Prompting Large Language Models for Italian Clinical Reports: A Benchmark Study

Livia Lilli[1,2], **Carlotta Masciocchi**[1], **Antonio Marchetti**[1], **Giovanni Arcuri**[1], **Stefano Patarnello**[1]
[1] Fondazione Policlinico Universitario Agostino Gemelli IRCCS
[2] Catholic University of the Sacred Heart of Rome
livia.lilli@policlinicogemelli.it

Abstract

Large Language Models (LLMs) have significantly impacted medical Natural Language Processing (NLP), enabling automated information extraction from unstructured clinical texts. However, selecting the most suitable approach requires careful evaluation of different model architectures, such as generative LLMs and BERT-based models, along with appropriate adaptation strategies, including prompting techniques, or fine-tuning. Several studies explored different LLM implementations, highlighting their effectiveness in medical domain, including complex diagnostics patterns as for example in rheumatology. However, their application to Italian remains limited, serving as a key example of the broader gap in non-English language research. In this study, we present a task-specific benchmark analysis comparing generative LLMs and BERT-based models, on real-world Italian clinical reports. We evaluated zero-shot prompting, in-context learning (ICL), and fine-tuning across eight diagnostic categories in the rheumatology area. Results show that ICL improves performance over zero-shot-prompting, particularly for Mixtral and Gemma models. Overall, BERT fine-tuning present the highest performance, while ICL outperforms BERT in specific diagnoses, such as renal and systemic, suggesting that prompting can be a potential alternative when labeled data is scarce.

1 Introduction

Recent advancements in Large Language Models (LLMs) have significantly impacted medical Natural Language Processing (NLP), enabling the extraction of structured information from unstructured clinical texts with increasing accuracy. Transformer-based architectures, such as BERT-based models and generative LLMs, have demonstrated strong potential in clinical text classification, named entity recognition, and medical concept extraction. However, selecting the most suitable model for a given task requires careful consideration of both model architecture and adaptation strategy, as different approaches offer varying levels of performance, efficiency, and practical feasibility.

LLMs, particularly generative architectures, can be adapted through zero-shot prompting (Sivarajkumar et al., 2024), where the model relies solely on its pre-trained knowledge, or in-context learning (ICL) (Liu et al., 2024), where domain-specific context is provided within the prompt. More advanced strategies include instruction fine-tuning (Tran et al., 2024; Li et al., 2024b), which refines the model's alignment with task-specific instructions. BERT-based models (Devlin et al., 2019), following a discriminative approach, typically require fine-tuning through supervised learning, though they can also be applied in Natural Language Inference (NLI) frameworks or used in few-shot and zero-shot settings by leveraging pre-trained embeddings. In all cases, pretraining on large domain-specific corpora can further enhance performance, though it remains computationally expensive and data-intensive.

In this work, we present a task-specific benchmark analysis tailored to a real-world clinical scenario, focusing on the necessity of extracting structured information from Italian clinical notes, in a real-world hospital setting. Our study evaluates generative LLMs in two different prompting strategies: zero-shot prompting, where the model relies solely on its pre-trained knowledge, and ICL, where additional domain-specific context is provided to guide the extraction process. To establish a strong comparative baseline, we also assess fine-tuned BERT-based models, which have traditionally been used for medical information extraction tasks (Lee et al., 2020; Muizelaar et al., 2024; Yang et al., 2024).

Our evaluation is based on a very general use case, which is the detection of complex diagnoses

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 190–200

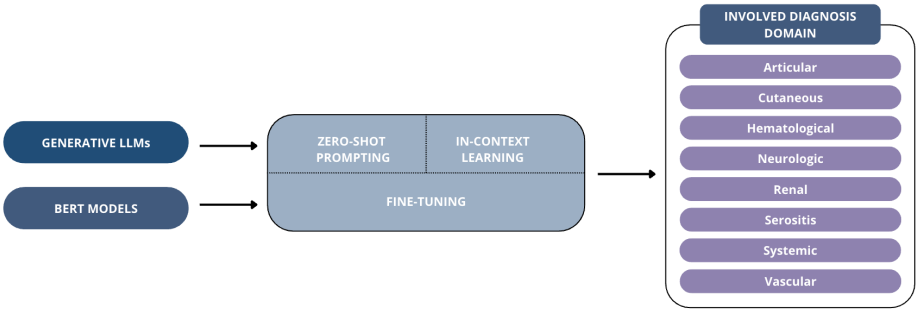

Figure 1: Study Framework: task-specific benchmark analysis comparing LLMs using zero-shot and in-context learning (ICL) strategies against fine-tuned BERT-based models, in an information extraction task.

in medical notes. The example we used covers the rheumatology domain, where very often a disease can impact several domains, each corresponding to a particular organ or system involvement (Figure 1). The use case of diagnosis extraction allows us to systematically compare generative LLMs and fine-tuned models, analyzing their strengths, limitations, and potential applications in real-world clinical workflows. By conducting this study in a practical hospital setting, we aim to provide insights into the feasibility of integrating LLM-based NLP solutions for automated information extraction in clinical practice. This process can support the development of decision-support tools and enable the creation of research datasets for predictive analytics, ultimately enhancing both clinical decision-making and medical research.

