Semantic Prompting for Large Language Models in Biomedical Named Entity Recognition

Erik Calcina
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Erik Novak
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Dunja Mladenić
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Abstract

Extracting structured medical information from unstructured clinical text remains a challenge for biomedical research and decision support. Recent advances in large language models (LLMs) suggest that prompt-based methods could provide a promising alternative to traditional supervised approaches for Named Entity Recognition (NER) in the biomedical domain. This study investigates whether adding semantic descriptions of entity labels can improve NER performance on clinical texts. Using a dataset of annotated case reports, we evaluate model performance in zero-shot, few-shot, and fine-tuned settings. Results show that semantic prompts enhance accuracy in low-supervision scenarios, while offering limited benefit once models are fine-tuned.

Keywords

Named entity recognition, large language models, semantic prompting, prompt engineering, medical domain, biomedicine

1 Introduction

Biomedical texts present a critical challenge for automated analysis. Clinical case reports, patient records, and related narratives are written in free text rather than in structured formats. While they contain essential medical knowledge, their unstructured nature makes it necessary to extract and organize information for systematic use in research and clinical decision support. Doing this manually is costly, time-consuming, and challenging to scale. Therefore, an automated approach to extract relevant information is required.

Named entity recognition (NER) models enable the identification and classification of clinically relevant entities, such as biological structures, diagnostic procedures, or symptoms. Recent advances in large language models (LLMs) show strong generalizing abilities, identifying relevant entities in both zero-shot and few-shot settings. However, in the biomedical domain, performance can be hindered by specialized terminology and subtle entity distinctions. To address this, we propose enriching prompts with semantic descriptions of entity labels, providing models with explicit context to improve their understanding of the task.

This study investigates the impact of semantically enhanced prompting in biomedical named entity recognition using large language models. We evaluate the effect of enriching entity labels with semantic descriptions on model performance across zero-shot, few-shot, and fine-tuned scenarios, using the MACCROBAT2020 dataset [3]. The contributions of this paper are threefold. First, we introduce the use of semantically enhanced prompts for biomedical NER by enriching entity labels with descriptions. Second, we provide a systematic evaluation of semantic prompting across zero-shot, few-shot, and fine-tuned scenarios, assessing its effectiveness under different levels of supervision. Third, we apply a statistical validation method, McNemar's test, to rigorously assess the reliability of observed performance differences between baseline and semantically enhanced prompts.

The remainder of the paper is structured as follows: Section 2 contains the overview of the related work. Next, we present the methodology in Section 3, and describe the experiment setting in Section 4. The experiment results are found in Section 5, followed by a discussion in Section 6. Finally, we conclude the paper and provide ideas for future work in Section 7.

2 Related Work

This section focuses on the related work on named entity recognition in biomedicine, as well as the use of semantic descriptions in prompting.

2.1 Prompting with semantic context

PromptNER introduced the idea of augmenting few-shot prompts with entity definitions, leading to substantial gains in F1 score on benchmarks like CoNLL, GENIA, and FewNERD, improving performance by 4–9 points compared to standard prompting [2]. Extending this idea, PromptNER unifies locating and typing into a single enriched prompt, enabling phrase extraction and entity classification simultaneously [7]. Similarly, the biomedical NER study demonstrated that "on-the-fly" inclusion of concept definitions enhances performance (+15% F1) in low-data settings [5].

2.2 Iterative and zero-shot semantic prompting

Recent work in zero-shot NER explores iterative prompt refinement to align model outputs with precise entity definitions. EvoPrompt uses an evolving definition-based framework to better distinguish between similar entity types, yielding improvements across benchmarks [9]. In a broader context, some studies found that while directly injecting semantic parses into LLM inputs can degrade performance, carefully designed semantic "hints" embedded in prompts can reliably boost outcomes [1].

2.3 Domain-specific prompt optimization

FsPONER optimizes few-shot prompts for industrial NER tasks by using semantic entity–enhanced meta prompts and task-specific exemplar selection, yielding F1 improvements of 5 to 13 points

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

in domain benchmarks [8]. In the biomedical domain, MPE³ integrates ontology-derived label semantics into prompts, improving performance in few-shot NER scenarios [10].

