LLMs in Biomedical: A Study on Named Entity Recognition

Anonymous ACL submission

Abstract

Large Language Models (LLMs) demonstrate remarkable versatility in various NLP tasks but encounter distinct challenges in biomedical due to the complexities of language and data scarcity. This paper investigates LLMs application in the biomedical domain by exploring strategies to enhance their performance for the NER task. Our study reveals the importance of meticulously designed prompts in the biomedical. Strategic selection of in-context examples yields a marked improvement, offering $\sim 15-20\%$ increase in F1 score across all benchmark datasets for biomedical few-shot NER. Additionally, our results indicate that integrating external biomedical knowledge via prompting strategies can enhance the proficiency of general-purpose LLMs to meet the specialized needs of biomedical NER. Leveraging a medical knowledge base, our proposed method, DiRAG, inspired by Retrieval-Augmented Generation (RAG), can boost the zero-shot F1 score of LLMs for biomedical NER. Code will be released upon acceptance.

1 Introduction

LLMs such as GPT4 have demonstrated exceptional capabilities across diverse tasks and domains (Espejel et al., 2023; Dai et al., 2023; Dong et al., 2022). These models could have a revolutionary impact on healthcare; however, their integration into medical research and practice has been slow (Zhou et al., 2023; Vaishya et al., 2023; Nori et al., 2023a) and it is crucial to examine the unique challenges presented by the biomedical field that contribute to this discrepancy. Specifically, LLMs encounter challenges in medical Information Extraction (Gutierrez et al., 2022; Moradi et al., 2021) due to the scarcity of high-quality biomedical data in their pretraining, and the need for a nuanced comprehension of the text for this task (Gu et al., 2023). Medical entities can have multiple synonyms and abbreviations, complicating their recognition by models (Grossman Liu et al., 2021). Furthermore, context sensitivity is even more critical in the biomedical compared to the general domain. The specificity of entity types and the complexity of their interrelations necessitate a level of background knowledge that standard prompts may fail to provide. LLMs are primarily exposed to vast amounts of generic text data limiting their effectiveness in managing the intricate nuances of medical language (Kumari et al., 2023; Karabacak and Margetis, 2023).

In this paper, we concentrate on NER, a foundational task for various applications such as recruiting patients for clinical trials, searching biomedical literature, or building models that predict the progression of disease based on free-text notes.

In our initial analysis, we broaden the scope of TANL (Paolini et al., 2021) and DICE (Ma et al., 2022), two text-to-text formats initially proposed for model training, adapting their use to prompt design specifically for biomedical NER. Our findings reveal that the relative effectiveness of the resulting prompt pattern varies based on specific dataset characteristics. Subsequently, we investigate the importance of example selection via In-Context Learning (ICL) and demonstrate the value of nearest neighbor example selection using pre-trained biomedical text encoders when performing biomedical NER. A key question that arises in the deployment of LLMs concerns the comparative advantage of closed-source LLMs versus open-source ones. In our third study, we shed light on this question by presenting an assessment of performance and cost across various experiments. Furthermore, we explore the integration of external medical knowledge to refine LLM capabilities (Gao et al., 2023c; Zakka et al., 2024). Leveraging the insights gained from these techniques, we present a novel data augmentation method incorporating a medical knowledge base, e.g., UMLS (Bodenreider, 2004), which substantially improves zero-shot biomedical NER.

2 Background and Preliminaries

Prompt engineering Prompt tuning (White et al., 2023; Lester et al., 2021; Ding et al., 2021) as its own research field shows that skillfully crafted prompts can significantly enhance LLM understanding for complex tasks (Lu et al., 2021; Kaddour et al., 2023; Webson and Pavlick, 2021). Researchers have explored different prompt formats for IE tasks with LLMs (Wang et al., 2023c; Gutierrez et al., 2022; Wang et al., 2023b) including more work around knowledge insertion for prompt augmentation (Seo et al., 2024; Chen et al., 2023) Another type of prompting is ICL (Brown et al., 2020), where LLMs use a limited set of "input-output" pairs within the prompt along with a query input as demonstrations of what the task output should be. In this realm, Liu et al. (2021); Min et al. (2022); Gao et al. (2023a) demonstrated that choosing targeted in-context examples over random sampling leads to more accurate model responses.