2 Background

Natural Language Processing (NLP) has gained increasing attention in medical application, with studies exploring its potential for extracting meaningful clinical insights from unstructured medical texts. A systematic review by Omar et al. (2024) provides a comprehensive analysis of NLP applications specifically for the reumathology domain, covering various techniques used to process electronic health records (EHRs), PubMed abstracts, FAQ and exams' questions for diseases such as rheumatoid arthritis (RA), gout, and systemic lupus erythematosus (SLE). Among the identified works, Li et al. (2022) explores named entity recognition (NER) in RA clinical notes, leveraging a BERT model enhanced with BiLSTM and CRF layers, achieving promising results in medical entity extraction. In the study of Osborne et al. (2021) NLP is used for gout flare detection, developing a fine-tuned BERT classifier based on annotated Emergency Department (ED) chief complaint notes, demonstrating that chief complaints alone are highly predictive of gout flares. Expanding on this approach, Oliveira et al. (2024) compares traditional NLP methods (e.g., tf-idf) with domain-specific LLMs, distinguishing between generative and discriminative models. Their study shows that generative models used as feature extractors can enhance performance when integrated with an SVM classifier, suggesting a hybrid approach for clinical text classification.

Focusing specifically on SLE and the Italian language, Lilli et al. (2024a) investigates the adaptation of BERT-based models for the extraction of Lupus-related diagnoses, symptoms, and treatments, demonstrating the feasibility of transformer-based NLP approaches in non-English medical corpora. Lilli et al. (2024b) also presents an NLP pipeline that integrates regular expression-based extraction with BERT-based topic detection, improving the structured identification of Lupus-related clinical features from Italian medical texts.

Beyond disease-specific applications, broader research has investigated the effectiveness of LLMs and BERT-based models in medical NLP tasks. Zhang et al. (2024) evaluates prompt engineering versus fine-tuning for clinical note classification, using metastatic cancer identification as a benchmark task. Their findings indicate that GPT-4 with structured prompts outperforms fine-tuned BERT-based models, suggesting that prompting can be an effective alternative to model retraining in clinical NLP. Meanwhile, Savage et al. (2024) examines whether LLMs can emulate clinical reasoning by structuring prompts to reflect differential diagnosis formation, intuitive reasoning, analytical reasoning, and Bayesian inference. Their results suggest that LLMs can provide interpretable rationales without compromising diagnostic accuracy, addressing the

"black box" issue that limits trust in AI-driven medical applications.

More recent studies have also advanced the understanding of prompting strategies in clinical NLP. Naguib et al. (2024) conducted a multilingual evaluation of few-shot prompting for clinical NER, showing that masked language models often outperform generative models, particularly in low-resource settings. Similarly, Nagar et al. (2024) benchmarked various prompting and retrieval strategies across structured biomedical tasks, highlighting the limitations of reasoning-augmented methods like Chain-of-Thought and RAG, especially for classification and NER. Hu et al. (2024) proposed a prompt engineering framework for GPT models in clinical NER, demonstrating that structured, task-specific prompting can substantially improve performance.

All the above studies highlight the evolution of NLP techniques in medical applications, and the increasing role of LLMs in replacing or complementing traditional fine-tuned models for clinical text analysis, classification, and decision support.

3 Method

3.1 Dataset

The dataset used in this study consists of a collection of outpatient visit reports written in Italian language, related to patients with a SLE diagnosis and treated in the Reumatology department of a real-world hospital. The outcome of the information task was to identify eight different types of diagnoses based on the specific organ or system involvement. The categories considered are: Articular, Cutaneous, Hematologic, Neurologic, Renal, Systemic, Serositis, and Vascular.

3.2 Generative Modeling

For the generative LLM experiments, we employed a set of open-source language models, either multilingual or specifically trained for the Italian language, leveraging the Ollama framework to optimize computational efficiency. The models were then accessed through the Ollama Python library[1], utilizing its `generate` function to process and analyze clinical texts. This approach allowed us to efficiently execute inference without the need for fine-tuning, making it a scalable and adaptable solution for medical NLP tasks. Input reports were preprocessed and analyzed at the paragraph level rather than as full documents. This approach was adopted to reduce text length, enabling a more focused and efficient processing of clinical information. At the end of the processing pipeline, a logical OR operation was applied to aggregate paragraph-level classifications into a final diagnosis at the Electronic Health Record (EHR) level. This means that if any paragraph within a patient's report indicated the presence of a specific Lupus subtype, that classification was assigned to the entire EHR. For both the zero-shot and in-context learning setups, we leveraged ChatGPT-4o (Hurst et al., 2024; Achiam et al., 2023) to generate appropriate prompts, ensuring well-structured and consistent instructions tailored for clinical information extraction. In both cases, the prompt was in English and designed to return a structured binary output (1 for presence, 0 for absence) for each Lupus category independently. However, the models did not always comply with this format, often including additional explanations or justifications alongside the binary response. To provide a standardized output, we applied a regular expression (regex) filter to isolate and extract the binary classification for each category separately, ensuring consistency in the final results. To improve the robustness of this approach, we manually reviewed a sample of LLM outputs to identify common patterns for the development of regex rules. This step helped us reduce misclassifications caused by unexpected output formats or embedded rationales.