Prior research has shown that enriching prompts with semantic context and label definitions can significantly boost LLM performance in both few-shot and zero-shot NER. Our work provides a systematic evaluation in the biomedical domain. By examining multiple supervision settings, benchmarking several model families, and validating differences through McNemar's test, we offer a comprehensive assessment of when semantically enriched prompts provide benefits.

3 Methodology

This study evaluates the impact of incorporating semantic information into prompts on the performance of LLMs in biomedical NER tasks. Three distinct approaches were employed: zero-shot prompting, few-shot prompting, and fine-tuning.

Zero-shot prompting. In the zero-shot setting, models were prompted to perform NER without any prior exposure to labeled examples. Two types of prompts were utilized: *baseline prompt*, a standard instruction to identify and classify entities without additional context, and *semantically enhanced prompt*, which includes detailed descriptions for each entity label, offering explicit semantic context to guide the model's understanding and classification.

Few-shot prompting. The few-shot approach involved providing the models with a limited number of annotated examples (k-shots) before performing NER on new texts. Similar to the zero-shot setting, both baseline and semantically enhanced prompts were employed to assess the influence of semantic information.

Fine-tuning. Fine-tuning was conducted to adapt the pre-trained LLMs to the specific biomedical NER task. Two fine-tuning strategies were explored: *standard fine-tuning*, where models are fine-tuned using the original dataset annotations without additional semantic information, and *semantically enhanced fine-tuning*, which fine-tunes models on data where annotations were supplemented with semantic descriptions of each entity label.

4 Experiment Setting

This section describes the experiment setting, which includes the dataset and prompt preparation, the fine-tuning procedure used, the evaluation metrics, and the statistical significance test description.

4.1 Dataset

The experiments were conducted using the MACCROBAT2020 [3] dataset, which comprises 200 clinical case reports sourced from PubMed Central. In total, it contains 4,542 sentences with an average of 22.7 sentences per document, which includes manual annotations of biomedical entities, events, and relations, provided in brat standoff format[1]. For this study, we focused on the five most frequent entity labels within the dataset. These are BIOLOGICAL STRUCTURE, DIAGNOSTIC PROCEDURE, LAB VALUE, SIGN SYMPTOM, and DETAILED DESCRIPTION supplemented by the AGE and SEX labels. The inclusion of AGE and SEX was motivated by their prevalence and clarity within clinical narratives, providing a basis for evaluating model performance on both complex and straightforward entity types.

Each document was segmented into individual sentences by splitting on full stops. Subsequently, each sentence, along with its associated entity annotations, was transformed into a JSON format to facilitate processing by the language models.

[1] https://brat.nlplab.org/standoff.html

4.2 Semantically enhanced prompts

To enhance the semantic understanding of entity labels, detailed descriptions were crafted for each. These descriptions were derived by combining information from the MACCROBAT2020 dataset documentation and definitions from the Oxford English Dictionary [6]. The integration of these sources was performed manually, ensuring that the descriptions were both accurate and contextually relevant.

Prompts were structured as plain text instructions, guiding the model to identify and classify entities within the provided sentences. For the semantically enhanced prompts, the detailed entity descriptions were included to provide additional context. Models were instructed to output their responses in a JSON format, explicitly focusing on the labels component. Below we present an example of the entity description, specifically for the label AGE.

Baseline prompt: `The age of the patient.`
Semantic enhanced prompt: `The duration of time a patient has lived, expressed numerically (e.g., '65-year-old', '20 years old') or categorically (e.g., 'newborn', 'teenage'), representing their age at the time of presentation.`

This added context is intended to improve the model's ability to distinguish and extract nuanced biomedical entities more accurately.

4.3 Fine-tuning procedure

Fine-tuning is carried out using parameter-efficient techniques, where only lightweight adapter modules are trained instead of modifying the full model. This strategy reduces memory usage, mitigates catastrophic forgetting, and accelerates training.