Named Entity Recognition GPT-NER (Wang et al., 2023b) was one of the first methods to incorporate a unique symbol to transform the sequence tagging task into text generation via ICL with GPT-3 (Brown et al., 2020), achieving performance on par with fully supervised baselines. Following this work, Gutierrez et al. (2022); Moradi et al. (2021) showed that LLMs are not skilled few-shot learners in the biomedical domain. However, recent advancements, such as GPT-4, have increased LLM performance on many tasks (Tian et al., 2024; Hu et al., 2024a; Nori et al., 2023a) including in the biomedical domain (Hu et al., 2024b). In the direction of knowledge distillation from LLMs (Wang et al., 2023c; Gu et al., 2023), Zhou et al. (2023) presented UniNER, a targeted distillation technique coupled with instruction tuning to develop an efficient open-domain NER model. Our research draws from these works and explores the capabilities of LLMs for biomedical NER, employing prompt design, strategic ICL example selection, and data augmentation via an external knowledge base to enhance performance.

Problem definition Assume data samples are represented as (X, Y) and the goal is to develop a model, denoted as $f : (X \times T) \rightarrow Y$, where X signifies the input set, T represents a predetermined set of entity types, and Y denotes the set of entity types. The task is to predict the entity type of each input word among the set T. We followed the standard practice of using the F1 score for evaluation purposes in both mention/token-level analyses.

Datasets We used three biomedical NER datasets with different entity types: I2B2 (Uzuner et al., 2011) which includes test, treatment, and problem entities, NCBI-disease (Doğan et al., 2014) consisting of the disease entity, and BC2GM (Smith et al., 2008) containing the gene entity.

3 Influence of Input-Output Format

Recent studies demonstrated the importance of prompt engineering for various tasks (Wang et al., 2023a; Gao et al., 2023b; Nori et al., 2023b). We studied the influence of input-output format by adapting TANL (Paolini et al., 2021) and DICE (Ma et al., 2022) for biomedical NER. In TANL, the task is framed as a translation task which involves augmenting the text by tagging entity types for each word directly within the text. The method is exemplified in Fig 1, showcasing how the text incorporates entity types.

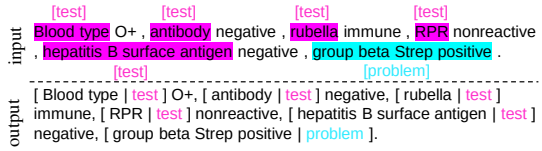

Figure 1: TANL input/output format for NER task.

Then, the generated output is decoded into the BIO format (Ramshaw and Marcus, 1999) for the assessment. In the refined DICE format, the input-output format involves adding a description for each entity type in a template following DEGREE (Hsu et al., 2021). Given an input text and corresponding labels, the desired output should be the input followed by the phrase "entity type is <entity_type>. <entity_description>. entity is <entity>" for each class label, e.g., *test*, *treatment*, and *problem* in the I2B2 dataset. Then, we expect the model to output the same template filling out the <entity> with the corresponding entities in the given text as demonstrated in Fig 2. For the entity types with no matched entities in the sentence, the output returns <entity> token in the output. Examples for the NCBI-disease and BC2GM datasets are presented in Appendix A.5.

Our experiments in Table 1 reveal that neither format consistently outperforms the other; rather, the effectiveness of each format varies depending on the complexity of the dataset and model size. To

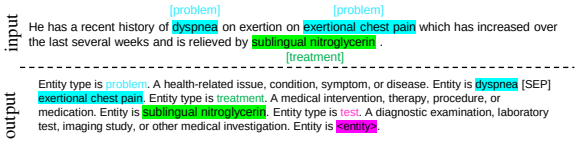

Figure 2: DICE input/output format for NER task.

Model	input-output format	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
GPT-3.5-turbo	DICE	41.2 /50.0	45.3 /**62.0**	**43.3** /**55.6**
	TANL	**52.9**/**59.7**	**46.5**/51.3	39.1/50.8
GPT-4	DICE	58.8/70.1	**68.1/77.8**	**57.1**/67.9
	TANL	**61.9/73.5**	67.5/70.0	56.4/**69.6**

Table 1: TANL vs. DICE format with GPT-3.5-turbo/GPT-4 . The superiority of any format varies with the complexity of the dataset and model size.

maintain consistency in the rest of our experiments, we opted for the TANL format, which offers a more straightforward pattern.

