Leveraging External Knowledge Bases: Analyzing Presentation Methods and Their Impact on Model Performance

Hui-Syuan Yeh[1], Thomas Lavergne[1], Pierre Zweigenbaum[1],
[1]Université Paris-Saclay, CNRS, LISN, Rue du Belvédère, 91405 - Orsay, France,
{yeh,lavergne,pz}@lisn.fr

Abstract

Integrating external knowledge into large language models has demonstrated potential for performance improvement across a wide range of tasks. This approach is particularly appealing in domain-specific applications, such as in the biomedical field. However, the strategies for effectively presenting external knowledge to these models remain underexplored. This study investigates the impact of different knowledge presentation methods and their influence on model performance. Our results show that inserting knowledge between demonstrations helps the models perform better, and improve smaller LLMs (7B) to perform on par with larger LLMs (175B). Our further investigation indicates that the performance improvement, however, comes more from the effect of additional tokens and positioning than from the relevance of the knowledge [1].

1 Introduction

While Large Language Models (LLMs) can potentially achieve strong performance in the medical domain, they are often difficult to run locally and hence raise significant data privacy concerns. Additionally, retraining and updating LLMs on biomedical corpora is a costly and resource-intensive process. Fortunately, the biomedical domain has established knowledge bases that can be leveraged to enhance LLMs without extensive retraining or exposing sensitive data.

Our approach is to integrate external knowledge from these knowledge bases into LLMs through natural language prompts. We use smaller LLMs which can be efficiently run locally. By incorporating additional knowledge as natural text, this method can be more effective than alternatives such as embedding-space integration or training models from scratch. As demonstrated, this approach outperforms graph-based models and knowledge embeddings for drug-drug interaction prediction (Zhu et al., 2023).

While the integration of external knowledge into LLMs in the prompt has been widely explored in general domains, its application in domain-specific settings, such as biomedical, remains understudied. Besides, existing guidelines for incorporating external knowledge are often intuitive rather than grounded in systematic experimentation.

In this work, we explore the use of the Unified Medical Language System (UMLS) Metathesaurus as a source of external knowledge to enhance prompts for biomedical relation extraction. We propose leveraging UMLS for two key reasons: (1) its background information can help highlight critical contextual details, and (2) it can potentially guide models toward specific relations with similar relationships.

This study aims to address the following research questions:

RQ 1. Which method is more effective to present the additional knowledge to the models, and how sensitive is model performance to the quality of the external knowledge provided?

RQ 2. If performance improves with presented external knowledge, is it truly due to the extra information, or could it result from other interconnected factors?

2 Experimental Design

To explore the integration of the Knowledge Base (KB), we modularize our experiments into three parts (see Figure 1). We first configure a good basic prompt with development sets (Section 3). On top of this foundation, we explore the use of external knowledge (Section 4). Finally, we integrate text from irrelevant knowledge sources to examine if

[1]Our code is available at: https://github.com/Dotkat-dotcome/umls-prompts

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 193–204

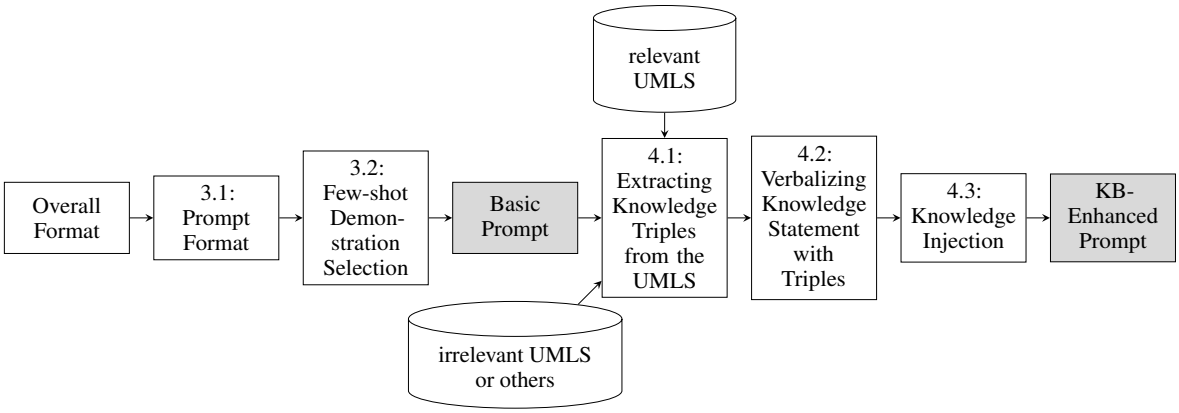


Figure 1: The flowchart of the experiment design.

the performance is influenced by the relevance of the knowledge (Section 5.3).

2.1 Datasets

We use four biomedical relation extraction datasets — two in English: ChemProt (Kringelum et al., 2016), DDI (Segura-Bedmar et al., 2013), and two in non-English: a subset of ADE in German (ADE-de) and French (ADE-fr) (Raithel et al., 2024) [2]. Details are described in Appendix A.