To further improve efficiency, models are quantized to 4-bit precision. Fine-tuning is supervised and focuses on the generated outputs; all non-target tokens (e.g., system prompts, input context) are masked during loss computation. This ensures that training adapts the model to the expected JSON label output format rather than to the input content or prompt structure.

4.4 Evaluation metrics

To evaluate entity recognition performance, we use two F1-based metrics. The Exact F1 score measures strict matches, requiring predicted entities to align perfectly with the reference text and label. The Relaxed F1 score allows partial matches, counting predictions as correct if they include the true entity as a substring with the correct label.

4.5 McNemar statistical significance test

While Exact and Relaxed F1 scores quantify the magnitude of performance differences, they do not establish whether these differences are statistically reliable. The McNemar test [4] complements the Exact F1 metric by verifying whether observed improvements can be attributed to the semantically enhanced

prompts rather than random variation. Following standard NER practice, we treat Exact F1 as the primary endpoint and therefore apply McNemar's test only to exact match predictions.

Let b denote the number of cases correctly predicted by the semantically enhanced model but missed by the baseline, and c the number of cases correctly predicted by the baseline but missed by the semantically enhanced model. Only discordant pairs (b, c) contribute to the test; agreements do not affect the statistic. Using the continuity-corrected version of the test, the statistic is computed as

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c},$$

which follows a chi-squared distribution with one degree of freedom. The corresponding p-value allows us to test the null hypothesis H_0: the two models have equal marginal probabilities (i.e., performance differences are due to chance). Conventionally, $p < 0.01$ is considered statistically significant.

5 Results

This section presents model performance under three experimental conditions: zero-shot, few-shot, and fine-tuned prompting. For each condition, we compare the impact of semantically enhanced prompts against standard prompts using Exact and Relaxed F1 scores on a subset of clinically relevant entity types.

5.1 Zero-shot prompting

Table 1 reports the Exact and Relaxed F1 scores for models evaluated in the zero-shot setting using semantically enhanced prompts. Without semantic descriptions, most models struggled to generate outputs in the required JSON format, and valid scores could not be computed. Even with semantically enhanced prompts, Meta-Llama-3.1-8B consistently failed to produce structured responses.

Among the evaluated models, Llama-3.1-8B-Instruct achieved the highest Exact F1 score, while txgemma-9b-chat attained the best Relaxed F1 score. Llama-3.2-3B-Instruct and DeepSeek-Qwen-7B also demonstrated non-trivial performance in both metrics. These results suggest that semantically enhanced prompts can effectively compensate for the absence of training examples in zero-shot scenarios by providing clearer task guidance and improving structured prediction output.

Table 1: Exact and Relaxed F1 scores in the zero-shot setting with semantically enhanced prompts. Bolded values indicate the highest score in each column. Results without valid JSON output are marked with /.

Model	Exact F1 Semantics	Relaxed F1 Semantics
Llama-3.1-8B-Instruct[2]	**0.2310**	0.3708
Meta-Llama-3.1-8B[3]	/	/
Llama-3.2-3B-Instruct[4]	0.1620	0.3254
DeepSeek-Qwen-7B[5]	0.1592	0.3217
txgemma-9b-chat[6]	0.2181	**0.4245**

[2]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[3]https://huggingface.co/meta-llama/Llama-3.1-8B
[4]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[5]https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
[6]https://huggingface.co/google/txgemma-9b-chat

5.2 Few-shot prompting

Table 2 summarizes the Exact and Relaxed F1 scores for few-shot prompting. The addition of semantic information consistently improved model performance across most models. Notably, txgemma-9b-chat achieved the highest Exact F1 score 0.3288 and Relaxed F1 score 0.4998 with semantic prompting, compared to 0.2732 and 0.4469 without.

