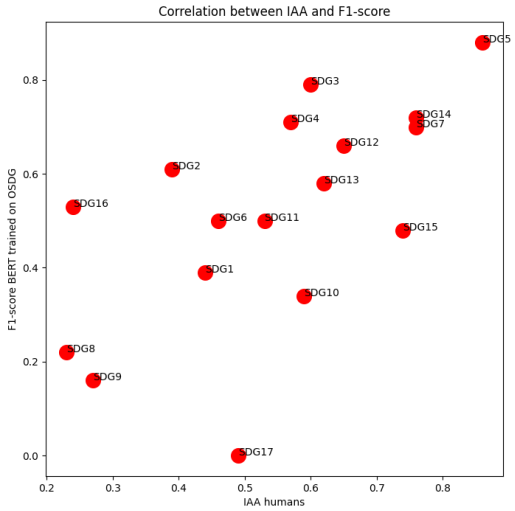
Figures 8.7a, 8.7b, 8.8a, and 8.8b refer to the correlation between the IAA among humans in Table 5.1 and the results obtained by the models fine-tuned on OSDG, visible in Table 5.7.



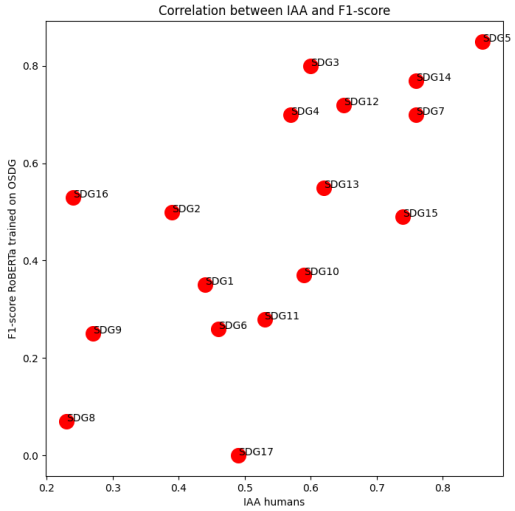
(a) Correlation between the IAA among humans and the F1-score obtained by LR fine-tuned on OSDG.



(b) Correlation between the IAA among humans and the F1-score obtained by SVC fine-tuned on OSDG.



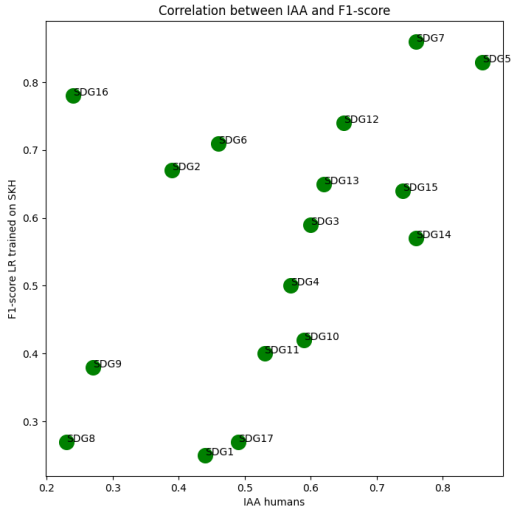
(a) Correlation between the IAA among humans and the F1-score obtained by BERT fine-tuned on OSDG.



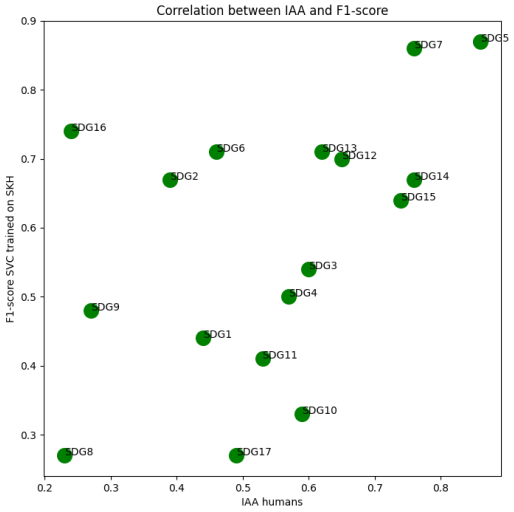
(b) Correlation between the IAA among humans and the F1-score obtained by RoBERTa fine-tuned on OSDG.

