LLMs in Biomedical: A study on clinical Named Entity Recognition

Anonymous ACL submission

Abstract

Large Language Models (LLMs) demonstrate remarkable versatility in various NLP tasks but encounter distinct challenges in biomedicine due to medical language complexities and data scarcity. This paper investigates the application of LLMs in the medical domain by exploring strategies to enhance their performance for the Named-Entity Recognition (NER) task. Specifically, our study reveals the importance of meticulously designed prompts in biomedical. Strategic selection of in-context examples yields a notable improvement, showcasing $\sim 15 - 20\%$ increase in F1 score across all benchmark datasets for few-shot clinical NER. Additionally, our findings suggest that integrating external resources through prompting strategies can bridge the gap between general-purpose LLM proficiency and the specialized demands of medical NER. Leveraging a medical knowledge base, our proposed method inspired by Retrieval-Augmented Generation (RAG) can boost the F1 score of LLMs for zero-shot clinical NER. We will release the code upon publication.

1 Introduction

LLMs such as OpenAI's GPT-4 have demonstrated exceptional capabilities across diverse tasks and domains, exhibiting remarkable performance (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 lags behind (Zhou et al., 2023; Vaishya et al., 2023; Nori et al., 2023a). Acknowledging this gap, it is crucial to examine the unique challenges presented by the medical field that contribute to this discrepancy. Specifically, LLMs encounter challenges in medical Information Extraction (IE) (Gutierrez et al., 2022; Moradi et al., 2021) due to the scarcity of high-quality clinical data in their pre-training, compounded by the requirement for a nuanced comprehension of the text for this task (Gu et al., 2023). Medical entities can have multiple synonyms and abbreviations which makes it difficult for models to recognize (Grossman Liu et al., 2021). Besides, context sensitivity takes on even greater significance in 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 in biomedical. 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 in In-Context Learning (ICL) and demonstrate the significant impact of choosing more contextually similar examples to the input data in ICL for clinical NER. A key question that arises in the deployment of LLMs concerns the comparative advantage of closed-source LLM versus fine-tuning an open-source one. In our third study, we shed light on this question by presenting an assessment of the two main factors of performance and cost across various experiments. Furthermore, we explore the integration of external clinical knowledge to further 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), substantially improves zero-shot clinical 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 LLMs understanding for complex tasks (Lu et al., 2021; Kaddour et al., 2023; Webson and Pavlick, 2021). Researchers have explored different formats to examine how LLMs can handle IE tasks (Wang et al., 2023c; Gutierrez et al., 2022; Wang et al., 2023b). 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 levels on par with fully supervised baselines. Following this work, Gutierrez et al. (2022); Moradi et al. (2021) have shown that LLMs are not skilled few-shot learners in the biomedical domain. However, recent advancements, such as GPT-4, have increased LLM performance on various biomedical tasks (Tian et al., 2024; Hu et al., 2024a; Nori et al., 2023a) including biomedical domain Hu et al. (2024b). In the direction of knowledge distillation from LLMs (Wang et al., 2023c; Gu et al., 2023), UniNER (Zhou et al., 2023) presented 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 clinical NER, employing prompt design, strategic ICL example selection, and data augmentation via an external knowledge base to enhance performance.

Problem definition Assume the data sample is 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 collection of entity types, and Y denotes the set of entities types within the provided input. 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.

Datasets We use 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) formats for the biomedical NER prompt designing. 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 7, showcasing how the text is integrated with entity types.

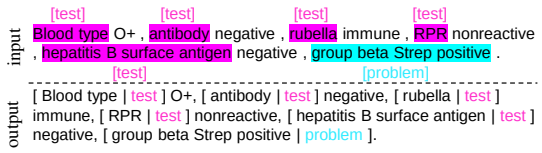

Figure 1: TANL input/output format for NER task.

Then, the generated output is decoded into the BIO format for the assessment. In the refined DICE format, the input-output format involves adding some description for each entity type in a template following DEGREE (Hsu et al., 2021). Given an input text and their 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 of NCBI-disease and BC2GM datasets are presented in Appendix A.

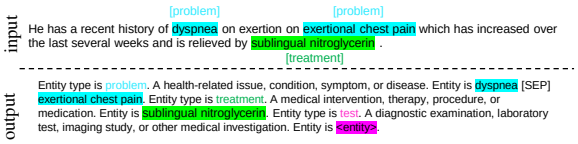

Figure 2: DICE input/output format for NER task.

	input-output format	I2B2	NCBI-disease	BC2GM
GPT-3.5-turbo	DICE	50.0	**62.0**	**55.6**
	TANL	**59.7**	51.3	50.8
GPT-4	DICE	70.1	**77.8**	67.9
	TANL	**73.5**	70.0	**69.6**

Table 1: Comparison of TANL and DICE using GPT-3.5-turbo/GPT-4 on three benchmark datasets. The superiority of any single format varies with the complexity of the dataset and model size.