4 In-Context Examples Selection: A Key to Improving ICL Outcomes

In-context examples can be randomly chosen from the training set; however, researchers have demonstrated that the performance of ICL depends on the order and similarity of ICL examples to the test samples (Liu et al., 2021; Min et al., 2022; Gao et al., 2023a). Liu et al. (2021) presented Knn-Augmented in-conText Example selection (KATE). KATE identifies in-context examples selectively using nearest neighbor search on example embeddings, leading to better performance than random example selection. We tested KATE on TANL formatted examples with 16-shot ICL using four different LM encoders (w/o fine-tuning) to produce example embeddings. We used MPNET (Song et al., 2020) for its popularity and performance on sentence embedding benchmarks (Reimers and Gurevych, 2019) , SimCSE (Gao et al., 2021) for its documented performance as an alternative to standard sentence transformers, and BioClinicalBERT (Alsentzer et al., 2019) and BioClinicalRoBERTa (Gururangan et al., 2020) for their dominance on clinical data tasks (Lehman et al., 2023).

Our results summarized in Table 2 show that strategic in-context example selection via KATE outperforms random selection. BioClinicalRoBERTa achieved the best results among all example encoders tested. The strong performance of BioClinicalBERT and BioClinicalRoBERTa underscores the importance of using LM encoders pretrained on biomedical text when applying KATE for biomedical NER.

Model	KATE vs RS	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
GPT-3.5-turbo (ICL)	RS	52.9/59.7	46.6/51.3	39.1/50.8
	BioClinicalRoBERTa	66.1/77.4	**68.0**/77.7	**61.6/72.5**
	BioClinicalBERT	**67.0/78.9**	67.6/**78.8**	60.9/72.0
	MPNET	65.3/76.7	63.7/76.7	59.1/70.0
	SimCSE	65.2/76.1	61.6/76.1	57.8/68.8
	(Hu et al., 2024b)	49.3/ -	-	-
GPT4 (ICL)	RS	67.7/73.5	62.6/70.0	59.2/69.6
	BioClinicalRoBERTa	81.2/**88.4**	**79.3**/**88.3**	**72.4/80.7**
	BioClinicalBERT	**81.7**/88.1	**79.3**/88.0	71.9/79.4
	MPNET	80.7/87.5	79.8/87.4	71.1/80.2
	SimCSE	79.6/86.6	77.3/86.5	69.9/77.9
	(Hu et al., 2024b)	59.3/ -	-	-
BioBERT	fully supervised	- /87.3	- /**89.1**	- /83.8
BioClinicBERT	fully supervised	- /87.7	- /89.0	- /81.7
BioClinicRoBERTa	fully supervised	- /**89.7**	- /89.0	- /**87.0**

Table 2: 16-shot ICL for Random example selection (RS) vs. KATE method Vs MLMs with Mention/Token-level (M/T) analysis. KATE significantly outperforms random sampling in all settings, and LMs pre-trained on biomedical text outperform general domain encoders.

5 In-Context Learning or Fine-Tuning?

Within the scope of LLMs for biomedical applications, an essential question is whether to prompt a closed-source LLM via ICL or fine-tune an open-source one. Comparing two different LLMs employing divergent strategies is not straightforward. To provide some insight into this dilemma, we examined two key factors, performance and cost, for biomedical NER, and presented a detailed analysis under various experiment settings. This comparison offers valuable perspective into the right strategy given the task and dataset attributes. For fine-tuning, we used LoRA (Hu et al., 2021). Details can be found in Appendix A.5. The cost of fine-tuning comes from training an LLM on a large labeled dataset while the cost of ICL mainly comes from calling an API for each input query. For 16-shot ICL experiments, we calculated the cost based on the number of processed and generated tokens considering the average text size based on current LLM API pricing.[1] The estimated cost for the entire test set of each benchmark dataset considering the input text, prompt, and generated text size using the TANL format is summarized in Table 3. Referring to the OpenAI API for fine-tuning pricing, we also estimated the cost for fine-tuning LLama2-7B, summarized in Table 3. Interestingly, for the I2B2 dataset, GPT-3.5-turbo with a much cheaper cost outperforms fine-tuning Llama2-7B.