Models trained on SKH

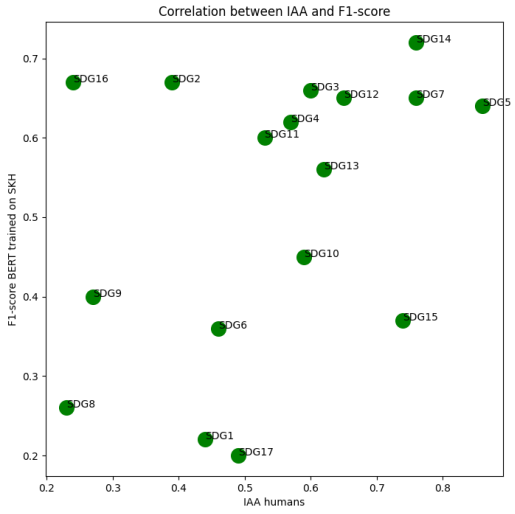
Figures 8.9a, 8.9b, 8.10a, and 8.10b refer to the correlation between the IAA among humans in Table 5.1 and the results obtained by the models fine-tuned on SKH, visible in Table 5.8.



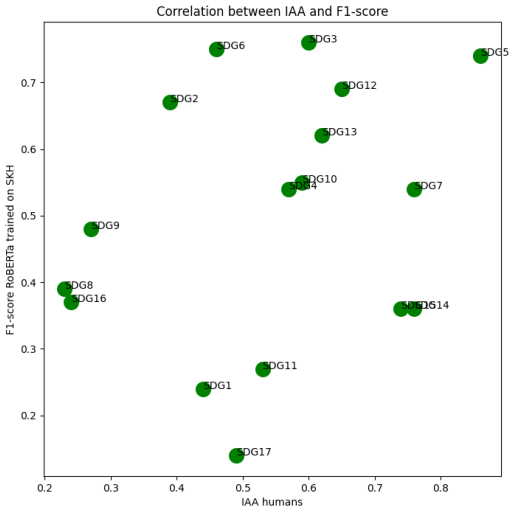
(a) Correlation between the IAA among humans and the F1-score obtained by LR fine-tuned on SKH.



(b) Correlation between the IAA among humans and the F1-score obtained by SVC fine-tuned on SKH.



(a) Correlation between the IAA among humans and the F1-score obtained by BERT fine-tuned on SKH.



(b) Correlation between the IAA among humans and the F1-score obtained by RoBERTa fine-tuned on SKH.

Zero-Shot Classification Results of Small Models on Administrative Acts

Images 8.11, 8.12a and 8.12b show the confusion matrices of the smaller models, BART, T5 and RoBERTa respectively, of the zero-shot classification of administrative acts from Regione Emilia-Romagna.

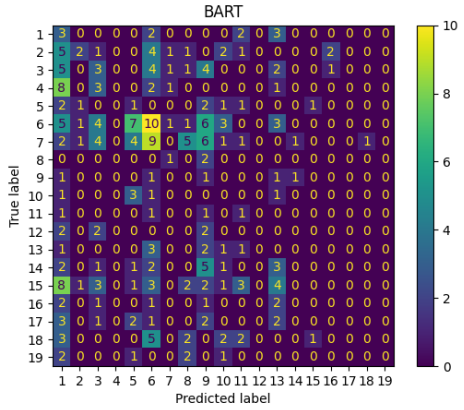
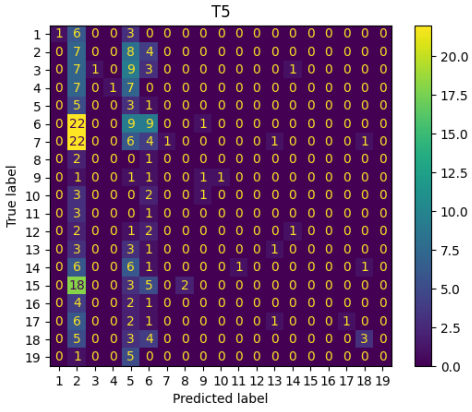
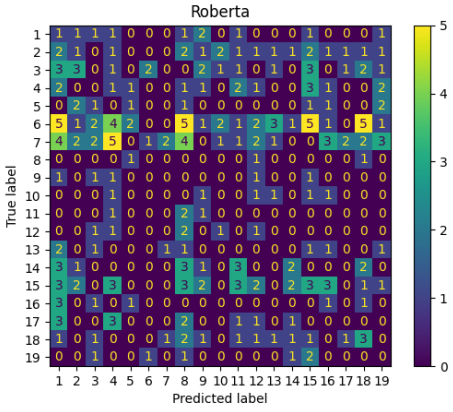


Figure 8.11: Confusion matrix of BART.