2.2 Models

Definition 2.1 (Demonstration) *A* demonstration *is a task sample provided to models during inference, included in the prompt to illustrate how the task should be performed.*

Definition 2.2 (In-Context Learning (ICL)) *The model is conditioned on a natural language instruction and/or a few* demonstrations *of the task and is then expected to complete further instances of the task simply by predicting what comes next. – (Brown et al., 2020)*

We use open source MISTRAL[3] within the In-Context Learning (ICL) framework for our experiments. The models can handle sequences of arbitrary length due to the use of sliding window attention. MISTRAL is an English model, but works well even on our non-English datasets. We also use BIOMISTRAL[4] for some experiments. BIOMISTRAL is a model based on MISTRAL and was further pre-trained on the PubMed Central corpus, primarily composed of English documents. It is shown that BIOMISTRAL underperforms its base model MISTRAL (Dorfner et al., 2024). We focus our experiments on MISTRAL and include BIOMISTRAL for further analysis.

As Causal Language Models (CLMs) do not always produce clean outputs for evaluation, we use simple pattern matching to extract answers from the models, discarding any responses that are out-of-label.

3 Basic Prompt Setup

3.1 Prompt Format

To ensure a good-quality prompt, we reference the prompt curated for the relation extraction task from the prompt source framework[5]. We pick the best one from a few trial runs on the development set. We then run variants presenting the entities of interest with different markers: *ordered markers*—Entities are masked in their appearing order with E1 and E2; *entity markers*—Entities are masked with their entity type; *decorated markers*—Entities are unmasked and enclosed in markers like [E1], [/E1] or [E2], [/E2]. (see Appendix Figure 7, Figure 5, and Figure 6 for full examples.).

Our results in Table 1 echoed the findings in (Zhang et al., 2024), that revealing the mention of interest (*decorated markers*) does not always perform better than masking out the mentions. Surprisingly, for DDI, *entity markers* perform best despite arbitrary entity order; while *ordered markers* works the best for ADE-de and ADE-fr, even with diverse entity types.

For the following experiments, we use *ordered markers* for ADE-de and ADE-fr, and *decorated markers* for ChemProt and DDI, based on our results. The latter choice ensures comparability with

[2]To ensure meaningful annotations, we take subsets that filter out relations with low inter-annotator agreement.".

[3]https://huggingface.co/mistralai/Mistral-7B-v0.1

[4]https://huggingface.co/BioMistral/BioMistral-7B

[5]https://github.com/bigscience-workshop/promptsource

Marker	Dataset			
	ADE-de	ADE-fr	DDI	ChemProt
decorated (~~~~ [E1]paracetamol[/E1] ~~ [E2]headache[/E2] ~~~~)	73.5	82.8	34.8	**59.0**
entity (~~~~ @DRUG$ ~~ @DISORDER$ ~~~~)	70.2	77.5	**40.4**	50.5
ordered (~~~~ E1 ~~ E2 ~~~~)	**74.5**	**85.7**	35.6	47.8

Table 1: A comparison of different prompt formats over the development set with MISTRAL on 1-shot (per relation) relation extraction.

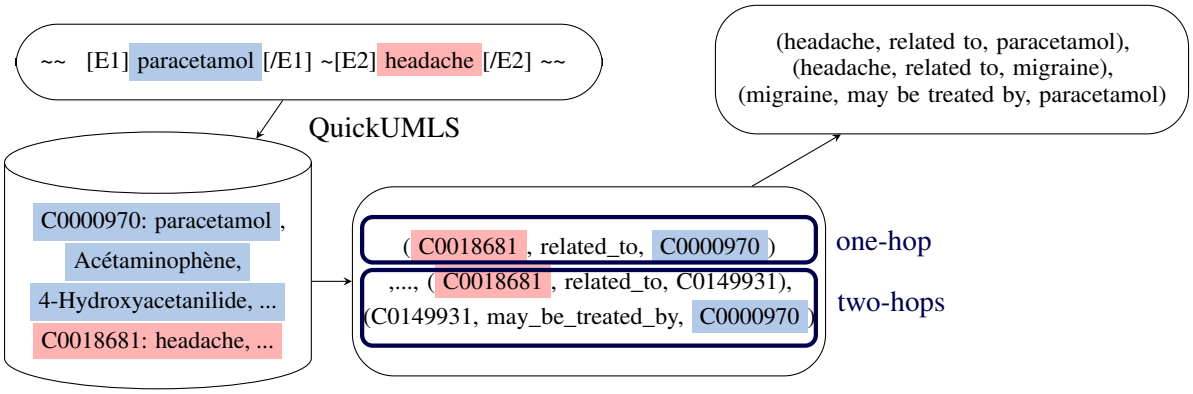


Figure 2: An illustration of extracting and verbalizing information from the UMLS.

prior work and adds task-relevant information.

3.2 Few-shot Demonstrations Selection

To improve performance over random demonstrations, we implement a retrieval module using similarity based on *bag of n-gram token*. The rationale is that selecting samples with similar relations to the inference sample increases the likelihood of correct predictions. In order to ensure a low-resource setting, for each dataset, we randomly select 10% of the training set to create a pool for drawing demonstrations. We map samples to bi-grams and tri-grams using NLTK toolkit[6], compute Jaccard similarity, and select the top-k most similar examples from all relations. Demonstrations are ordered inversely by similarity, placing the most similar samples near the model's output.