3.2.1 Zero-Shot Prompting

In the zero-shot prompting setup, the models were prompted without any additional contextual guidance or predefined medical terms. The prompt structure followed a direct query format, instructing the model to determine the presence of a Lupus diagnosis based on the involvement of a specific organ or system. The exact prompt used was:

```
Given the following Italian medical report,
return "1" if there is evidence of lupus with
{category} involvement, otherwise, return "0".

Report:
{text}
```

By relying solely on the model's pre-trained knowledge, this approach aimed to evaluate the intrinsic capability of generative LLMs to extract structured medical information without external lexical or contextual augmentation.

[1] https://github.com/ollama/ollama-python

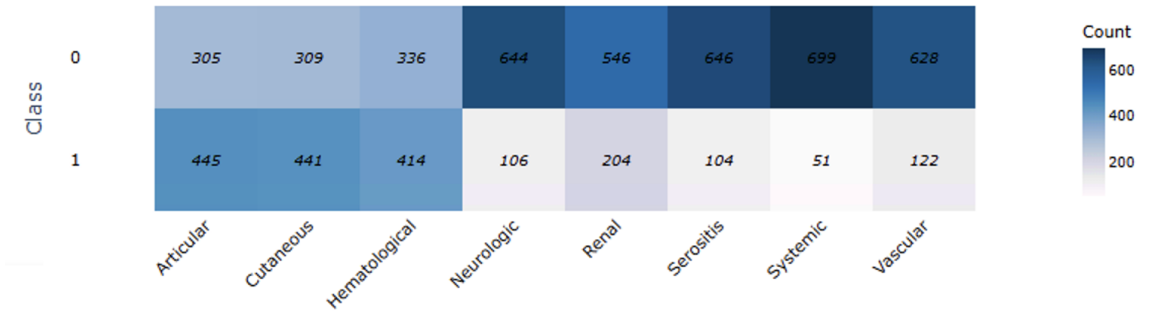

Figure 2: Dataset Composition.

3.2.2 In-Context Learning

In the in-context learning setup, we provided additional domain-specific context by including an Italian dictionary of medically relevant terms related to each category. These terms consisted of synonyms and alternative expressions commonly used in clinical texts to describe the specific type of organ or system involvement. By integrating this lexical knowledge directly into the prompt, we aimed to guide the model toward more accurate information extraction while still leveraging its generative capabilities. Unlike the zero-shot prompting, the following prompt structure was used to incorporate the dictionary of terms:

```
Given the following Italian medical report,
return "1" if there is evidence of lupus with
{category} involvement, otherwise, return "0".
To determine this, check if the report contains
relevant terms associated with {category}
involvement. Below there is a list of medically
relevant terms that indicate {category} involvement:

Relevant terms:
{list_of_terms}

Report:
{text}
```

This setup allowed the model to leverage both its pre-trained knowledge and the medical terminology provided, creating a context-augmented approach that aimed to improve classification accuracy. The list of medically relevant terms used in the ICL prompts is derived from a domain-specific dictionary originally developed for a rule-based information extraction system, as described in our previous work (Lilli et al., 2024b). In that study, the dictionary served as the foundation for a classifier based on pattern-matching within clinical texts. Further details about the dictionary of terms used in the prompt is provided in Appendix A.

3.3 BERT-based Fine-Tuning

To compare the performance of generative language models with fine-tuned approaches, we also included results from a set of fine-tuned BERT-based models, where each Lupus category was treated as an independent binary classification task, with separate classifiers trained for each type of organ involvement. The experimental setup and classification approach are consistent with a prior benchmark study on fine-tuned BERT models (Lilli et al., 2024a), ensuring a direct comparison with the LLM experiments, where a training set of 1000 labelled texts was used.

As in the generative LLM setup, the BERT fine-tuning process followed a paragraph-level approach, respecting the token limit constraints of BERT-based architectures. At inference, each paragraph was classified independently, with the final Electronic Health Record (EHR) classification determined using a logical OR operation.

Additionally, the evaluation set remained the same across all techniques, ensuring a fair and consistent comparison between fine-tuned BERT models and generative LLMs. This methodological alignment allowed us to analyze their relative strengths and limitations under identical conditions.