(a) Confusion matrix of T5.



(b) Confusion matrix of RoBERTa.

Prompting Techniques Results for Classification of Administrative Acts

Classes Description

This is the description of every class extracted from the *Titolario* for the classification of administrative acts from Regione Emilia-Romagna. As the task was performed on Italian

documents and the Titolario is in Italian, so were all the prompts given to the LLMs. Given the length of the documents, we could not use the full description of the classes as is in the Titolario, we had to create a summary that includes the key points for every class.

01: ORGANI DI GOVERNO, AFFARI GIURIDICO-ISTITUZIONALI E COMUNICAZIONE Comprende le attività legate alla configurazione e al funzionamento degli organi di governo e amministrativi della Giunta, i rapporti istituzionali con enti nazionali e internazionali, la gestione degli organismi intersettoriali, la comunicazione istituzionale, la rappresentanza, gli affari legali e il contenzioso, nonché la pubblicazione di atti ufficiali e la cura del cerimoniale.

02: ORGANIZZAZIONE, PATRIMONIO E RISORSE STRUMENTALI Comprende le attività di supporto e organizzazione interna dell'Ente, assicurandone il funzionamento e l'operatività. Include la gestione strutturale e logistica, la sicurezza sul lavoro, la gestione del demanio e del patrimonio immobiliare e mobiliare, le attività negoziali e contrattuali, le locazioni passive, la fornitura di beni e servizi, gli interventi manutentivi e i servizi ausiliari.

03: RISORSE UMANE Comprende le attività di gestione del personale, dalla costituzione alla cessazione del rapporto di lavoro, inclusi reclutamento, stato giuridico, trattamento economico, previdenza, mobilità, formazione, valutazione, rapporti sindacali e pari opportunità. Includendo anche la gestione di contratti atipici, la disciplina e i servizi a favore del personale.

04: RISORSE FINANZIARIE, GESTIONE CONTABILE E FISCALE Comprende le attività di formazione e gestione del bilancio, rendicontazione, gestione economale, entrate e tributi, debito e rapporti finanziari con enti e istituzioni. Include la fiscalità attiva e passiva, il controllo contabile su enti regionali e la consulenza normativa e programmatoria di settore.

05: SISTEMA INFORMATIVO Riguarda la gestione dell'informatizzazione dell'ente, includendo progettazione, realizzazione e manutenzione delle componenti tecnologiche e applicative del sistema informativo regionale. Comprende anche la gestione documentale, il protocollo, gli archivi, le relazioni con il pubblico, la normativa sulla privacy, la biblioteca regionale, e i servizi statistici e informativi.

06: PROGRAMMAZIONE, COORDINAMENTO E CONTROLLO Si occupa della programmazione generale dell'ente, della gestione di progetti e affari europei in ambito intersettoriale, del coordinamento e del controllo interno ed esterno. Include la programmazione regionale e comunitaria, la gestione di strumenti programmatori, il monitoraggio e la valutazione delle politiche, il controllo della gestione e il supporto tecnico-giuridico per enti, autonomie locali e terzo settore.

07: AGRICOLTURA E ALLEVAMENTO Include le attività relative all'agricoltura, allevamento, pesca e settori collegati, dalla programmazione alle relazioni con le Agenzie regionali. Comprende la regolamentazione, i finanziamenti e i servizi rivolti alle imprese agricole e zootecniche, la gestione delle foreste, la promozione dell'agriturismo e la tutela della pesca e della fauna ittica, oltre ad attività di studio, monitoraggio e divulgazione.

08: ARTIGIANATO Comprende le attività di disciplina, tutela, promozione e sviluppo dell'artigianato, inclusi finanziamenti e servizi di supporto. Si occupa della programmazione settoriale, della gestione di commissioni e osservatori, della promozione dell'artigianato artistico e dell'eccellenza, del riconoscimento delle imprese e del sostegno tecnico per innovazione e certificazione, in collaborazione con istituzioni locali, nazionali e comunitarie.

09: COMMERCIO Si occupa della programmazione e pianificazione della rete distributiva regionale, delle autorizzazioni per le strutture di vendita, delle attività relative a fiere e mercati, dell'erogazione di incentivi e servizi alle imprese e della tutela dei consumatori. Include attività di studio, ricerca e monitoraggio settoriale, gestione di organismi e osservatori, promozione, regolamentazione degli orari e delle modalità di vendita, e vigilanza su fiere, mercati e commercio su aree pubbliche, con interventi per agevolazioni, contributi e progetti comunitari.