4 KB-Enhanced Prompt Setup

The external knowledge source we use is the Unified Medical Language System (UMLS)[7], a rich biomedical resource. In this section, we introduce the setup for applying the UMLS knowledge to enhance the prompts for the biomedical relation extraction task, as illustrated in Figure 2.

[6]https://github.com/nltk/nltk

[7]https://www.nlm.nih.gov/research/umls/index.html

4.1 Extracting Knowledge Triples from the UMLS

To access the relevant part of the ontologies recorded in the UMLS, we use the QuickUMLS [8] to map the two entities to be classified in a sample to their CUIs (Concept Unique Identifiers)[9]. In Figure 2, "paracetamol" and "headache" are mapped to "C0000970" and "C0018681" respectively for looking up the associated relationships. This mapping is an entity-linking process that uses an approximate dictionary-based approach to find the best match of concept identifiers in the UMLS for input strings.

From the two CUIs of the associated entities in one sample, we extract both direct and one-hop relationships between these CUIs from the UMLS table MRREL[10]. For instance, one of the two-hops relationships extracted between "C0000970" and "C0018681" is "(C00018681, related_to, C0149931), (C0149931, may_be_treated_by, C0000970)".

[8]https://github.com/Georgetown-IR-Lab/QuickUMLS

[9]CUIs are the key to obtaining information from the UMLS. In the UMLS, terminology is mapped to the CUIs for disambiguating the concepts, and for documenting relationships.

[10]https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T.related_concepts_file_mrrel_rrf/

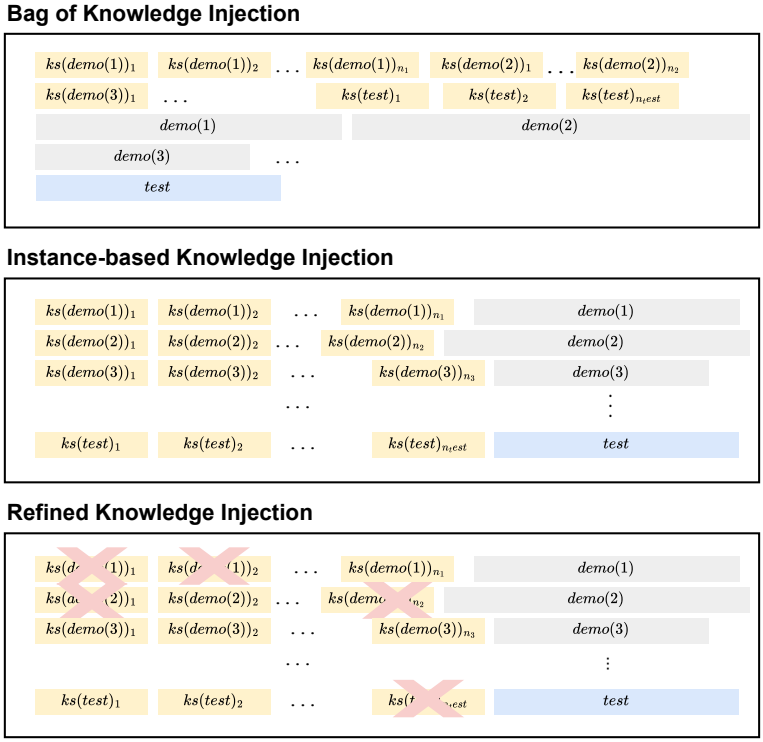


Figure 3: An illustration of the three knowledge injection methods, showcasing increasing levels of refinement from top to bottom. Knowledge statements are in yellow. Top: *Bag of Knowledge Injection* has all knowledge statements prepended altogether to the prompt. Middle: *Instance-based Knowledge Injection* has knowledge statements prepended to each instance. Bottom: *Refined Knowledge Injection* has low-quality knowledge statements removed from *Instance-based Knowledge Injection*.

4.2 Knowledge Statement: Verbalizing Knowledge Statement with Triples

After extracting the relevant triples from the UMLS, we process them to be more natural language-like as it was demonstrated to help the model perform tasks better (Gonen et al., 2023).

For instance, the extracted triples (C00018681, related_to, C0149931) and (C0149931, may_be_treated_by, C0000970) are processed to "(headache, related to, migraine), (migraine, may be treated by, paracetamol)". The CUIs of the intermediate concepts are mapped to their preferred terms[11] using UMLS table `MRCONSO`[12], C0000970 is mapped to "migraine". On the other hand, the CUIs of the entities are mapped to their original mentions from their corresponding samples. We select preferred terms in the same language as the dataset[13]. In this way, we allow the external knowledge to be possibly more integrated into the model's reasoning process as in the case where the preferred terms exist in the sample sentences.

We refer to the processed triples as *knowledge statements*— knowledge expressed as natural language statements. The knowledge statements are then injected into the prompt in different ways, which we will introduce in the next section.