	KATE method vs random sampling	I2B2	NCBI-disease	BC2GM
GPT-3.5-turbo	Random sampling	59.7	51.3	50.8
	BioClinicalRoBERTa	77.4	77.7	**72.5**
	BioClinicBERT	**78.9**	**78.8**	72.0
	MPNET	76.7	76.7	70.0
	SimCSE	76.1	76.1	68.8
	(Hu et al., 2024b)	49.3	-	-
GPT4	Random sampling	73.5	70.0	69.6
	BioClinicalRoBERTa	**88.4**	**88.3**	**80.7**
	BioClinicBERT	88.1	88.0	79.4
	MPNET	87.5	87.4	80.2
	SimCSE	86.6	86.5	77.9
	(Hu et al., 2024b)	59.3	-	-

Table 2: The influence of KATE approach with medical/non-medical encoders on ICL compared with random sampling and (Hu et al., 2024b). KATE significantly outperforms random sampling in all settings

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 maintain consistency in the rest of our experiments, we opted for the TANL format, in which the input-output relationship exhibits 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 find the performance of ICL sensitive to the order and similarity of ICL examples to the test examples (Liu et al., 2021; Min et al., 2022; Gao et al., 2023a; Zhang et al., 2023). Liu et al. (2021) presented Knn-Augmented in-conText Example selection (KATE). KATE identifies in-context examples selectively using the 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[1], 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.

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.

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. On the other hand, 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 clinical 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, and its details can ben found in Appendix A. The cost of fine-tuning comes from training an LLM on a large labeled dataset while the cost of ICL mainly comes from calling the 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 the current LLM API pricing[2], 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 which is summarized in Table 3. Interestingly, for the I2B2 dataset, GPT-3.5-turbo with a much cheaper cost outperforms fine-tuning Llama2-7B.

6 Dictionary-Infused RAG

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) is a technique to enhance the ca-

[1] https://www.sbert.net/docs/pretrained_models.html

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

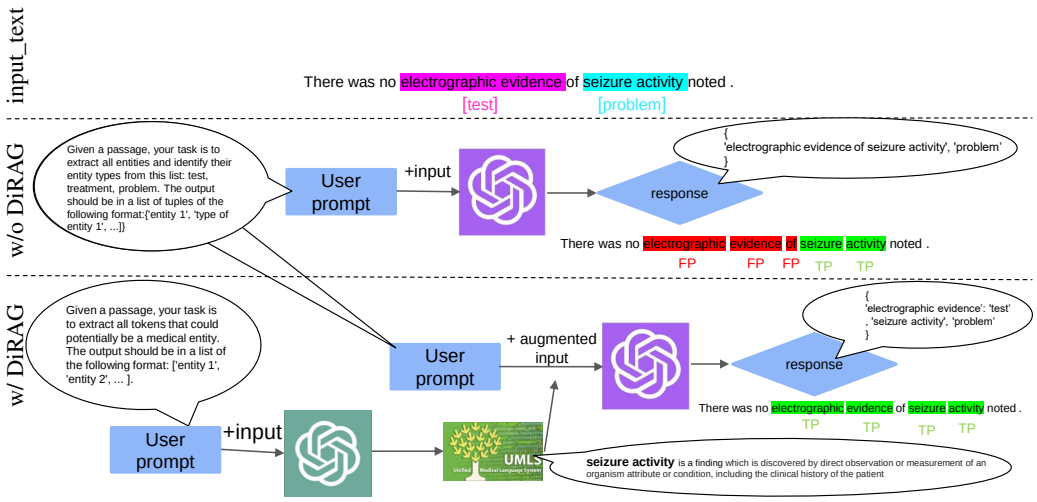

Figure 3: An overview of Dictionary-Infused RAG

	Language model	I2B2	NCBI-disease	BC2GM
Performance	GPT-3.5-turbo w/ KATE	78.9	78.8	72.5
	GPT4 w/ KATE	**88.4**	88.3	**80.7**
	Llama2-7B	76.2	**91.3**	75.1
Cost	GPT3.5-turbo w/ KATE	$0.35	$0.11	$1.34
	GPT4 w/ KATE	$10.42	$3.12	$40.13
	Llama2-7B	$5.23	$5.24	$5.69

Table 3: 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.