10: INDUSTRIA E ATTIVITA' ESTRATTIVE Riguarda le politiche regionali per il settore industriale, i servizi alla produzione e le attività minerarie ed estrattive. Include la programmazione, l'elaborazione normativa, la concertazione con istituzioni e categorie, lo studio e monitoraggio del settore, la gestione di comitati e conferenze tecniche, e attività promozionali. Si occupa di agevolazioni, contributi, supporto all'innovazione, creazione d'impresa, infrastrutture produttive, internazionalizzazione e ricerca scientifica. Per le attività estrattive, gestisce autorizzazioni, vigilanza, coordinamento e concessioni relative a cave, miniere e risorse naturali.

11: TURISMO E STRUTTURE RICETTIVE Riguarda le politiche, l'organizzazione territoriale e gli incentivi per il turismo, includendo l'elaborazione normativa, la programmazione, lo studio e il monitoraggio del settore. Comprende la gestione di commissioni, attività promozionali in Italia e all'estero, il supporto alle strutture territoriali e agli enti turistici, le autorizzazioni per agenzie di viaggio e professioni turistiche, e la regolamentazione delle strutture ricettive alberghiere ed extralberghiere attraverso classificazioni, vigilanza, autorizzazioni e contributi.

12: PIANIFICAZIONE TERRITORIALE, URBANISTICA ED EDILIZIA Rriguarda lo sviluppo urbanistico del territorio, la pianificazione dello sviluppo edilizio pubblico e privato e il soddisfacimento del fabbisogno abitativo. Include attività normative, programmazione, monitoraggio, produzione di cartografia tecnica, vigilanza sugli strumenti urbanistici e sugli appalti pubblici, oltre al finanziamento e alla gestione di interventi per l'edilizia residenziale, scolastica, sanitaria e storico-culturale. Comprende anche la promozione, il recupero dei centri storici, la riqualificazione urbana, l'abbattimento di barriere architettoniche e il supporto agli enti locali.

13: INFRASTRUTTURE E TRASPORTI Riguarda la pianificazione, il coordinamento e lo sviluppo delle reti e delle infrastrutture di comunicazione, della mobilità urbana ed extraurbana, della rete viaria, ferroviaria, portuale e aeroportuale. Comprende attività normative, programmazione, monitoraggio, progettazione e finanziamento per strade, trasporto pubblico, porti, aeroporti e linee ferroviarie. Include la gestione della sicurezza, dei servizi logistici, delle concessioni sul demanio marittimo e lacuale, delle opere infrastrutturali e dei collegamenti con le isole, oltre alla promozione e al sostegno alla mobilità sostenibile e inclusiva.

14: TUTELA DELL'AMBIENTE Si occupa dello studio, programmazione, attuazione e monitoraggio degli interventi per la salvaguardia e valorizzazione dell'ambiente. Include attività normative, promozionali e di educazione ambientale; gestione e controllo della qualità dell'aria, acqua, suolo e risorse idriche; prevenzione e mitigazione dell'inquinamento acustico, elettromagnetico e marino; gestione del ciclo dei rifiuti; tutela paesaggistica, parchi naturali e aree protette; promozione di fonti energetiche rinnovabili; pianificazione idrogeologica e costiera; valutazione dell'impatto ambientale e protezione civile in caso di emergenze naturali o antropiche.

15: SANITA', IGIENE e VETERINARIA Riguarda la programmazione, organizzazione, coordinamento e controllo del sistema sanitario, includendo la prevenzione e tutela della salute, igiene pubblica e veterinaria. Comprende attività normative, programmazione sanitaria, gestione economico-finanziaria, controllo di qualità, gestione del personale sanitario, assistenza sanitaria e socio-sanitaria, farmaceutica e veterinaria. Si occupa della prevenzione di malattie, sicurezza alimentare, igiene ambientale e protezione da rischi sanitari, oltre a garantire il benessere e la gestione degli animali, inclusi quelli d'allevamento, domestici e selvatici.

16: POLITICHE SOCIALI Si occupa della programmazione, organizzazione, coordinamento e controllo delle attività finalizzate a migliorare la qualità della vita, prevenendo e riducendo situazioni di disabilità, disagio e bisogno. Include interventi di sostegno per famiglie, minori, anziani, persone disabili e soggetti in condizioni di particolare vulnerabilità. Cura la gestione del personale socio-assistenziale, la promozione della rete dei servizi sociali, attività normative, di ricerca e monitoraggio, oltre a iniziative di sensibilizzazione e prevenzione. Promuove l'integrazione sociale ed educativa, garantendo supporto e accoglienza per i più fragili.