4.3 Knowledge Injection (KI)

We present the extracted *knowledge statements* into the prompt at varying levels of quality, where quality is defined by *granularity*—the degree of association between the task sample and the knowledge statements. As illustrated in Figure 3, lower granularity corresponds to less refined knowledge, requiring minimal pre-processing but placing a greater reasoning burden on the model to achieve strong task performance. Conversely, higher granularity involves more carefully curated knowledge, reducing the model's reasoning load.

[11]The string preferred in a source or in the Metathesaurus as the name of a concept, lexical variant, or string.

[12]https://www.ncbi.nlm.nih.gov/books/NBK9685/table/ch03.T.concept_names_and_sources_file_mr/

[13]We use the column "TTY" in the `MRCONSO` table to select preferred terms.

#Train	Model	Method	ChemProt	DDI	ADE-de	ADE-fr
1-shot	MISTRAL	ICL w/o KI	42.2	39.6	78.5	73.9
		Bag of KI	53.9	40.8	74.7	71.6
		Instance-based KI	<u>60.1</u>	**44.9**	<u>79.9</u>	**77.3**
		Refined KI	**60.2**	<u>44.3</u>	**80.0**	**77.3**
1-shot [1]	GPT-3.5-TURBO (Zhang et al., 2024)	ICL w/o KI	68.5	-	-	-
	GPT-3.5 (Jahan et al., 2024)	ICL w/o KI	-	46.43	-	-
full-shot [2]	PUBMEDBERT	finetuned	73.2	75.9	-	-
	XLM-ROBERTA	finetuned	-	-	66.3	76.4

Table 2: Macro F_1 across different methods on datasets ChemProt, DDI, ADE-de, and ADE-fr, aggregating over five random seeds. Within the MISTRAL experiments, we highlight the **best** score with bold, and <u>second-best</u> score with underline.

[1] We collect the GPT-3.5-TURBO and GPT-3.5 results from the benchmarking papers. [2] We train a classifier using PUBMEDBERT for tasks in English, i.e., ChemProt and DDI; and XLM-ROBERTA for ADE-fr in French and ADE-de in German. To address the issue of imbalanced class distribution, we employ a resampling technique while training XLM-ROBERTA. [3] We set the similarity threshold for the Refined KI to 0.85 for ChemProt and DDI and 0.9 for ADE-de and ADE-fr.

- **Bag of KI** We prepended all extracted descriptions to the beginning of the prompt as a *bag*. This presentation requires the models to associate and reason with the evidence.
- **Instance-based KI** We prepended extracted descriptions to each associated instance. In this presentation, the relevant information is directly before the *instance*.
- **Refined KI** We prepended only the high-quality, semantically relevant triples to associated instances. We used PUBMEDBERT to encode samples and knowledge statements, pruning irrelevant knowledge statements based on cosine similarity of CLS embeddings and a similarity threshold.

5 Results

5.1 Knowledge Injection vs. Baselines

In this section, we discuss our experiment results, summarized in Table 2.

***ICL w/o KI* > full-shot finetuned BERT-based models on user-generated datasets** Our base setup (*ICL w/o KI*) performs better than the full-shot fine-tuned BERT-based models on the user-generated dataset, ADE-de (∼+5% F_1) and perform almost on-par on ADE-fr (∼+1% F_1); while performs worse than the full-shot fine-tuned PUBMEDBERT on the scientific dataset, ChemProt (∼-30% F_1) and DDI (∼-40% F_1).

While we argued previously that ADE-de and ADE-fr are more familiar to the models, it is still surprising that MISTRAL works well on them (even better than the fine-tuned XLM-ROBERTA) despite not having any external knowledge nor entity type information.

***ICL w/o KI* < full-shot finetuned BERT-based models in English scientific dataset** ChemProt and DDI, on the other hand, are more challenging for CLMs with ICL, including our base setup and the state-of-the-art. GPT-3.5-TURBO, a very strong baseline, yields lower performance than fine-tuned models on ChemProt (∼-10% F_1) and GPT-3.5 underperforms on DDI (∼-35% F_1).

Our *ICL w/o KI* with Mistral yields lower performance than GPT-3.5 models on ChemProt (-20% F_1) and DDI (- 5% F_1). GPT-3.5 is a larger CLM with a parameter size of 175B, while our model has 7B. To our knowledge, there is no study with 7B CLMs on ChemProt and DDI for reference here.

Bag of KI* < *ICL w/o KI Compared to our base setup without any external knowledge (*ICL w/o KI*), *Bag of KI* does not show consistent improvement across the datasets; while ChemProt (∼+10% F_1) and DDI (∼+1% F_1) show improvement, ADE-de (∼-3% F_1) and ADE-fr (∼-1% F_1) show a decrease in performance. In the cases where *Bag of KI* underperforms (ADE-de and ADe-fr), the performance is very high to begin with, and the additional knowledge might not be very helpful, since it is background information that still requires models to associate it to the respective instances.