	I2B2	NCBI-disease	BC2GM
UniversalNER (Zhou et al., 2023)	40.4	61.4	47.2
(Rohanian et al., 2023) w/ GPT-3.5	-	33.4	32.0
(Hu et al., 2024b) w/ GPT-3.5-turbo	43.6	-	-
(Hu et al., 2024b) w/ GPT-4	52.6	-	-
GPT-3.5-turbo w/o DiRAG	33.12	38.1	31.5
GPT-3.5-turbo w/ DiRAG	43.55	45.0	28.8
GPT-4 w/o DiRAG	54.48	61.64	**57.2**
GPT-4 w/ DiRAG	**60.04**	**62.17**	53.83

Table 4: Zero-shot clinical NER using GPT-3.5-turbo and GPT-4 w/ and w/o DiRAG in comparison to SOTA. Our method improved zero-shot NER significantly for I2B2 and NCBI-disease datasets.

pabilities 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 clinical NER task. The process with detailed prompts is visualized in Fig 3. Unlike traditional RAG techniques that rely on embedding similarities to retrieve relevant documents, our approach initially employs GPT-3.5-turbo to tackle a more straightforward task: identifying all words within an input text that could potentially qualify as 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 definition. Finally, we call GPT-4 with augmented input text. A similar prompting format to UniNER is used to do a fair comparison. We tested this approach on zero-shot NER and compared it with SOTA in Table 4. Our proposed strategy with GPT-4 outperforms the baseline on I2B2 and NCBI-disease datasets with an F1 score improvement of 5.56% and 0.53% respectively. Additionally, the approach enhanced the performance of GPT-3.5-turbo on both the I2B2 and NCBI-disease datasets by 10.43% and 6.9% respectively. It 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 if a more relevant knowledge base is employed.

7 Conclusion

In this paper, we explored LLMs by customizing various prompting techniques specifically for biomedical NER. Through a detailed comparative analysis, we have highlighted the vital role of ICL and the selection of contextually pertinent examples, demonstrating their profound effect on enhancing clinical NER tasks. Moreover, our investigation into incorporating external clinical knowledge bases resulted in proposing a novel data augmentation approach, considerably advancing the capabilities of zero-shot clinical NER.

Limitations

While we have shown the potential of enhancing LLM performance for clinical NER, the experiments in this paper are limited in two aspects mainly due to computational constraints. (1) TANL format has a more straightforward text-to-text while DICE uses additional descriptions in its format. It is unclear how redesigning TANL by incorporating additional descriptions as in DICE could work in our experiment setting. (2) Our RAG-based method exclusively utilizes UMLS as the knowledge base which is limited in its vocabulary. For medical terms not covered by UMLS, we did not augment the input text, though other knowledge bases such as Wikipedia could potentially 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.

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.

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).

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*.

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

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

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*.

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.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023. What makes good examples for visual in-context learning? *arXiv preprint arXiv:2301.13670*.

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

Fine-tunign setting for Llama We fine-tuned Llama2-7B 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 an approximate of 2 days.

TANL and DICE more examples In this section, we visualize two examples of each format for NCBI-disease and BC2GM datasets for demonstration.

input

[Disease]

Heterozygous mutations in the human PAX6 gene result in various phenotypes , including aniridia , Peters anomaly , autosomal dominant keratitis , and familial foveal dysplasia .

[Disease] [Disease] [Disease]

output

Heterozygous mutations in the human PAX6 gene result in various phenotypes, including [aniridia | Disease], [Peters anomaly | Disease], [autosomal dominant keratitis | Disease], and [familial foveal dysplasia | Disease].

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

input

[Disease]

Heterozygous mutations in the human PAX6 gene result in various phenotypes , including aniridia , Peters anomaly , autosomal dominant keratitis , and familial foveal dysplasia .

[Disease] [Disease] [Disease]

output

Entity type is Disease. A health condition with specific symptoms and often a known cause that disrupts normal body functions. Entity is aniridia [SEP] Peters anomaly [SEP] autosomal dominant keratitis [SEP] familial foveal dysplasia.

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

input

[GENE] [GENE] [GENE]

Using the same approach we have shown that hFIRE binds the stimulatory proteins Sp1 and Sp3 in addition to CBF .

[GENE]

output

Using the same approach we have shown that [hFIRE | GENE] binds the stimulatory proteins [Sp1 | GENE] and [Sp3 | GENE] in addition to [CBF | GENE].

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

input

[GENE] [GENE] [GENE]

Using the same approach we have shown that hFIRE binds the stimulatory proteins Sp1 and Sp3 in addition to CBF .

[GENE]

output

Entity type is GENE. A unit of heredity encoded in DNA that dictates the structure of proteins and regulates specific biological processes. Entity is hFIRE [SEP] Sp1 [SEP] Sp3 [SEP] CBF.

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