17: ISTRUZIONE, FORMAZIONE E LAVORO Si occupa della programmazione, organizzazione, coordinamento e monitoraggio di interventi e incentivi volti a promuovere lo sviluppo dell'istruzione, della formazione professionale e del lavoro. Garantisce il diritto allo studio, sostiene la formazione continua e permanente, promuove l'occupazione e la sicurezza sul lavoro, tutela le pari opportunità e coordina i programmi comunitari. Comprende attività normative, di ricerca, monitoraggio e informazione, nonché il supporto tecnico e la gestione di servizi per l'impiego e la formazione, favorendo l'integrazione tra istruzione, formazione tecnica e mercato del lavoro.

18: BENI E ATTIVITA' CULTURALI Si occupa della programmazione, organizzazione, coordinamento e monitoraggio di interventi e incentivi volti alla tutela, conservazione e valorizzazione del patrimonio culturale e delle attività artistiche. Comprende la gestione di musei, biblioteche, archivi, siti archeologici, complessi monumentali e istituti culturali, nonché la promozione di eventi teatrali, musicali, cinematografici e editoriali. Svolge attività normative, di ricerca, catalogazione e monitoraggio, favorendo la diffusione della cultura attraverso progetti, accordi e attività divulgative.
19: ATTIVITA' SPORTIVE E RICREATIVE Comprende la programmazione, il coordinamento e il monitoraggio degli interventi e degli incentivi per promuovere lo sport e le attività ricreative, includendo aspetti normativi, pianificazione, promozione, gestione di eventi, infrastrutture sportive, associazioni di settore e regolamentazione di caccia e pesca.

Prompting Techniques Details

All the prompting techniques contained the description of the classes explained in section 8.1, which will not be repeated here. Again, as the task is performed on Italian documents, all the prompts are in Italian.

The first prompting technique, the *One Call*, simply asked the model to choose one of the possible classes, and showed it an example:

Ti fornirò il testo di un atto amministrativo e il compito è classificarlo tra una delle seguenti 19 classi. Ogni classe è identificata da un numero, un nome e una descrizione. Di seguito trovi l'elenco delle classi:

Classes description, see 8.1

Quando ti fornirò un testo, valuta la somiglianza semantica con ciascuna classe e restituisci il titolo della classe scelta, nel formato:
05: COMMERCIO.
Rispetta rigorosamente questo formato e restituisci esclusivamente il risultato.

Esempio di input:
Full text of a real administrative act

Esempio di output:
19: ATTIVITA' SPORTIVE E RICREATIVE

Ora analizza il seguente documento e restituisci il risultato:
Full text of the document to classify.

The second prompting technique, the *Multiple Calls*, called the model once for every class and asked it to express how related the document was to that class on a scale from 0 to 100:

Il tuo ruolo è di capire se il testo del documento è relazionato alla seguente classe. La risposta dovrà contenere solamente il punteggio di corrlazione tra il documento e la classe (da 0 a 100), e il titolo della classe comprensivo del suo numero identificativo, nella forma 'punteggio - titolo'. Per esempio:
57 - 12: PIANIFICAZIONE TERRITORIALE, URBANISTICA ED EDILIZIA

Classe: *Description of the class under analysis.*

Testo del documento:
Full text of the document to classify.

The third prompting technique, *Two Steps*, gave the model the description of all the 19 classes and an example as in *One Call*, but asked it to assign a correlation score to all the classes and return only the ones that had this correlation above a certain threshold, that we found worked best at 50. Then, on a second call, the model was asked to choose one class among the ones that it returned the first time:

First call: Ti fornirò il testo di un atto amministrativo e il tuo compito è identificare tutte le classi pertinenti tra le seguenti 19 classi. Ogni classe è identificata da un numero, un nome e una descrizione. Di seguito trovi l'elenco completo delle classi:

Classes description, see 8.1

Compito Quando ti fornirò un documento, segui questi passaggi:
Valuta la correlazione semantica tra il contenuto del documento e ciascuna delle 19 classi.
Per ogni classe pertinente, assegna un punteggio di correlazione (da 0 a 100) che indica quanto il contenuto del documento è rilevante per quella classe.
Restituisci solo le classi con una percentuale di correlazione uguale o superiore a THRESHOLD.
Il formato di output deve essere il seguente:
Una riga per ogni classe pertinente.
Ogni riga deve seguire questo formato esatto: [punteggio] - [numero classe]: [nome classe]

Esempio di input:
Full text of a real administrative act

Esempio di output:
95 - 09: COMMERICIO
60 - 13: 13: INFRASTRUTTURE E TRASPORTI

Nota:
Includi solo le classi con correlazione ≥ 50.
Rispetta esattamente il formato indicato. Non aggiungere spiegazioni o commenti. Non cambiare il titolo delle classi, rispetta tutti i caratteri di quello fornito nella descrizione.
Non includere alcuna analisi aggiuntiva oltre a quella richiesta.