[1] https://openai.com/pricing

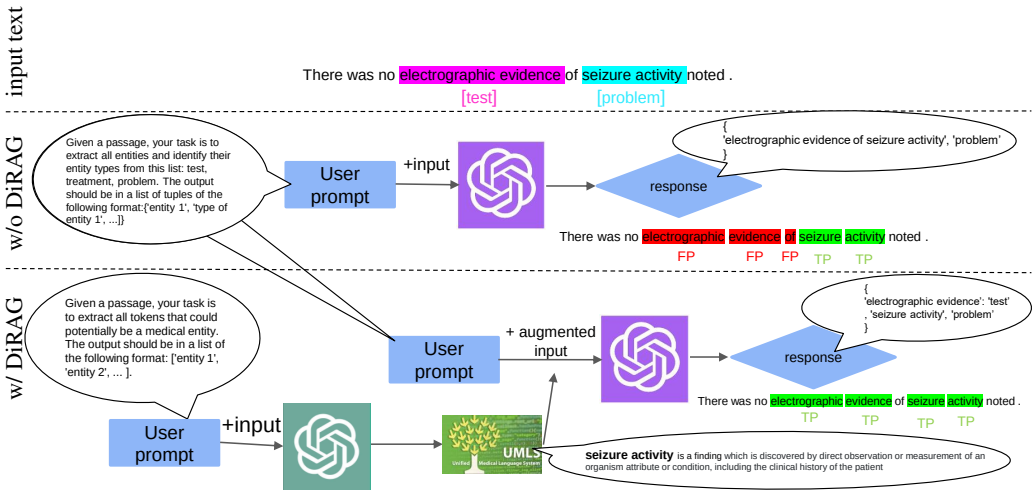

Figure 3: An overview of Dictionary-Infused RAG

	Model	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
Performance	GPT-3.5-turbo w/ KATE (ICL)	67.0/78.9	68.0/78.8	61.6/72.5
	GPT4 w/ KATE (ICL)	**81.7/88.4**	79.3/88.3	**72.4/80.7**
	Llama2-7B (FT)	61.2/76.2	**80.4/91.3**	68.1/75.1
Cost (T+I)	GPT3.5-turbo w/ KATE (ICL)	($0.35)	($0.11)	($1.34)
	GPT4 w/ KATE (ICL)	($10.42)	($3.12)	($40.13)
	Llama2-7B (FT)	($47.85+$7.4)	($23.5+$1.2)	($69.7+$12.9)

Table 3: Analysis of ICL vs fine-tuning LLMs: assessing performance and cost (Training + Inference) implications. Fine-tuning LLama2 exhibits superior outcomes on NCBI-disease, whereas GPT-4, enhanced by KATE using a biomedical encoder, achieves more favorable results on both the I2B2 and BC2GM datasets.

Model	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
UniversalNER (Zhou et al., 2023)	40.4/ -	60.4/ -	47.2/ -
(Rohanian et al., 2023) w/ GPT-3.5	-	33.4 / -	32.0 / -
(Hu et al., 2024b) w/ GPT-3.5-turbo	39.3/ -	-	-
(Hu et al., 2024b) w/ GPT-4	52.6/ -	-	-
GPT-3.5-turbo w/o DiRAG	41.9 /54.7	38.2 / 49.4	38.6 / 28.7
GPT-3.5-turbo w/ DiRAG	43.0 / 55.7	44.7 / 50.0	30.45 / 22.5
GPT-4 w/o DiRAG	46.3 / 59.1	55.7 /60.5	**52.1 / 58.4**
GPT-4 w/ DiRAG	**53.1 /62.8**	**61.0 /66.2**	51.1 / 55.0

Table 4: Zero-shot NER with GPT models w/ and w/o DiRAG vs. SOTA. DiRAG improved zero-shot NER significantly for I2B2 and NCBI-disease datasets for both GPT models. Results with confidence intervals are in the appendix.

6 Dictionary-Infused RAG

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) is a technique to enhance the capabilities of LLMs by integrating external information or knowledge into the generation process. This method involves retrieving relevant documents from a large corpus and providing this external knowledge in the input context to improve the quality and relevance of the generated text. Inspired by RAG, we developed a new method, DiRAG, to utilize UMLS as an external resource to augment the input data for the biomedical NER task. The process with detailed prompts is visualized in Fig 3, while an expanded view of the UMLS component is depicted in Fig 8. Unlike traditional RAG techniques that rely on embedding similarities to retrieve relevant documents, our approach initially employs the LLM to tackle a more straightforward task: identifying all words that could potentially qualify as medical named entities. Then, we look up each selected word in an external knowledge base, e.g., UMLS to augment the input data with useful information such as term definition. Then, we call the LLM with augmented input text. The process is visualized in Fig 8. We tested the approach on zero-shot NER and compared it with SOTA in Table 4. Our proposed approach enhanced the performance of both GPT versions on the I2B2 and NCBI-disease datasets significantly. DiRAG with GPT-4 achieved SOTA for zero-shot NER. Our approach proved ineffective for the BC2GM dataset due to the nature of the UMLS knowledge base which is predominantly tailored to medical terminology rather than biogenetics. We expect our approach to outperform GPT-4 on BC2GM with a more relevant knowledge base.