Both Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct showed improvements in both Exact and Relaxed F1 scores when provided with semantically enhanced prompts. For instance, Llama-3.1-8B-Instruct improved from 0.2509 to 0.3005 (Exact) and from 0.3526 to 0.3948 (Relaxed), while Llama-3.2-3B-Instruct increased from 0.2300 to 0.2439 (Exact) and from 0.3769 to 0.3948 (Relaxed). These gains highlight the benefit of enriching prompt instructions when training data is limited. However, not all models responded positively. For example, Meta-Llama-3.1-8B experienced a drop in Exact F1 from 0.2698 to 0.2210 and in Relaxed F1 from 0.3537 to 0.2799, indicating that semantically enhanced prompts do not universally improve performance and may be less effective for some models.

To assess the reliability of these differences, we conducted McNemar tests on Exact paired predictions. The tests revealed that performance differences between baseline and semantically enhanced prompts were statistically significant for all models except Llama-3.2-3B-Instruct. It is important to note, however, that significance here indicates that the two variants produce systematically different predictions, but does not itself imply improvement. For instance, while the difference for Meta-Llama-3.1-8B was highly significant, the semantically enhanced model in fact performed worse in terms of F1 scores.

5.3 Fine-tuned performance

In the fine-tuning scenario, results were more nuanced. As shown in Tables 2, most models performed strongly even without semantic enhancements. For instance, Meta-Llama-3.1-8B attained the highest Exact F1 score (0.7099) with semantic input, only slightly outperforming its baseline (0.7076), and this difference was not statistically significant ($p \approx 0.64$).

Some models, such as Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct, even showed small performance drops when semantic descriptions were included, with McNemar tests confirming that these differences were not significant ($p \approx 0.75$ and $p \approx 0.88$). This suggests that in settings where the model is already exposed to sufficient task specific supervision, additional prompt-level context may offer limited benefit or even introduce redundancy.

In contrast, TxGemma-9B-Chat exhibited the most notable improvement, with Exact and Relaxed F1 scores increasing from 0.6837 to 0.7092 and from 0.7483 to 0.7686, respectively; the McNemar test confirmed this difference as statistically significant ($p \approx 9.7 \times 10^{-5}$). By comparison, DeepSeek-Qwen-7B also showed a significant difference ($p \approx 6 \times 10^{-3}$), but in this case the semantically enhanced model performed worse (Exact F1: $0.7013 \rightarrow 0.6879$).

5.4 Overall observations

The largest performance improvements from semantically enhanced prompts appeared in zero-shot and few-shot settings, where gains in F1 scores were often statistically significant. In contrast, fine-tuned models showed smaller and mixed effects:

Table 2: Exact (left) and Relaxed (right) F1 scores for selected labels in few-shot and fine-tuned settings, with and without semantically enhanced prompts. Bolded values indicate the highest score in each column. We use symbols ○ and ● to denote whether the differences between using the baseline or semantically enhanced prompts are statistically significant (●) or not (○) according to the McNemar test at a significance level of $p = 0.01$.

Model	Exact F1				Relaxed F1			
	Few-Shot		Fine-Tuned		Few-Shot		Fine-Tuned	
	/	Semantic	/	Semantic	/	Semantic	/	Semantic
Llama-3.1-8B-Instruct	0.2509	0.3005 ●	0.7053	0.7004 ○	0.3526	0.3948	0.7660	0.7645
Meta-Llama-3.1-8B	0.2698	0.2210 ●	**0.7076**	**0.7099** ○	0.3537	0.2799	**0.7670**	**0.7765**
Llama-3.2-3B-Instruct	0.2300	0.2439 ○	0.6881	0.6867 ○	0.3769	0.3948	0.7629	0.7622
DeepSeek-Qwen-7B	0.1423	0.2270 ●	0.7013	0.6879 ●	0.2465	0.3891	0.7584	0.7521
txgemma-9b-chat	**0.2732**	**0.3288** ●	0.6837	0.7092 ●	**0.4469**	**0.4998**	0.7483	0.7686

for most, differences were not significant, though TxGemma-9B-Chat benefited reliably while DeepSeek-Qwen-7B showed a significant decrease. These results indicate that semantic prompting is most effective in low-resource conditions, while its impact under full supervision is limited and model-dependent.