Ora analizza il seguente documento e restituisci il risultato:
Full text of the document to classify.

Second Call: Ti fornirò il testo di un atto amministrativo e il tuo compito è identificare la classe a cui esso appartiene. Ogni classe è identificata da un numero, un nome e una descrizione. Di seguito trovi l'elenco completo delle classi:

Description of the classes returned on the first call.

Quando ti fornirò un testo, valuta la somiglianza semantica con ciascuna classe e restituisci il titolo della classe scelta, nel formato: 05: SISTEMA INFORMATIVO.
Rispetta rigorosamente questo formato e restituisci esclusivamente il risultato. Non aggiungere spiegazioni.

Esempio di input:
Full text of a real administrative act

Esempio di output:
19: ATTIVITA' SPORTIVE E RICREATIVE

Ora analizza il seguente documento e restituisci il risultato:
Full text of the document to classify.

The final prompting technique, *Second Chance*, had two different calls, the first one exactly like the One Call technique, the second one told the model that it had previously classified the document in the class resulting from the first call, then showed it all the possible classes again and asked it if it was confident of its choice:

Second Call:
Hai classificato il seguente documento come appartenente alla classe:
Class obtained through the first call.

All the possible classes are:
Classes description, see 8.1.

Sei sicuro della classificazione originale?
Rispondi solamente con il nome della classe più corretta.

Esempio di output:
19: ATTIVITA' SPORTIVE E RICREATIVE

Testo del documento:
Full text of the document to classify.

Confusion Matrices of the Different Prompting Techniques

Images 8.13, 8.14a, 8.14b, 8.15a, 8.15b show the confusion matrices of the different prompting techniques described in Section 6.1.1 alongside those of Gemini, Image 8.16a, and Llama, Image 8.16b.

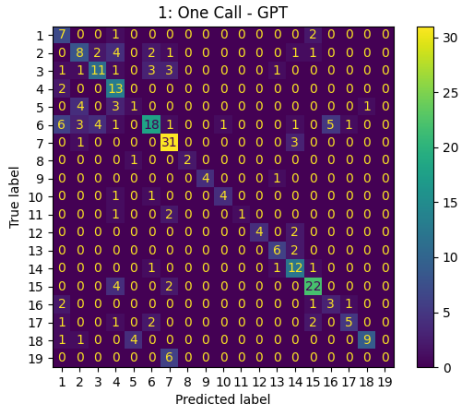
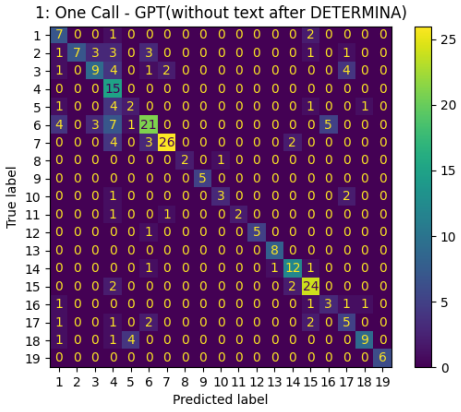
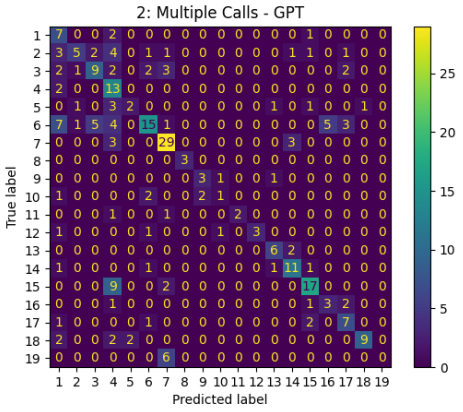


Figure 8.13: Confusion matrix of prompting technique 1: One Call using GPT, with all text.