4 Experiments

4.1 Data

The dataset used for the study evaluation consists of 790 outpatient visit reports, collected from the SLE Data Mart of an Italian hospital. To facilitate processing and improve classification efficiency, in all the experiments each report was segmented at the paragraph level, resulting in a total of 6,024 paragraphs. On average, each paragraph contains 111.5 BERT tokens and 303.7 characters. This seg-

Experiment	Model	Articular	Cutaneous	Hematologic	Neurologic	Renal	Serositis	Systemic	Vascular	Overall
Zero-Shot Prompting	Gemma3-1B	0.00	0.22	0.27	0.08	0.18	0.08	**0.48**	0.04	0.17
	Gemma3-4B	**0.80**	**0.80**	0.78	0.30	0.53	0.29	0.14	0.30	0.49
	Llama3.2-1B	0.62	0.60	0.65	0.28	0.42	0.22	0.14	0.25	0.40
	Llama3.2-3B	0.76	0.76	0.76	0.38	0.55	0.30	0.16	**0.36**	0.50
	Mistral-7B	0.77	0.70	**0.80**	**0.40**	0.64	0.34	0.15	0.33	0.52
	Mixtral-8x7B	0.78	0.74	0.79	0.34	**0.69**	**0.39**	0.16	0.32	**0.53**
In-Context Learning	Gemma3-1B	0.41	0.44	0.34	0.26	0.18	0.60	**0.52**	0.16	0.36
	Gemma3-4B	0.84	**0.88**	0.91	0.76	0.77	0.71	0.17	0.44	0.68
	Llama3.2-1B	0.67	0.67	0.66	0.24	0.41	0.25	0.13	0.27	0.41
	Llama3.2-3B	0.74	0.73	0.71	0.25	0.43	0.24	0.13	0.28	0.44
	Mistral-7B	0.86	0.61	0.87	0.73	0.88	0.77	0.31	0.37	0.68
	Mixtral-8x7B	**0.91**	0.73	**0.95**	**0.83**	**0.96**	**0.80**	0.23	**0.52**	**0.74**
BERT-based Fine-Tuning	Alberto	0.90	0.87	**0.98**	0.86	**0.94**	0.81	**0.29**	**0.69**	**0.79**
	Albert2	0.85	0.80	0.96	0.57	0.85	0.65	0.28	0.55	0.69
	Albert1	**0.92**	0.81	0.94	0.86	0.92	0.51	0.07	0.58	0.70
	Biobit	**0.92**	0.88	0.93	0.81	0.85	**0.87**	0.12	0.63	0.75
	Medbit	0.83	**0.92**	0.96	0.79	0.90	0.66	0.13	0.61	0.73
	Medbit-plus	**0.92**	0.90	0.90	**0.88**	0.85	0.72	0.07	0.63	0.73

Table 1: Comparison of Generative LLMs and BERT-Based Models Across Different Experimental Setups (Zero-Shot Prompting, In-Context Learning, and BERT-based Fine-Tuning), in terms of F1-Score.

mentation ensures that text segments remain within the acceptable token limits of BERT-based models, preserving sufficient clinical context for classification. Additionally, this approach is beneficial for generative LLMs, as it enables them to process shorter and more concise text inputs, optimizing computational efficiency and response accuracy. The dataset includes eight distinct types of Lupus diagnoses, each corresponding to a specific organ or system involvement. Since multiple categories can co-occur in the same report, a single document may be associated with more than one diagnosis. For privacy reasons, we can't report practical examples of the dataset, but we provide an overview of its composition in Figure 2.

4.2 Generative LLMs

For the generative experiments, we tested a range of open-source language models using the Ollama framework to ensure efficient inference. The models evaluated included Llama 3.2 (1B and 3B parameters), Gemma 3 (1B and 4B parameters), Mixtral (8x7B) and Mistral (7B). Each model was evaluated in both zero-shot prompting and ICL setups, on the SLE information extraction task.

Llama 3.2, developed by Meta (Grattafiori et al., 2024), is an optimized version of the Llama family, known for its improved efficiency and multilingual capabilities. The 1B and 3B parameter versions provide a balance between computational cost and performance, making them suitable for real-world scenarios. Gemma 3, released by Google DeepMind (Team et al., 2024), is a lightweight transformer-based model optimized for low-resource settings while maintaining strong reasoning abilities. The 1B model is designed for efficiency, whereas the 4B version offers enhanced performance with increased computational requirements. Mixtral, a mixture-of-experts model from Mistral AI (Jiang et al., 2024), activates only two out of eight expert networks per inference, allowing for improved efficiency while retaining strong language understanding. Finally, Mistral 7B (also from Mistral AI (Jiang et al., 2023)) is a dense transformer model known for its superior performance compared to similarly sized models, making it a potential alternative to Llama and Gemma for various NLP tasks. By selecting models with different architectures, sizes, and capabilities, we ensured a comprehensive evaluation of generative approaches for NLP in medical domain. Table 1 shows performances in terms of F1-Score metric for the zero-shot prompting and the in-context learning scenarios, respectively.