Instance-based KI* ≈ *Refined KI* > *ICL w/o KI *Instance-based KI* and *Refined KI* show consis-

tent improvement compared to *Bag of KI* and *ICL w/o KI* on ChemProt (∼+20% F_1), DDI (∼+5% F_1), ADE-de (∼+2% F_1), and ADE-fr (∼+3% F_1). These results suggest that positioning the knowledge closer to the instances is more beneficial for the models to make the right prediction. Nevertheless, comparing *Refined KI* to *Instance-based KI*, we can see that the performance is barely increasing. We do not know if the insignificant improvement is due to the *Refined KI* sometimes removing the knowledge statements of good quality, or it is due to that the quality of the knowledge statements is not as important for the performance. Therefore, we further explore the effect of the quality of the knowledge statements in the next section, when we explore the similarity threshold for *Refined KI*.

***Instance-based KI* boosts the performance of smaller CLMs to be more on par with the big CLMs** *Instance-based KI* with MISTRAL obtain better results than, the large CLM GPT-3.5 on DDI (∼+1% F_1) and much closer to GPT-3.5-TURBO on ChemProt (∼-5% F_1) than the base setup. These results, as GPT-3.5s results, are still behind the full-shot fine-tuned BERT-based models on ChemProt and DDI by a noticeable margin, but the gap is much smaller than the base setup. Small BERT-based models are still highly effective for biomedical relation extraction tasks due to their ease of fine-tuning. Additionally, smaller CLMs with appropriate knowledge injections can also achieve competitive results and are significantly more efficient to run than the larger CLMs.

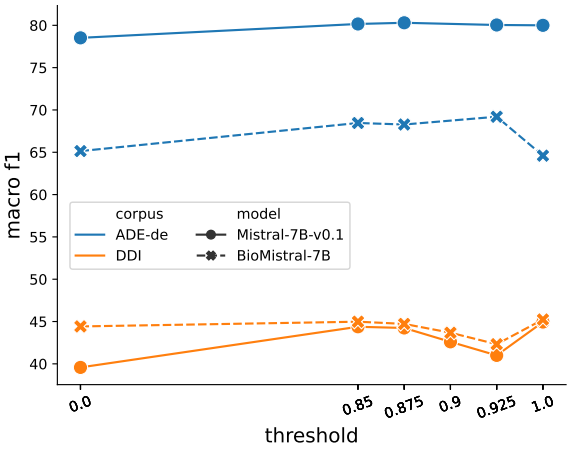


Figure 4: macro F_1(%) over similarity threshold with MISTRAL and BIOMISTRAL on ADE-de and DDI. The x-axis similarity threshold runs from 0, which corresponds to Instance-based KI, to 1, which corresponds to the w/o KI.

5.2 Effect of Similarity Threshold

The semantic similarity between the knowledge statements and the instances is high, with scores ranging from 0.80 to 0.95 (see Appendix Figure 8). We examine the effect with MISTRAL and BIOMISTRAL on ADE-de and DDI (see Figure 4). With all additional knowledge (threshold=0), BIOMISTRAL performs worse than MISTRAL on ADE-de (-15% F_1), but slightly better in DDI (∼+5% F_1). This discrepancy is likely due to BIOMISTRAL's medical training resources being predominantly in English, hence making it less effective on multilingual datasets. Although MISTRAL initially performs worse on DDI, enforcing a similarity threshold brings MISTRAL to perform on par with BIOMISTRAL. This result demonstrates that general models can be improved by high-quality knowledge statements to match the capacity of biomedical models trained with additional large corpora.

The results show that the performance improves with increasing similarity thresholds and that the performance is saturated around 0.85 for DDI, and 0.9 for ADE-de; followed by a decline. These results suggest that while higher-quality knowledge statements enhance performance, excessively high thresholds may reduce the number of usable knowledge statements, thereby hurting the overall performance.

5.3 Effect of Knowledge Source and Position

The additional knowledge statements help all datasets, yet they also change the prompt layout, which could affect the model performance. Our goal here is to investigate if the observed performance gains are contributed by external knowledge rather than extra tokens that changed the prompt format. We, therefore, swap the knowledge statements from the extracted UMLS triples.

- *UMLS instance-unrelated*: UMLS triples relevant to the corpus (extracted as described in Section 4) but irrelevant to the sample.
- *UMLS corpus-unrelated*: UMLS triples that are completely irrelevant to the corpus. We extract triples that do not involve any CUI from the entities in the corpus.
- *Bible*: We take text from the Wikipedia page of the Bible[14] as our generation pool. This ex-

[14]https://en.wikipedia.org/wiki/Bible

Position	Triple Source	*macro* F_1
-	ICL w/o KI	42.2
task background	Bag of KI	53.9
close-to instances	Instance-based KI (UMLS instance-related)	60.2
close-to instances	UMLS instance-unrelated	60.0
	UMLS corpus-unrelated	**61.4**
	Bible	60.1
	Empty	60.0

Table 3: macro F_1 (%) with different adversarial knowledge statements on ChemProt.

periment serves as a totally irrelevant knowledge source.

- *Empty*: We discard the content in the triplet template using just the placeholder, i.e., (, ,).