7 Conclusion

We explored LLMs for biomedical NER by customizing various prompting techniques. Through a detailed comparative analysis, we highlighted the vital role of ICL and the selection of contextually pertinent examples with biomedical text encoders for biomedical NER tasks. Moreover, our investigation into incorporating external medical knowledge resulted in a novel data augmentation approach, considerably advancing the capabilities of zero-shot biomedical NER with LLMs.

Limitations

While we have shown the potential of enhancing LLM performance for biomedical NER, the experiments in this paper are limited in two aspects mainly due to computational constraints. (1) TANL uses a straightforward text-to-text format while DICE uses additional descriptions. Future work could attempt to simplify DICE or combine it with TANL. Ablation studies on components of each format could help researchers design new prompt formatting strategies. (2) Our RAG-based method exclusively utilizes UMLS as the knowledge base, though it is limited in its vocabulary. For medical terms not covered by UMLS, we did not augment the input text. Other knowledge bases such as Wikipedia could serve as an alternative.

References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiusi Chen, Jyun-Yu Jiang, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, and Wei Wang. 2023. Minprompt: Graph-based minimal prompt data augmentation for few-shot question answering. *arXiv preprint arXiv:2310.05007*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032.

Shuzheng Gao, Xin-Cheng Wen, Cuiyun Gao, Wenxuan Wang, Hongyu Zhang, and Michael R Lyu. 2023a. What makes good in-context demonstrations for code intelligence tasks with llms? In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 761–773. IEEE.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M Churpek, and Majid Afshar. 2023b. Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv preprint arXiv:2308.14321*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023c. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Lisa Grossman Liu, Raymond H Grossman, Elliot G Mitchell, Chunhua Weng, Karthik Natarajan, George Hripcsak, and David K Vawdrey. 2021. A deep database of medical abbreviations and acronyms for natural language processing. *Scientific Data*, 8(1):149.

Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, et al. 2023. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. *arXiv preprint arXiv:2307.06439*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410*.

I Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, et al. 2021. Degree: A data-efficient generation-based event extraction model. *arXiv preprint arXiv:2108.12724*.

Danqing Hu, Bing Liu, Xiaofeng Zhu, Xudong Lu, and Nan Wu. 2024a. Zero-shot information extraction from radiological reports using chatgpt. *International Journal of Medical Informatics*, 183:105321.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024b. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15(5).

Amita Kumari, Anita Kumari, Amita Singh, Sanjeet K Singh, Ayesha Juhi, Anup Kumar D Dhanvijay, Mohammed Jaffer Pinjar, Himel Mondal, and Anoop Kumar Dhanvijay. 2023. Large language models in hematology case solving: a comparative study of chatgpt-3.5, google bard, and microsoft bing. *Cureus*, 15(8).

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? *arXiv preprint arXiv:2302.08091*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Mingyu Derek Ma, Alexander K Taylor, Wei Wang, and Nanyun Peng. 2022. Dice: data-efficient clinical event extraction with generative models. *arXiv preprint arXiv:2208.07989*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2021. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Omid Rohanian, Mohammadmahdi Nouriborji, and David A Clifton. 2023. Exploring the effectiveness of instruction tuning in biomedical language processing. *arXiv preprint arXiv:2401.00579*.

Minju Seo, Jinheon Baek, James Thorne, and Sung Ju Hwang. 2024. Retrieval-augmented data augmentation for low-resource domain tasks. *arXiv preprint arXiv:2402.13482*.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9:1–19.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Raju Vaishya, Anoop Misra, and Abhishek Vaish. 2023. Chatgpt: Is this version good for healthcare and research? *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 17(4):102744.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023a. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023b. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023c. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2021. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinformatics*, 37(17):2792–2794.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

A Appendix

A.1 TANL/DICE more examples

In Fig 4-8, we visualize two examples of each format for NCBI-disease and BC2GM datasets for more demonstration.

A.2 Benchmark datasets

We studied LLMs on I2b2, NCBI-disease, and BC2GM dataset. In the following, we provide some details about each.

I2B2: I2B2 is a collection of annotated clinical records that are used primarily for Clinical NER.The task involves identifying clinical terms such as medical problems, treatments, and tests from patient records. The dataset typically includes a large number of annotated clinical narratives that are de-identified to protect patient confidentiality. This makes it a rich resource for training and testing NER models.