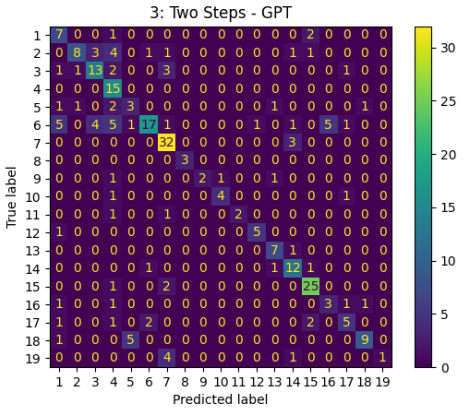
All the following matrices refer to the prompting techniques without giving the model the text after the keyword *DETERMINA*.



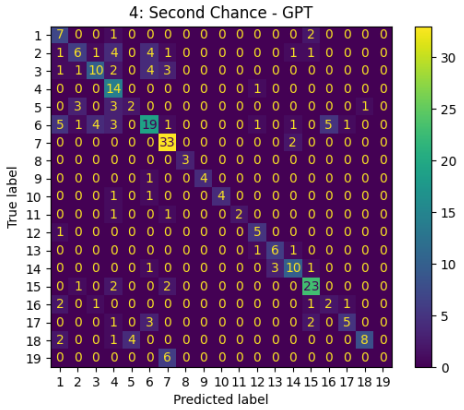
(a) Confusion matrix of prompting technique 1: One Call using GPT.



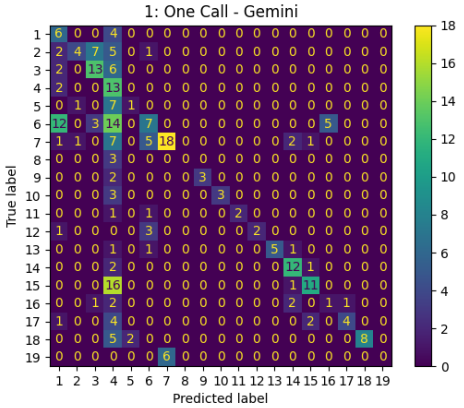
(b) Confusion matrix of prompting technique 2: Multiple Calls using GPT.



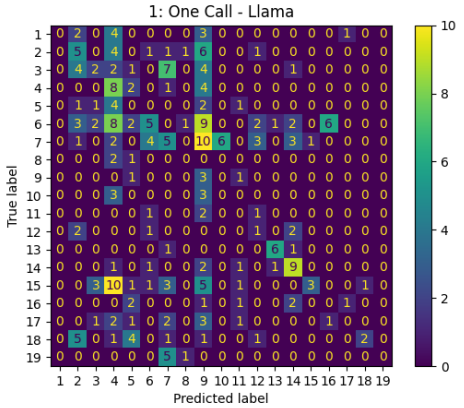
(a) Confusion matrix of prompting technique 3: Two Steps using GPT.



(b) Confusion matrix of prompting technique 4: Second Chance using GPT.



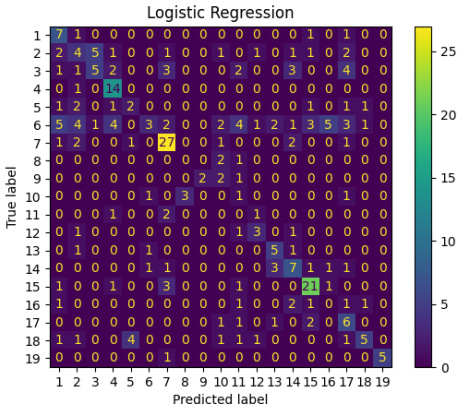
(a) Confusion matrix of prompting technique 1: One Call using Gemini.



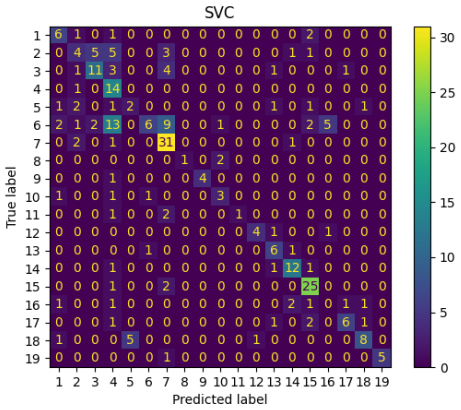
(b) Confusion matrix of prompting technique 1: One Call using Llama.

LR and SVC results for Classification of Administrative Acts

Images 8.17a and 8.17b show the confusion matrices obtained training Logistic Regression and SVC on the silver dataset of administrative acts from Emilia Romagna and testing on the gold dataset, as described in Section 6.1.3.