4.3 BERT-based Models

The BERT-based fine-tuning was performed using the PyTorch Trainer from the Hugging Face Transformers library (Wolf et al., 2020), running for 10 epochs (for further implementation details, see Appendix B). The models considered in this study include BioBIT3, MedBIT4, MedBIT-r3-plus5, AlBERTo, and two base versions of ALBERT.

BioBIT3, MedBIT4, and MedBIT-r3-plus5, developed by Buonocore et al. (2023), are BERT models pretrained on Italian biomedical corpora, making them particularly suitable for clinical NLP tasks. AlBERTo, originally proposed by Polignano

et al. (2019), is an Italian-adapted version of ALBERT, trained on Italian tweets. In addition, we included two base versions of ALBERT Lan et al. (2019), which serve as the foundation of AlBERTo. Table 1 shows F1-Score metric values for the BERT-based experiments.

4.4 Results and Discussion

The results of the zero-shot and in-context learning (ICL) prompting experiments, compared to BERT fine-tuning, are presented in Table 1. For each scenario, the table reports the F1-scores of all tested models across the eight categories, along with the overall F1-score, calculated as the mean value. To better interpret these results, we structure our analysis into two key perspectives. First, we provide an overall comparison of performance across different methods (zero-shot prompting, in-context learning, and fine-tuned BERT models) to assess their general effectiveness. Second, we examine model-specific performance patterns across different disease categories, identifying strengths and limitations in extracting various diagnostic domains.

Regarding overall model performance across different disease categories, BERT-based classification models achieve the highest scores, with Alberto obtaining the best average F1-score of 0.79. However, it is noteworthy that even a limited degree of adaptation through In-Context Learning (ICL) significantly improves LLM performance. Mixtral-8x7B, with an average F1-score of 0.74, performs comparably to the best BERT-based models, demonstrating the effectiveness of ICL in enhancing generative models for structured information extraction. In contrast, Zero-Shot Prompting shows the weakest performance, with Mixtral-8x7B achieving the highest overall F1-score at just 0.53. This performance gap is likely due to the lack of contextual guidance, which makes it more challenging for the model to differentiate between diagnostic categories. In the absence of domain-specific cues, semantic differences across diagnoses reduce the model's discriminative power, leading to lower classification accuracy.

Moving to an in-depth analysis of performance across different diagnostic categories, the zero-shot setting reveals notable variations among models. Mixtral-8x7B, with the highest overall performance (F1-Score=0.53), specifically outperforms the other models in Renal (F1-Score=0.69) and Serositis (F1-Score=0.39) diagnoses. While Mistral-7B, with a slightly lower F1-Score of 0.52, presents the highest F1-Score in Hematologic (0.80) and Neurologic (0.40) categories. Meanwhile, Gemma3-4B, with the best F1-scores in the Articular and Cutaneous categories (0.80), shows an overall F1 performance near to Mixtral-8x7B and Mistral7B, equal to 0.49. In general, zero-shot performance is particularly weak for Neurologic, Renal, Serositis, Systemic, and Vascular diagnoses, with F1-scores ranging from 0.36 for Vascular (with Llama3.2-3b), to 0.69 for Renal (with Mixtral-8x7b).

In the ICL setting, Mixtral-8x7B achieves the highest scores on most of the categories, with the highest in the Renal, with a F1 value equal to 0.96. However, Gemma3-4B and Gemma3-1B outperform Mixtral-7B in two specific cases: Cutaneous (F1-Score=0.88) and Systemic (F1-Score=0.52). A particular improvement is observed in Neurologic, Renal, and Serositis diagnoses, where zero-shot prompting had shown extremely weak performance: with ICL, these categories experience a substantial boost, with Mixtral-8x7B achieving the highest scores, ranging from 0.80 for Serositis to 0.96 for Renal. On the other hand, classification for Systemic and Vascular categories remains weak, with the best performances achieved by Gemma3-1B for Systemic and Mixtral-8x7B for Vascular (F1-score = 0.52).