NCBI-disease: This dataset is specifically curated for disease name recognition and normalization in biomedical texts. It comprises abstracts from PubMed annotated for disease mentions and linked to the NCBI disease database. The corpus is relatively smaller compared to i2b2 but is densely annotated, providing high-quality, fine-grained annotations of disease entities, which are crucial for models aimed at medical literature.

BC2GM: This dataset focuses on the recognition of gene and gene product mentions in PubMed abstracts that is a suitable dataset for biological NER. The BC2GM dataset is extensively annotated to include a wide range of gene and gene product mentions, reflecting the complex and varied ways these entities are referred to in scientific literature.

A.3 PEFT setting of Llama for fine-tuning

We fine-tuned Llama2-7B on the entire training set of each dataset for three epochs and maintained a batch size of 16, learning rate of 2e-4, and cap the maximum sequence length at 512, truncating

any sequences that exceeded this limit. The LoRA dropout rate is adjusted to 0.1, and the LoRA α and rank parameters are also set at 16 and 32 respectively. The training was done on 4 NVIDIA Tesla V100 GPUs for approximately 24, 12, and 63 hours for I2B2, NCBI-disease, and BC2GM respectively.

A.4 Few-shot and Zero-shot performances with Confidence Interval

We introduced both few-shot and zero-shot settings to comprehensively evaluate the versatility and generalization capabilities of our study across different levels of data availability. While it's true that the performance in the zero-shot setting is generally lower compared to the few-shot setting, this approach offers valuable insights into the model's behavior when no training examples are provided. The zero-shot setting, leveraging techniques like Retrieval-Augmented Generation (RAG), demonstrates the model's potential to utilize pre-existing knowledge embedded in its parameters and external sources effectively. This is particularly important for scenarios where labeled data is scarce or unavailable, making zero-shot learning a critical area of study to ensure broader applicability of the model in real-world applications. Moreover, the inclusion of both methodologies allows us to highlight the performance trade-offs and strengths of the model under different instructional paradigms, contributing to a more robust and nuanced understanding of its capabilities. We ran all experiments with different random seeds and reported the full results of Table 2, 3, and 4 with confidence Intervals In Tables 6, 7, and 8.

A.5 UMLS detail

In Fig 8, we visualize the process by which potential words suggested by the LLM are searched within the UMLS and demonstrate how the input is augmented to enhance zero-shot prompting in LLMs.

