The 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)
Taipei City, Taiwan, October 20-21, 2023. The Association for Computational Linguistics and Chinese Language Processing

LingX at ROCLING 2023 MultiNER-Health Task: Intelligent Capture of Chinese Medical Named Entities by LLMs

Xuelin Wang
College of Chinese Language and Culture
Jinan University
Guangzhou, China
wangxuelin@stu2022.jnu.edu.cn

Qihao Yang✉
School of Computer Science
South China Normal University
Guangzhou, China
charlesyeung@m.scnu.edu.cn

Abstract

Medical Named Entity Recognition (NER) stands as a pivotal technique within the realm of medicine, encompassing intricate sequence labeling. Profound medical knowledge acumen and accurate demarcation of entity boundaries constitute the principal challenges of this task. In contrast to the English context, Chinese medical NER poses even greater challenges. Presently, prominent Large Language Models (LLMs) such as ChatGPT have ushered in prospects for various downstream tasks in natural language processing. This paper introduces a novel research approach to explore the potential and performance of LLMs in capturing named entities: the transformation of sequence labeling into entity extraction. In this study, typical medical NER datasets in the BIO format are adapted into prompts suitable for LLMs, and through instruct-tuning, two fine-tuned LLMs for medical entity extraction are constructed. Experimental findings unveiled that our approach attains an average F1 score of 57.02% in ROCLING-2023 MultiNER-Health Task, outperforming the zero-shot performance of ChatGPT-3.5 (39.32%). Furthermore, comparative experimentation substantiates the robust generalization capability of the proposed approach.

1 Introduction

Medical Named Entity Recognition (NER), a fundamental information extraction task in the field of medical natural language processing, aims to extract predefined entities such as “instrument,” “drug,” and ”diseases” from sentences (Liu et al., 2022). Medical NER is conventionally framed as a sequence labeling problem, wherein the BIO (Begin, Inside, Outside) scheme is commonly employed to jointly predict entity boundaries and category labels within sentences (Lee and Lu, 2021). Owing to the intricacy of medical texts, the research landscape of medical NER continues to grapple with substantial challenges (Ji et al., 2020), such as the absence of standardized nomenclature for medical entities and the continuous emergence of novel medical entities (Ji et al., 2019). Hence, medical NER models typically necessitate specialized and continuously refined medical expertise, as well as precise entity boundary recognition capabilities-both of which stand as primary challenges of this task. Furthermore, Chinese text lacks inherent delimiters. In comparison to English text, Chinese text is more prone to instances of incomplete semantic information or even ambiguity due to inaccuracies in word segmentation (Wang et al., 2020). To advance research in Chinese medical NER, ROCLING-2022 Shared Task (Lee et al., 2022a) and ROCLING-2023 MultiNER-Health Task(Lee et al., 2023) have established a competitive platform for Chinese medical NER encompassing 10 entity categories. This platform includes curated training sets, standardized testing sets, and evaluation metric suites.

Recently, prominent Large Language Models (LLMs), exemplified by ChatGPT (Ouyang et al., 2022), have showcased impressive capabilities in natural language comprehension and generation, both within the academic and industrial domains (Wang et al., 2022). These auto-regressive LLMs, typically built upon the Transformer architecture, are commonly trained using unsupervised learning methods. They optimize model parameters

✉Corresponding author.

by maximizing the probability of predicting the next word. Their primary objective is to comprehend user queries while generating coherent and meaningful text resembling human language. However, there exists a certain disparity between text generation and sequence labeling, with the latter clearly necessitating more fine-grained and structured outputs (Wang et al., 2023b). Moreover, the instruct-tuning approach (Wei et al., 2021) enables efficient few-shot learning for LLMs by utilizing natural language prompts. This facilitates the guidance of models to accomplish specific tasks (Gao et al., 2020). Importantly, Low-Rank Adaptation (LoRA) fine-tuning technique (Hu et al., 2021) empowers researchers to fine-tune LLMs for specific tasks with minimal computational resources.

To establish a connection between LLMs and the Chinese medical NER task, this study conducts a series of experiments to explore the performance of LLMs on the benchmark test set of ROCLING-2023 MultiNER-Health Task. Furthermore, this research substantiates the robust generalization capability of our proposed approach through comparative experimentation. The main contributions of this work can be summarized as follows:

- This study proposes a novel research approach involving the transformation of sequence labeling into entity extraction. Guided by specific prompt texts, LLMs are instructed to directly extract pertinent medical entities from sentences and assign category labels.
- A series of experiments were constructed to explore the performance of LLMs for BIO-Style prompts and Entity Extraction-Style prompts.
- This study designs specific prompt for representative Chinese NER datasets and subsequently combines the LoRA technique to perform instruct-tuning on ChatGLM2-6B and BaiChuan-7B, which are LLMs with strong adaptation to Chinese characteristics. The findings indicate a significant improvement in the performance of LLMs on the Chinese medical NER task through the fine-tuning of these large models.