The results of both zero-shot and in-context learning (ICL) experiments highlight the significant role that contextual adaptation plays in enhancing generative models' performance. While the previous analysis examined each approach across different diagnostic categories, it is equally important to assess how ICL compares directly to zero-shot prompting across models and disease types. In general, moving from zero-shot to in-context learning (ICL) mostly leads to improved performance, as evident in Figure 3. Each bar plot in the figure represents the F1-score of different models across the eight categories, with the maximum F1-score from either the zero-shot or ICL scenario displayed. The colored margins, green and red, indicate the difference between the two approaches. A green margin means a positive difference, meaning ICL outperforms zero-shot prompting, enhancing information extraction. Conversely, a red margin indicates cases where zero-shot prompting achieved better results. From the figure, the majority of cases show an improvement with ICL, particularly for the Gemma and Mistral models. For instance, in the Serositis category, the F1-score of the Gemma3-4B model increases from 0.29 in the zero-shot setting

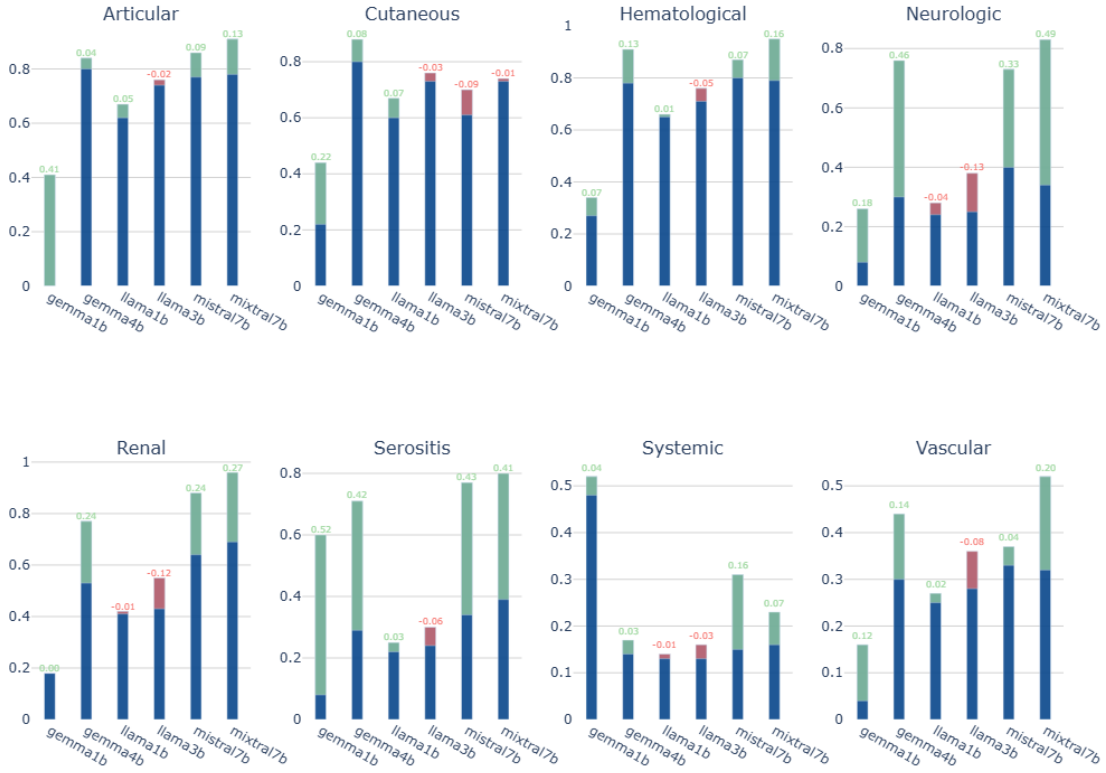

Figure 3: Comparison of Zero-Shot Prompting versus In-Context Learning (ICL) F1-Scores across different diagnoses. The bars represent the maximum score between the two methods. The green and red margins indicate the effect of ICL: green for improvements and red for declines compared to Zero-Shot Prompting.

to 0.71 with ICL, as shown by the green margin of 0.42. Similarly, in the Neurologic diagnosis, the Mixtral-8x7B model returns F1-Score values of 0.34 and 0.83 in the zero-shot and ICL scenarios respectively, with a final margin of improvement equal to 0.49. However, there are instances where ICL does not improve performance: the presence of red margins in at least one model for all categories suggests that semantic complexity alone is not the primary cause. Instead, it appears that certain models, particularly Llama3.2-3b, consistently perform better in the zero-shot scenario, and also Llama3.2-1B frequently shows negative margins. This indicates that for the Llama models, the type of contextual information introduced in ICL does not provide additional knowledge but instead has a confounding effect, hindering information detection.

To fully assess the effectiveness of prompting strategies, we compared them with a fully supervised fine-tuned approach of BERT-based models, which serve as a benchmark for structured information extraction. As already reported, BERT fine-tuning achieves the highest overall performance, with the Alberto model obtaining the highest F1-score of 0.79. In terms of individual categories, BERT models excel particularly in Hematologic diagnosis, where Alberto reaches the highest F1-score equal to 0.98. However, not all categories benefit the most from BERT fine-tuning. Some achieve better performance in the ICL scenario, such as Renal diagnosis, where Mixtral-8x7B reaches a 0.96 F1-Score, and Systemic diagnosis, where Gemma3-1B achieves a 0.52 F1-Score. These cases suggest that while BERT fine-tuning is generally effective, ICL can provide better results for specific types of medical information extraction.