Position Compared to *Bag of KI* where all knowledge statements are collected as task background altogether at the beginning of all instances, all methods that place knowledge statements close to instances show better performance (see Table 3), regardless of whether the knowledge statements are relevant or irrelevant. These results suggest that the models can effectively benefit from relevant knowledge statements when they are closely positioned to the instances. However, when the knowledge statements are distanced from the instances, the models struggle to recognize and leverage the knowledge.

Knowledge Source All knowledge sources improve the base setup (∼+20% F_1), including *Empty* (see Table 3). The results suggest that these additional tokens in between the instances improve the performance.

6 Discussion and Conclusion

For our experiments on ADE-de and ADE-fr, the prompts contain two languages: the instructions, relations from the UMLS, and the ground truth label of the sample—known as the *verbalizer*—are in English, while the samples, entities, and entities linked to the UMLS are respectively in French and German. The mixing of languages in prompts was studied in multilingual relation classification tasks (Chen et al., 2022) and cross-lingual natural language inference (XNLI) (Zhou et al., 2023). These studies concluded that directly translating the verbalizers to the target language for inference is not helpful. However, the effect of other parts of the prompt is still to be understood. Our results show that the mixed-language prompts still achieve competitive results in our tasks.

In this work, we explored the integration of external knowledge for the extraction of biomedical relations within the context of in-context learning. We extracted triples from the UMLS based on the entities involved in the relations and injected them into the prompt with different *granularity*.

Our experiments for configuring a basic prompt revealed that different entity markers are effective across different datasets, showing that entity mentions are not always more beneficial for the models than marking with entity types or order. Our experiments showed that MISTRAL with ICL performs very well on the user-generated datasets in non-English; however, the model still performs poorly on more difficult tasks in English. With knowledge integration, the performance of ICL is boosted to be more on par with the larger autoregressive models.

The knowledge statements help the model perform better across all datasets. Additionally, applying a suitable similarity threshold for further refining the knowledge statements further helps the models, especially for models trained only on general corpora. We observed that the performance was even more affected by the positioning and the addition of tokens. When the additional knowledge is positioned close to the instances, the models can effectively identify relevant knowledge statements.

Limitations

There are limitations to be noted for this work. Firstly, in the experiment setup, the hyperparameters are tuned in a cascaded manner, which is less computationally expensive yet suboptimal. Secondly, entity linking can be a bottleneck for this method, especially considering the typos and informal language of user-generated datasets. Third, the effect of prompt length is still to be understood. We found that the additional tokens can possibly help, even if carrying irrelevant knowledge, however, the effect of inserting irrelevant tokens and how one places them in the prompt also require further investigation. While related work has studied this direction (Levy et al., 2024), domain-specific tasks remain understudied and require more research.

Acknowledgments

This work was financially supported by the trilateral project KEEPHA, Grant Number JST-JPMJCR20G9, DFG-442445488, and ANR-20-

IADJ-0005-01. The experiments of this work were performed on Jean-Zay's V100 and A100 partitions. We also want to thank the reviewers for their feedback on this paper.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yuxuan Chen, David Harbecke, and Leonhard Hennig. 2022. Multilingual relation classification via efficient and effective prompting. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1059–1075, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, et al. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, page 108189.

Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. 2016. ChemProt-3.0: a global chemical biology diseases mapping. *Database*, 2016.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Lisa Raithel, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Philippe Thomas, Tomohiro Nishiyama, Sebastian Möller, Eiji Aramaki, Yuji Matsumoto, Roland Roller, and Pierre Zweigenbaum. 2024. A dataset for pharmacovigilance in German, French, and Japanese: Annotating adverse drug reactions across languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 395–414, Torino, Italia. ELRA and ICCL.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jeffrey Zhang, Maxwell Wibert, Huixue Zhou, Xueqing Peng, Qingyu Chen, Vipina K Keloth, Yan Hu, Rui Zhang, Hua Xu, and Kalpana Raja. 2024. A study of biomedical relation extraction using gpt models. *AMIA Summits on Translational Science Proceedings*, 2024:391.

Meng Zhou, Xin Li, Yue Jiang, and Lidong Bing. 2023. Enhancing cross-lingual prompting with dual prompt augmentation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11008–11020, Toronto, Canada. Association for Computational Linguistics.

Fangqi Zhu, Yongqi Zhang, Lei Chen, Bing Qin, and Ruifeng Xu. 2023. Learning to describe for predicting zero-shot drug-drug interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14855–14870, Singapore. Association for Computational Linguistics.