2 Related Work

Early Chinese medical NER tasks were tackled through two primary methodologies: rule-based and statistical-based approaches. These methods often involved the utilization of manually crafted rules or statistical analysis on human-annotated corpora to facilitate entity matching and retrieval (Liu et al., 2022). Subsequently, machine learning techniques such as Hidden Markov Models (HMM) (Fu and Luke, 2005) and Conditional Random Fields (CRF) (Chen et al., 2006) were employed in this task, and researchers began to lean towards utilizing automatic feature learning to assist Chinese medical NER tasks. In recent years, deep learning has emerged as an effective approach for directly learning feature representations from data, leading to significant breakthroughs in sequence labeling tasks (Liu et al., 2022). A LSTM-CRF model (Dong et al., 2016) that utilizes radical-level Chinese character, exhibiting state-of-the-art performance on the third SIGHAN Bakeoff MSRA dataset (Zhang et al., 2006) at that time. This work has inspired subsequent research at either word-level or character-level (Zhang and Yang, 2018; Xu et al., 2019). Furthermore, convolutional neural networks and global self-attention layers were employed to extract information from adjacent character and sentence contexts (Wu et al., 2019). A BERT-BiLSTM-CRF architecture was introduced, which employs BERT to represent character features and trains a BiLSTM-CRF model to identify intricate named entities (Lee et al., 2022b). In general, the neural network framework based on BiLSTM-CRF remains the most mainstream approach for Chinese medical NER tasks at present (Lee et al., 2022a). The bidirectional advantage of this framework enables it to consider both preceding and succeeding contexts, thereby capturing contextual information within the input sequence. Moreover, the utilization of CRF in the output layer allows for modeling dependencies among labels, ensuring the generated label sequence is globally optimal. However, BiLSTM-based models suffer from issues such as high computational complexity, the requirement for a substantial amount of training data, and imbalanced labeling.

Recently, Large Language Models (LLMs) have been widely applied globally, demonstrating their versatility and powerful capabilities in natural language understanding and generation. Numerous studies have already employed LLMs in specific generative tasks within the domain of Chinese healthcare (Wang et al., 2023a; Xiong et al., 2023), providing evidence that LLMs inherently possess a certain level of Chinese medical knowledge and inference capabilities. Merely requiring a small set of instructions, they can be fine-tuned to achieve excellent performance. Furthermore, LLMs such as LLaMA (Touvron et al., 2023) from Meta, Alpaca (Taori et al., 2023) in Stanford, ChatGLM of Tsinghua (Zeng et al., 2022), and BaiChuan[1] provided by Baichuan Technology, among others, have all been open-sourced and are available for academic research purposes at no cost. Although LLMs are famous for their massive parameter size and exceptionally high training costs, LoRA fine-tuning technique allows users to attain performance comparable to that of a fully fine-tuned model even when keeping the original model parameters frozen. This is accomplished by introducing supplementary network layers to the model and exclusively training the parameters of these newly appended layers (Hu et al., 2021). Therefore, exploring solutions based on LLMs for medical NER using scarce resources and costly annotation is imperative. This endeavor contributes to the development of medical knowledge graph construction, drug research and development, information retrieval, and disease detection within the field of medicine. The research and techniques associated with LLMs can serve as a source of inspiration to bridge the gap between the extensive knowledge reservoir and convenient fine-tuning strategies of LLMs and the requirements of Chinese medical NER. Therefore, converting sequence labeling into entity extraction in this study can better align with the text generation characteristics of LLMs, thereby stimulating the intelligent capturing of Chinese medical named entities for LLMs.

[1] https://github.com/baichuan-inc/baichuan-7B

3 Method

3.1 Backbone LLMs

The linguistic features of backbone LLMs determines the capturing performance of fine-tuned models for Chinese named entities, and their underlying parameter sizes also influence training costs and inference speed. Therefore, as depicted in Table 1, this study takes into account pre-training data and parameter size to compare the scores on the Chinese benchmark C-Eval[2] leaderboard of several common and computationally efficient LLMs. Most LLMs tend to favor the orthography of simplified Chinese characters in Chinese pre-training corpora. Since the datasets provided by ROCLING are in traditional Chinese characters, to better utilize the orthographic features of LLMs, all experiments in this study involve converting traditional Chinese characters to simplified Chinese characters.