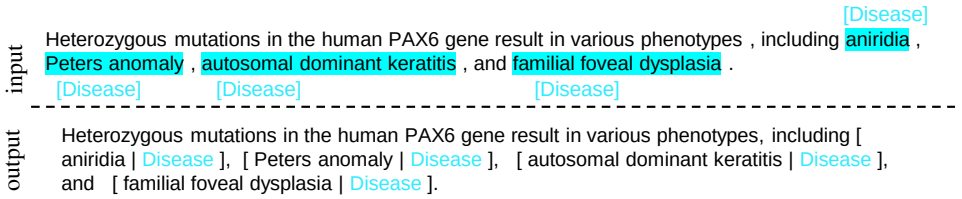

Figure 4: TANL input-output format example for NCBI-disease dataset

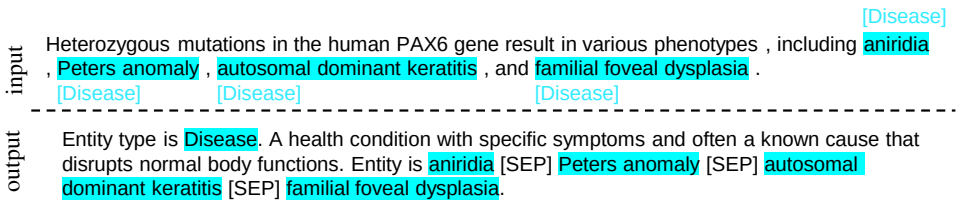

Figure 5: DICE input-output format example for NCBI-disease dataset

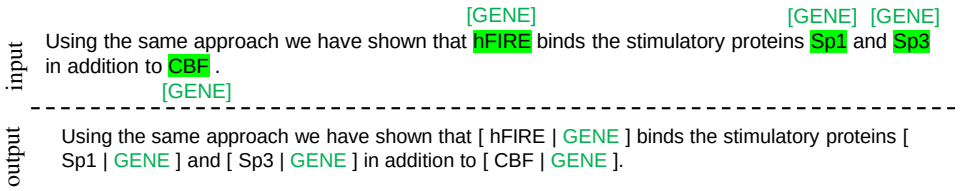

Figure 6: TANL input-output format example for BC2GM dataset

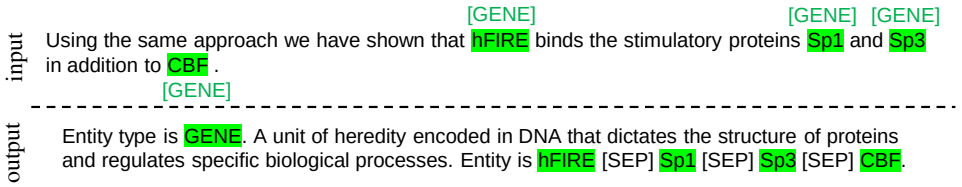

Figure 7: DICE input-output format example for BC2GM dataset

Model	input-output format	I2B2	Mention/Token NCBI-disease	BC2GM
GPT-3.5-turbo	DICE	41.2 ± 0.2 /50.0 ± 0.1	45.3 ± 0.2 **/62.0 ± 0.3**	**43.3 ± 0.5 /55.6 ± 0.4**
	TANL	**52.9 ± 0.3/59.7 ± 0.4**	**46.5 ± 0.5**/51.3 ± 0.5	39.1 ± 0.4/50.8 ± 0.5
GPT-4	DICE	58.8 ± 0.4 /70.1 ± 0.3	**68.1 ± 0.9/77.8 ± 1.1**	**57.1 ± 0.6**/67.9 ± 0.5
	TANL	**61.9 ± 0.3/73.5 ± 0.5**	67.5 ± 0.8/70.0 ± 0.6	56.4 ± 0.2/**69.6 ± 0.3**

Table 5: TANL vs. DICE format with GPT-3.5-turbo/GPT-4 with confidence intervals

Model	KATE vs RS	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
GPT-3.5-turbo (ICL)	RS	52.9 ± 0.3 / 59.7 ± 0.4	46.6 ± 0.5 / 51.3 ± 0.5	39.1 ± 0.4 / 50.8 ± 0.5
	BioClinicalRoBERTa	66.1 ± 0.4 / 77.4 ± 0.6	**68.0 ± 0.3** / 77.7 ± 0.2	**61.6 ± 0.5** / **72.5 ± 0.6**
	BioClinicalBERT	**67.0 ± 0.6/78.9 ± 0.5**	67.6 ± 0.1/**78.8± 0.1**	60.9 ± 0.7 /72.0 ± 0.5
	MPNET	65.3 ± 0.3 / 76.7 ± 0.2	63.7 ± 0.3 / 76.7 ± 0.3	59.1 ± 0.4 / 70.0 ± 0.4
	SimCSE	65.2 ± 0.2 / 76.1 ± 0.3	61.6 ± 0.4 / 76.1 ± 0.3	57.8 ± 0.5 / 68.8 ± 0.4
	(Hu et al., 2024b)	49.3/ -	-	-
GPT4 (ICL)	RS	67.7 ± 0.3 / 73.5 ± 0.5	62.6 ± 0.8 /70.0 ± 0.6	59.2 ± 0.2 / 69.6 ± 0.3
	BioClinicalRoBERTa	81.2 ± 0.3/**88.4 ± 0.6**	**79.3 ± 0.9/88.3 ± 0.8**	**72.4 ± 0.6/80.7 ± 0.5**
	BioClinicalBERT	**81.7 ± 0.4**/88.1 ± 0.3	**79.3 ± 0.4**/88.0 ± 0.3	71.9 ± 0.3 / 79.4 ± 0.3
	MPNET	80.7 ± 0.4 / 87.5 ± 0.5	79.8 ± 0.9 /87.4 ± 0.9	71.1 ± 1.1 / 80.2 ± 1.0
	SimCSE	79.6 ± 0.5 / 86.6 ± 0.4	77.3 ± 0.5 / 86.5 ± 0.8	69.9 ± 0.8 / 77.9 ± 0.5
	(Hu et al., 2024b)	59.3/ -	-	-
BioBERT	fully supervised	- /87.3	- **/89.1**	- /83.8
BioClinicBERT	fully supervised	- /87.7	- /89.0	- /81.7
BioClinicRoBERTa	fully supervised	- **/89.7**	- /89.0	- **/87.0**

Table 6: Random example selection (RS) vs. KATE with medical/non-medical encoders vs. fully supervised models with Mention/Token-level (M/T) analysis. KATE significantly outperforms random sampling in all settings, and LMs pre-trained on the biomedical text outperform strong, general domain encoders. HunFlair is added to the paper