A Dataset

Dataset	Source	#Relation	Relations	#Test
ChemProt	PubMed abstracts	6	activation (CPR:3) inhibition (CPR:4) agonist (CPR:5) antagonist (CPR:6) substrate (CPR:9) false (none of above)	16,943
DDI	MedLine abstracts	5	DDI-advise DDI-effect DDI-int DDI-mechanism DDI-false	5,761
ADE-de	Patient Forum	7	caused experienced_in has_dosage has_time signals_change_of treatment_for false	3,285
ADE-fr	Patient Forum	7	caused experienced_in has_dosage has_time signals_change_of treatment_for false	551

Table 4: Dataset Overview

B Prompt Examples

ordered markers E1-E2

Out of the possible relations: [CAUSED, EXPERIENCED_IN, HAS_DOSAGE, HAS_TIME, SIGNALS_CHANGE_OF, TREATMENT_FOR, NONE]
###

Given the sentence, J'ai aussi E1 pour la première fois de ma vie E2.
What is the semantic relation between the two nominals (nouns or noun phrases) DISORDER E1 and TIME E2 in the sentence: HAS_TIME

Given the sentence, et de la dominance en œstrogène ! Depuis six mois, je prends E2 de E1.
What is the semantic relation between the two nominals (nouns or noun phrases) DRUG E1 and MEASURE E2 in the sentence: HAS_DOSAGE

Given the sentence, J'ai aussi E1 pour la première fois de ma vie au cours des six derniers mois. Moi aussi, je suis désespérée par mon E2 et, enfin .
What is the semantic relation between the two nominals (nouns or noun phrases) DISORDER E1 and ANATOMY E2 in the sentence: EXPERIENCED_IN

Given the sentence, et de la dominance en œstrogène ! Depuis six mois, je prends 50 mg de E1....La fluoxétine est connue pour faire perdre du poids....J'ai E2 au début.
What is the semantic relation between the two nominals (nouns or noun phrases) DRUG E1 and DISORDER E2 in the sentence: CAUSED

Given the sentence, E2, je résistais à tout ! Mais quand rien n'allait plus, j'ai accepté d'en prendre. J'ai aussi E1 pour la première fois de ma vie au cours des six derniers mois.
What is the semantic relation between the two nominals (nouns or noun phrases) DISORDER E1 and TIME E2 in the sentence: NONE

Given the sentence, J'ai pris E2 après avoir pris 3 hormones différentes, ça a bien marché, mais j'ai dû E1 parce que j'avais des saignements abondants (janvier).
What is the semantic relation between the two nominals (nouns or noun phrases) CHANGE_TRIGGER E1 and DRUG E2 in the sentence: SIGNALS_CHANGE_OF

Given the sentence, Je prends maintenant Trisequens (depuis 2 mois) et E1 pour E2 et l'humeur.
What is the semantic relation between the two nominals (nouns or noun phrases) DRUG E1 and DISORDER E2 in the sentence: TREATMENT_FOR

###

Given the sentence,De plus, j'ai commencé à avoir des nausées, des E1 de E2, des muqueuses sèches, etc.
What is the semantic relation between the two nominals (nouns or noun phrases) DISORDER E1 and ANATOMY E2 in the sentence:

Figure 5: An example of the prompt with ordered markers.

entity-type markers @TYPE$

Out of the possible relations: [CAUSED, EXPERIENCED_IN, HAS_DOSAGE, HAS_TIME, SIGNALS_CHANGE_OF, TREATMENT_FOR, NONE]
###

Given the sentence, J'ai aussi @DISORDER$ pour la première fois de ma vie @TIME$.
What is the semantic relation between the two nominals (nouns or noun phrases) @DISORDER$ and @TIME$ in the sentence: HAS_TIME

Given the sentence, et de la dominance en œstrogène ! Depuis six mois, je prends @MEASURE$ de @DRUG$.
What is the semantic relation between the two nominals (nouns or noun phrases) @DRUG$ and @MEASURE$ in the sentence: HAS_DOSAGE

Given the sentence, J'ai aussi @DISORDER$ pour la première fois de ma vie au cours des six derniers mois. Moi aussi, je suis désespérée par mon @ANATOMY$ et, enfin .
What is the semantic relation between the two nominals (nouns or noun phrases) @DISORDER$ and @ANATOMY$ in the sentence: EXPERIENCED_IN

Given the sentence, et de la dominance en œstrogène ! Depuis six mois, je prends 50 mg de @DRUG$....La fluoxétine est connue pour faire perdre du poids....J'ai @DISORDER$ au début.
What is the semantic relation between the two nominals (nouns or noun phrases) @DRUG$ and @DISORDER$ in the sentence: CAUSED

Given the sentence, @TIME$, je résistais à tout ! Mais quand rien n'allait plus, j'ai accepté d'en prendre. J'ai aussi @DISORDER$ pour la première fois de ma vie au cours des six derniers mois.
What is the semantic relation between the two nominals (nouns or noun phrases) @DISORDER$ and @TIME$ in the sentence: NONE

Given the sentence, J'ai pris @DRUG$ après avoir pris 3 hormones différentes, ça a bien marché, mais j'ai dû @CHANGE_TRIGGER$ parce que j'avais des saignements abondants (janvier).
What is the semantic relation between the two nominals (nouns or noun phrases) @CHANGE_TRIGGER$ and @DRUG$ in the sentence: SIGNALS_CHANGE_OF

Given the sentence, Je prends maintenant Trisequens (depuis 2 mois) et @DRUG$ pour @DISORDER$ et l'humeur.
What is the semantic relation between the two nominals (nouns or noun phrases) @DRUG$ and @DISORDER$ in the sentence: TREATMENT_FOR