Overall, ICL improves performance over zero-shot prompting, though some models, like Llama models, struggle with added context. BERT fine-tuning remains the most reliable approach for this Italian use case, achieving the highest scores. However, prompting is a viable alternative, as it allows

adaptation with minimal data and no dedicated training, making it useful when resources are limited.

5 Conclusions

This study provides a comparative analysis of generative LLMs and fine-tuned BERT models for Italian clinical NLP, focusing on the extraction of diagnostic patterns within an outpatient setting. Our results demonstrate that while ICL significantly enhances generative models' performance over zero-shot prompting, fine-tuned BERT-based models still achieve the best overall results, providing structured and reliable classification solutions. However, ICL performances show that in-context adaptation techniques have great potential for iterative improvement. This is also confirmed by the results of this paper, where certain diagnostic categories, such as renal involvement, show better performance with ICL, indicating that supervised prompting can effectively overcome certain semantics complexities.

Beyond performance, model selection in healthcare applications must also consider privacy, data protection, and control on adaptation. For this reason we believe that a study focused on the comparison of open-source types of models provides a new perspective to complement GPT-based works, which are largely explored in current literature (Li et al., 2024a).

Future work should explore larger generative models, which may offer insight into the upper-bound performance achievable through prompting strategies alone. Additionally, future studies should conduct a more in-depth analysis of computational costs and trade-offs, particularly when considering prompting-based methods versus full fine-tuning, to guide practical decisions in clinical deployment scenarios.

By conducting this study in a real-world hospital setting, we aim to provide insights into the feasibility of integrating LLM-based NLP solutions for automated clinical information extraction. This could aid in the development of decision-support tools, facilitate the creation of research datasets for predictive analytics, and ultimately improve both clinical decision-making and medical research. Furthermore, by focusing on Italian clinical texts, this study expands NLP applications beyond English-language datasets, addressing the need for real-world solutions in underrepresented languages.

Limitations

While generative models show potential for medical information extraction, they do not always produce structured responses, requiring post-processing. We extracted binary classifications using regular expressions (regex), but this method can be imprecise, making BERT-based architectures more reliable for structured tasks. Additionally, due to the constraints of a real-world clinical setting in terms of computing resources, lighter versions of the models have been implemented. Future work could explore larger versions of Gemma and Llama running on more powerful computing environments, to achieve potential performance gains. Furthermore, in-context learning (ICL) proves effective for an initial adaptation, but its performance could be enhanced by incorporating labeled examples alongside the current dictionary. Further research should explore alternative adaptation techniques, such as instruction-tuning or a massive fine-tuning, to better compare different strategies for optimizing medical NLP models.

Ethics Statement

The use of data for this study has been implemented in full compliance with ethics and GDPR requirements. Specifically, data usage has been approved by the Ethics Committee of our hospital to conduct the presented research and the de-identification of sensitive data has been performed. Approval protocol number from the relevant Ethics Committee can be provided on request.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. 2023. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dingxin Hu, Xuanyu Zhang, Xingyue Zhang, Yiyang Li, Dongsheng Chen, Marina Litvak, Natalia Vanetik, Qing Yang, Dongliang Xu, Yanquan Zhou, Lei Li, Yuze Li, and Yingqi Zhu. 2024. Improving factual consistency in abstractive summarization with sentence structure pruning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8792–8803, Torino, Italia. ELRA and ICCL.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024a. Chatgpt in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, 245:108013.

Meiting Li, Feifei Liu, Ran Zhang, Yi Qin, Dongping Gao, et al. 2022. Model-based clinical note entity recognition for rheumatoid arthritis using bidirectional encoder representation from transformers. *Quantitative Imaging in Medicine and Surgery*, 12(1):184.

Rumeng Li, Xun Wang, and Hong Yu. 2024b. LlamaCare: An instruction fine-tuned large language model for clinical NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10632–10641, Torino, Italia. ELRA and ICCL.

Livia Lilli, Laura Antenucci, Augusta Ortolan, Silvia Laura Bosello, Maria Antonietta D'agostino, Stefano Patarnello, Carlotta Masciocchi, and Jacopo Lenkowicz. 2024a. Lupus alberto: A transformer-based approach for SLE information extraction from Italian clinical reports. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 510–516, Pisa, Italy. CEUR Workshop Proceedings.

Livia Lilli, Silvia Laura Bosello, Laura Antenucci, Stefano Patarnello, Augusta Ortolan, Jacopo Lenkowicz, Marco Gorini, Gabriella Castellino, Alfredo Cesario, Maria Antonietta D'Agostino, et al. 2024b. A comprehensive natural language processing pipeline for the chronic lupus disease. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, pages 909–913. IOS Press.

Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. 2024. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8721–8730.