	Model	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
Performance	GPT-3.5-turbo w/ KATE	67.0 ± 0.6 / 78.9 ± 0.5	68.0 ± 0.3 / 78.8 ± 0.1	61.6 ± 0.1 / 72.5 ± 0.6
	GPT4 w/ KATE	**81.7 ± 0.4 / 88.4 ± 0.6**	79.3 ± 0.4 / 88.3 ± 0.8	**72.4 ± 0.6 / 80.7 ± 0.4**
	Llama2-7B	61.2 ± 1.8 / 76.2 ± 1.3	**80.4 ± 0.9/ 91.3 ± 1.1**	68.1 ± 1.4 / 75.1 ± 1.3
Cost (T+I)	GPT3.5-turbo w/ KATE	($0.35)	($0.11)	($1.34)
	GPT4 w/ KATE	($10.42)	($3.12)	($40.13)
	Llama2-7B	($47.85+$7.4)	($23.5+$1.2)	($69.7+$12.9)

Table 7: Analysis of ICL vs fine-tuning LLMs: assessing performance and cost implications. Fine-tuning LLama2 exhibits superior outcomes on NCBI-disease, whereas GPT-4, enhanced by KATE using a biomedical encoder, achieves more favorable results on both the I2B2 and BC2GM datasets.

Model	I2B2 M/T	NCBI-disease M/T	BC2GM M/T
UniversalNER (Zhou et al., 2023)	40.4/ -	60.4/ -	47.2/ -
(Rohanian et al., 2023) w/ GPT-3.5	-	33.4 / -	32.0 / -
(Hu et al., 2024b) w/ GPT-3.5-turbo	39.3/ -	-	-
(Hu et al., 2024b) w/ GPT-4	52.6/ -	-	-
HunFlair (Weber et al., 2021)	0.0 / 0.0	24.8 / 36.1	28.2 / 22.7
GPT-3.5-turbo w/o DiRAG	41.9 ± 1.4 /54.7 ± 1.9	38.2 ± 1.7 / 49.4 ± 2.6	38.6 ± 1.0 / 28.7 ± 1.9
GPT-3.5-turbo w/ DiRAG	43.0 ± 0.9 / 55.7 ± 1.5	44.7 ± 0.5 / 50.0 ± 2.1	30.45 ± 1.6 / 22.5 ± 2.1
GPT-4 w/o DiRAG	46.3 ± 1.9 / 59.1 ± 2.7	55.7 ± 0.8 /60.5 ± 0.9	**52.1 ± 3.64 / 58.4 ± 1.3**
GPT-4 w/ DiRAG	**53.1 ± 1.1 / 62.8 ± 1.2**	**61.0 ± 0.6 /66.2 ± 0.5**	51.1 ± 2.0 / 55.0 ± 2.2

Table 8: Full results of Zero-shot NER with GPT-3.5-turbo and GPT-4, w/ and w/o DiRAG, and their comparision with SOTA. Our method improved zero-shot NER significantly for I2B2 and NCBI-disease datasets.

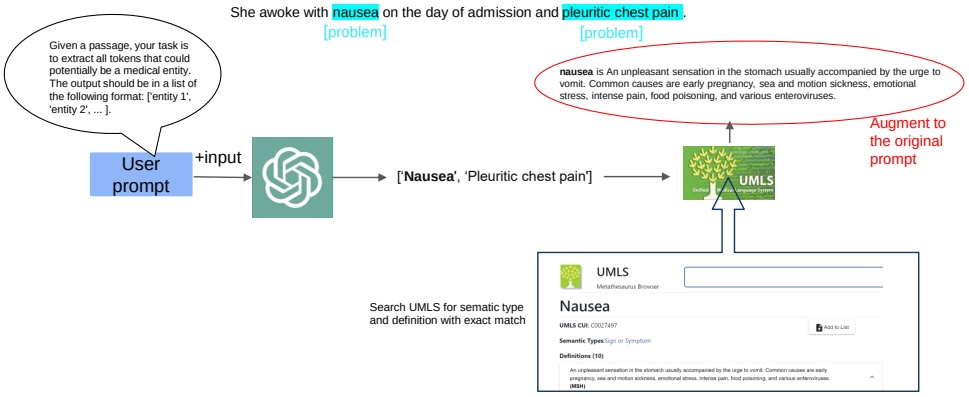

Figure 8: UMLS search. The GPT model is prompted for a simpler task of identifying all words that could potentially be a named entity. Then, the retrieved information from UMLS will augment the original input text for recalling the LLM