###

Given the sentence,De plus, j'ai commencé à avoir des nausées, des @DISORDER$ de @ANATOMY$, des muqueuses sèches, etc.What is the semantic relation between the two nominals (nouns or noun phrases) @DISORDER$ and @ANATOMY$ in the sentence:

Figure 6: An example of the prompt with entity-type markers.

decorated markers [E]ENTITY_T[/E]

Out of the possible relations: [CAUSED, EXPERIENCED_IN, HAS_DOSAGE, HAS_TIME, SIGNALS_CHANGE_OF, TREATMENT_FOR, NONE]
###

Given the sentence, J'ai aussi [E1]pris du poids[/E1] pour la première fois de ma vie [E2]au cours des six derniers mois[/E2].
What is the semantic relation between the two nominals (nouns or noun phrases) DISORDER pris du poids and TIME au cours des six derniers mois in the sentence: HAS_TIME

Given the sentence, et de la dominance en œstrogène ! Depuis six mois, je prends [E2]50 mg[/E2] de [E1]fluoxétine[/E1].
What is the semantic relation between the two nominals (nouns or noun phrases) DRUG fluoxétine and MEASURE 50 mg in the sentence: HAS_DOSAGE

Given the sentence, J'ai aussi [E1]pris du poids[/E1] pour la première fois de ma vie au cours des six derniers mois. Moi aussi, je suis désespérée par mon [E2]ventre[/E2] et, enfin .
What is the semantic relation between the two nominals (nouns or noun phrases) DISORDER pris du poids and ANATOMY ventre in the sentence: EXPERIENCED_IN

Given the sentence, et de la dominance en œstrogène ! Depuis six mois, je prends 50 mg de [E1]fluoxétine[/E1]....La fluoxétine est connue pour faire perdre du poids....J'ai [E2]perdu immédiatement 3 kg[/E2] au début.
What is the semantic relation between the two nominals (nouns or noun phrases) DRUG fluoxétine and DISORDER perdu immédiatement 3 kg in the sentence: CAUSED

Given the sentence, [E2]Jusqu'à l'année dernière[/E2], je résistais à tout ! Mais quand rien n'allait plus, j'ai accepté d'en prendre. J'ai aussi [E1]pris du poids[/E1] pour la première fois de ma vie au cours des six derniers mois. What is the semantic relation between the two nominals (nouns or noun phrases) DISORDER pris du poids and TIME Jusqu'à l'année dernière in the sentence: NONE

Given the sentence, J'ai pris [E2]Kliogest[/E2] après avoir pris 3 hormones différentes, ça a bien marché, mais j'ai dû [E1]arrêter[/E1] parce que j'avais des saignements abondants (janvier).
What is the semantic relation between the two nominals (nouns or noun phrases) CHANGE_TRIGGER arrêter and DRUG Kliogest in the sentence: SIGNALS_CHANGE_OF

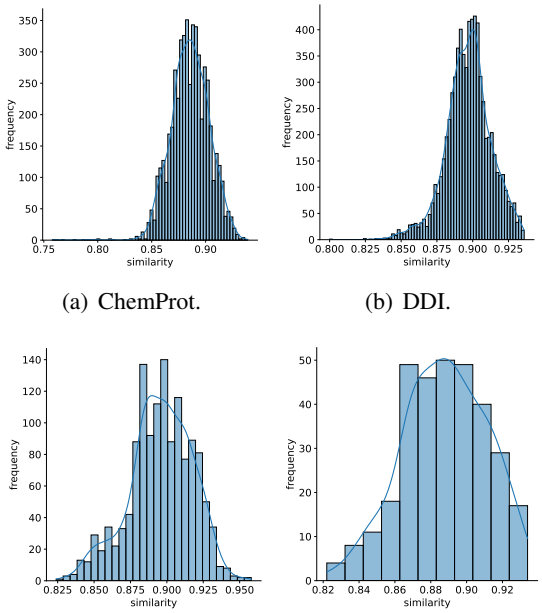
Given the sentence, Je prends maintenant Trisequens (depuis 2 mois) et [E1]Insidon[/E1] pour [E2]l'anxiété[/E2] et l'humeur.
What is the semantic relation between the two nominals (nouns or noun phrases) DRUG Insidon and DISORDER l'anxiété in the sentence: TREATMENT_FOR

###

Given the sentence,De plus, j'ai commencé à avoir des nausées, des [E1]inflammations[/E1] de [E2]l'estomac[/E2], des muqueuses sèches, etc.
What is the semantic relation between the two nominals (nouns or noun phrases) DISORDER inflammations and ANATOMY l'estomac in the sentence:

Figure 7: An example of the prompt with decorated markers.

C Similar Distribution of Knowledge Statements



(a) ChemProt. (b) DDI.

(c) ADE-de. (d) ADE-fr.

Figure 8: Similarity distribution of knowledge statements for different datasets. (a) ChemProt, (b) DDI, (c) ADE-de, and (d) ADE-fr.