Hielke Muizelaar, Marcel Haas, Koert van Dortmont, Peter van der Putten, and Marco Spruit. 2024. Extracting patient lifestyle characteristics from dutch clinical text with bert models. *BMC medical informatics and decision making*, 24(1):151.

Aishik Nagar, Viktor Schlegel, Thanh-Tung Nguyen, Hao Li, Yuping Wu, Kuluhan Binici, and Stefan Winkler. 2024. Llms are not zero-shot reasoners for biomedical information extraction. *CoRR*.

Marco Naguib, Xavier Tannier, and Aurelie Neveol. 2024. Few-shot clinical entity recognition in english, french and spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852.

Lucas Lopes Oliveira, Xiaorui Jiang, Aryalakshmi Nellippillipathil Babu, Poonam Karajagi, and Alireza Daneshkhah. 2024. Effective natural language processing algorithms for early alerts of gout flares from chief complaints. *Forecasting*, 6(1):224–238.

Mahmud Omar, Mohammad E Naffaa, Benjamin S Glicksberg, Hagar Reuveni, Girish N Nadkarni, and Eyal Klang. 2024. Advancing rheumatology with natural language processing: insights and prospects from a systematic review. *Rheumatology Advances in Practice*, 8(4):rkae120.

John D Osborne, James S Booth, Tobias O'Leary, Amy Mudano, Giovanna Rosas, Phillip J Foster, Kenneth G Saag, and Maria I Danila. 2021. Identification of gout flares in chief complaint text using natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2020, page 973.

Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, Valerio Basile, et al. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CEUR workshop proceedings*, volume 2481, pages 1–6. CEUR.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Bioinstruct: instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association*, 31(9):1821–1832.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yumeng Yang, Soumya Jayaraj, Ethan Ludmir, and Kirk Roberts. 2024. Text classification of cancer clinical trial eligibility criteria. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1304.

Xiaodan Zhang, Nabasmita Talukdar, Sandeep Vemulapalli, Sumyeong Ahn, Jiankun Wang, Han Meng, Sardar Mehtab Bin Murtaza, Dmitry Leshchiner, Aakash Ajay Dave, Dimitri F Joseph, et al. 2024. Comparison of prompt engineering and fine-tuning strategies in large language models in the classification of clinical notes. *AMIA Summits on Translational Science Proceedings*, 2024:478.

A In-Context Learning Dictionary

For the in-context learning experiments, we used a dictionary of terms covering each category of diagnosis to be extracted. This aimed to give additional context to the model, making the clinical concepts more understandable. Table 2 details the terms used for each category, reported in their original Italian language.

B BERT-based Fine-Tuning Setup

The fine-tuning was implemented using the PyTorch Trainer[2] of the Hugging Face Transformers library (Wolf et al., 2020), leveraging a desktop GPU Nvidia RTX 5000 Graphics Processing with 16GB of RAM. The 20% of training set was used as `eval_dataset`, while the remaining was employed as `train_dataset`. The learning rate was set to 2e-5, the batch size to 16, and the weight decay to 0.01.

[2]https://huggingface.co/docs/transformers/main/en/training

Category	Terms
Articular	articolare, artralgia, artrite, artrosica, gonartrite, jaccoud, miosite, monoartrite, oligoartrite, osteartrosi, osteoarticolare, poliartrite, polimiosite, rhupus, spondiloartrite.
Cutaneous	afta, aftosi, alopecia, cutaneo, discoide, eczematoso, effluvium capillorum, eritema, eritemato-crostosa, eritemato-desquamativa, eritemato-papulare, eritemato-papulosa, pomfo, fotosensibilità, gottron, led, muco-cutaneo, mucocutaneo, papula, percutaneo, perdita di capelli, porpora.
Hematologic	anemia, anemia emolitica, disturbo della coagulazione, ematico, ematologico, leucolinfopenia, leucopenia, linfopenia, neutropenia, pancitopenia, piastrinopenia.
Neurologic	cerebellare, cerebrale, encefalite, epilettico, epilessia, ictus, mononeurite, multineuropatia, neurite, neurologico, neuropatia, polineuropatia, snc, tia.
Renal	glomerulonefrite, irc, nefrite, nefritemembranosa, nefrosi, renale.
Serositis	ascite, miocardite, pericardite, peritonite, pleurite, pleuro-parenchimale, pleuro-polmonare, pleuropericardite, polmonare, polisierosite, sierosite, sierositico.
Systemic	febbre, astenia, linfoadenopatia, linfoadenite, mialgia, febbricola, linfoadenomegalia, polimialgia
Vascular	acrocianosi, alveolite emorragica, embolia, embolia polmonare, ep, fdr, ischemia, livedo reticularis, pitting, raynaud, trombo, tromboflebite, trombosi, trombosi venosa profonda, tvp, ulcera acrale, ulcera agli arti, vascolare, vasculite.

Table 2: Dictionary of terms used for in-context learning experiments for each category of diagnosis.