6 Discussion

This section discusses the experiment findings and highlights the advantages and disadvantages of the different approaches.

6.1 Model pretraining and domain adaptation

TxGemma-9B-Chat, based on the Gemma 2 architecture and further fine-tuned on therapeutic development data, outperformed general-purpose models in a few-shot scenario. This suggests that domain-specific pretraining can significantly improve performance when supervision is limited. However, in the full fine-tuning setting, its advantage diminished. In fact, general models like Meta-Llama-3.1-8B achieved comparable but slightly better results, indicating that once sufficient task-specific supervision is provided, prior domain specialization offers limited additional benefit.

6.2 Prompt quality matters

The structure and clarity of prompts are critical to model performance. Poorly designed prompts often resulted in JSON formatting errors or reduced accuracy, particularly in zero-shot and few-shot settings. While adding semantic context improves task understanding by making objectives and entity definitions more explicit, excessive length or ambiguity can offset these gains.

6.3 Prompt length vs. model response

Semantic enrichment inevitably increases prompt length, which can slow response time and raise computational overhead. It may also overwhelm smaller models when excessive detail is included. In practical applications, this must be weighed against the potential gains in entity extraction accuracy.

7 Conclusion

This study investigated the impact of a semantically enhanced prompt design on LLM-based NER in the clinical domain. Our experiments on the MACCROBAT2020 dataset demonstrated that adding semantic label descriptions significantly improves model performance in zero-shot and few-shot scenarios, with notable gains in both Exact and Relaxed F1 scores. In contrast, fine-tuned models already exposed to task-specific data showed only marginal improvement.

Future work could explore adaptive semantic prompting strategies, such as ontology-driven label enrichment, and further investigate the trade-offs between prompt length and inference efficiency. Additionally, this method could be tested on larger datasets and across different models.

In summary, semantically enhanced prompts offer a straightforward yet effective way to boost clinical NER performance in low-data regimes, but their impact diminishes as models are exposed to more supervised training.

Acknowledgements

This work was supported by the Slovenian Research Agency. Funded by the European Union. UK participants in Horizon Europe Project PREPARE are supported by UKRI grant number 10086219 (Trilateral Research). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA) or UKRI. Neither the European Union nor the granting authority nor UKRI can be held responsible for them. Grant Agreement 101080288 PREPARE HORIZON-HLTH-2022-TOOL-12-01.

References

[1] Kaikai An, Shuzheng Si, Yuchi Wang, et al. 2024. Rethinking semantic parsing for large language models. *arXiv preprint arXiv:2409.14469*.
[2] Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.
[3] J. Harry Caufield, Yichao Zhou, Yunsheng Bai, David A. Liem, Anders O. Garlid, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2019. A comprehensive typing system for information extraction from clinical narratives. *medRxiv*. Preprint. DOI: 10.1101/19009118.
[4] Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 2, (June 1947), 153–157. DOI: 10.1007/bf02295996.
[5] Monica Munnangi, Sergey Feldman, Byron C. Wallace, et al. 2024. On-the-fly definition augmentation of llms for biomedical ner. *arXiv preprint arXiv:2404.00152*.
[6] 2025. Oxford english dictionary. https://www.oed.com/. Accessed: 2025-06-17. (2025).
[7] Yongliang Shen, Zeqi Tan, Shuhui Wu, et al. 2023. Promptner: prompt locating and typing for named entity recognition. In *ACL (Long Papers)*.
[8] Yongjian Tang, Rakebul Hasan, and Thomas Runkler. 2024. Fsponer: few-shot prompt optimization for named entity recognition. *arXiv preprint arXiv:2407.08035*.
[9] Zeliang Tong, Zhuojun Ding, and Wei Wei. 2025. Evoprompt: evolving prompts for enhanced zero-shot named entity recognition. In *COLING*.
[10] Yuwei Xia, Zhao Tong, Liang Wang, et al. 2023. Learning meta-prompt with entity-enhanced semantics for few-shot ner. *SSRN*.