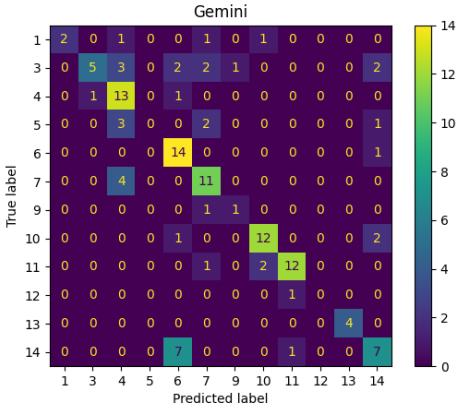
(a) Confusion matrix of the Logistic Regression model.



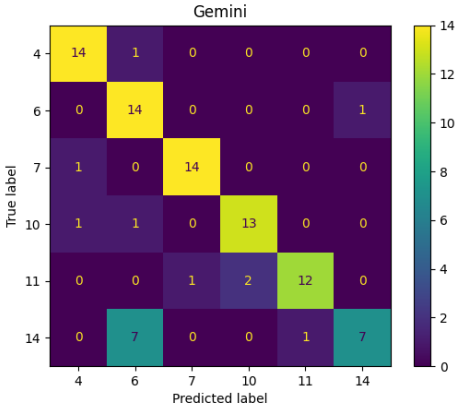
(b) Confusion matrix of the SVC model.

Gemini results for the classification of Bulgarian Case Law documents

Image 8.18a shows the confusion matrix of the zero-shot classification performance of Gemini on all the classes of Bulgarian Supreme Court of Cassation documents, while Image 8.18b shows Gemini's performance on only a portion of the classes.



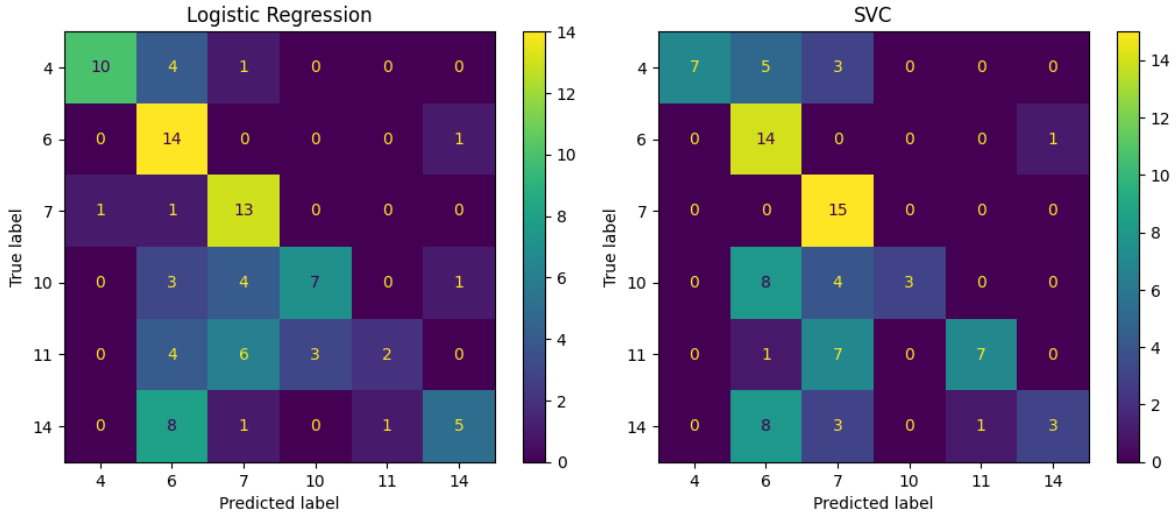
(a) Confusion matrix of Gemini on all the 14 possible classes.



(b) Confusion matrix of Gemini on a subset of classes.

LR and SVC results for classification of Bulgarian Case Law documents

Images 8.19a and 8.19b show the confusion matrices obtained training Logistic Regression and SVC on a subset of the silver dataset of Bulgarian Supreme Court of Cassation documents and testing on the gold dataset, as described in Section 6.2.4.



(a) Confusion matrix of the Logistic Regression model.

(b) Confusion matrix of the SVC model.

Funded by the European Union - Next Generation EU, Mission 4, Component 2, Investment 3.3 (Ministerial Decree 117/2023) CUP 38-412-03-DOT1303952-254