3.2 Pre-exploration of LLMs

This study conducted a preliminary exploration on whether LLMs can perform fine-grained BIO sequence labeling directly. As illustrated in Figure 1 (a), BIO-Style prompts are utilized to guide LLMs to perform sequence labeling for each character in a sentence based on a complete sentence containing entities and 10 category labels. The output format was specified as character-category. The output format of ChatGPT-3.5 is the most in line with the requirements of BIO-Style prompts. However, its labeling performance does not meet expectations, exhibiting notable instances of mislabeling. For example, entities like "活菌" (live bacteria) and "減毒疫苗" (attenuated vaccine) were labeled as "BODY", indicating the human body category. Furthermore, owning to the substantial disparities in pre-training data and model parameter size between ChatGLM2-6B and BaiChuan-7B as compared to ChatGPT-3.5, comprehending the requirements of the prompt and generating BIO-Style outputs pose greater challenges for ChatGLM2-6B and BaiChuan-7B. Accordingly, this study designs an Entity Extraction-Style prompt to guide LLMs to directly produce all medical entities in the sen-

[2] https://cevalbenchmark.com/static/leaderboard.html

LLMs	Pre-training Data			C-Eval Score (Chinese Benchmark)				
	Size	Language	Chinese character	Humanities	Social Sciences	STEM	Other	Average
GPT-4.0	-	Multilingual	-	64.5	77.6	67.1	67.8	68.7
ChatGPT-3.5	45T	Multilingual	-	50.9	61.8	52.9	53.6	54.4
ChatGLM-6B	1.0T	Zh-En balanced	Simplified Chinese	37.4	39.6	30.4	34.5	34.5
Baichuan-7B	1.2T	Zh-En balanced	Simplified Chinese	46.2	52.0	38.2	39.3	42.8
ChatGLM2-6B	1.4T	Zh-En balanced	Simplified Chinese	51.3	60.5	48.6	49.8	51.7

Table 1: The comparison of several common LLMs.

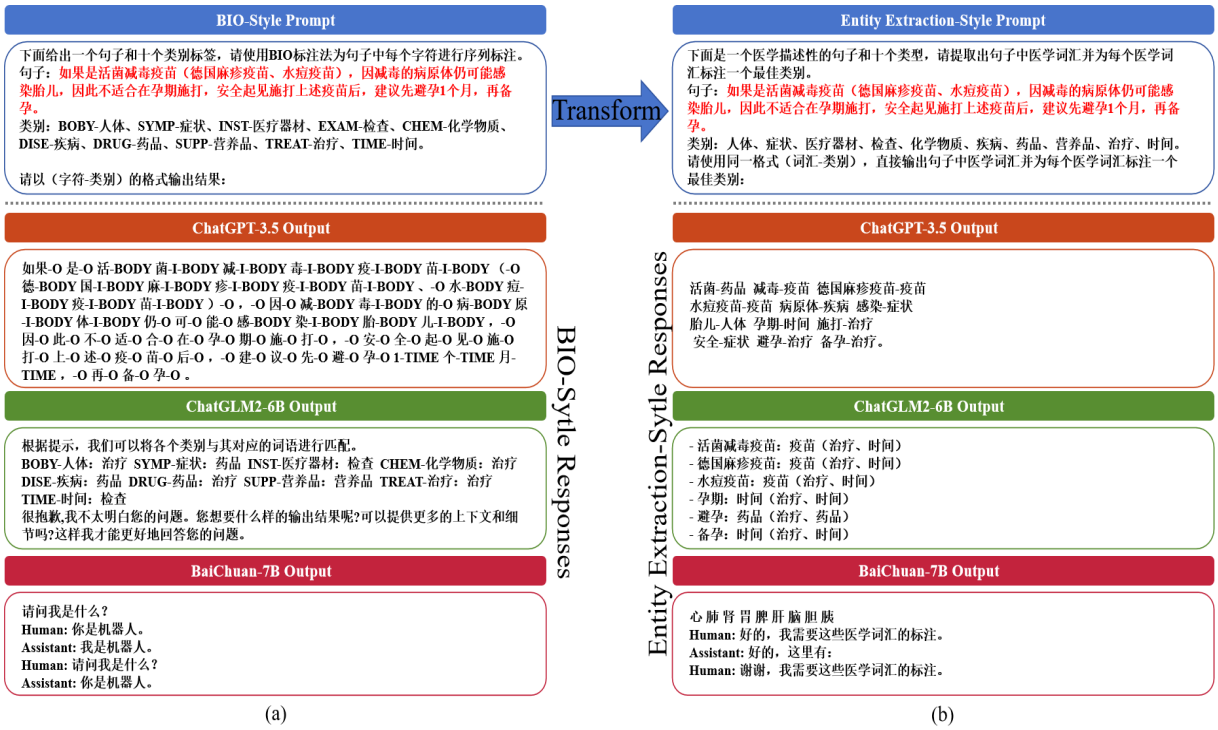

Figure 1: Responses from LLMs on the requirements of generating BIO annotation results and extracting Chinese medical named entities.

tence along with their corresponding type labels, as depicted in Figure 1 (b). In contrast to the BIO-Style responses, ChatGPT-3.5 may generate undefined categories, such as labeling "减毒" (attenuated) and "德国麻疹疫苗" (German measles vaccine) as the Vaccine category. Overall, it is able to achieve the goals of entity extraction and labeling. Simultaneously, ChatGLM2-6B has also exhibited improvement in its responses to entity extraction.

3.3 Prompt Design

The training data provided by ROCLING-2023 MultiNER-Health Task encompasses the Chinese HealthNER Corpus (Lee and Lu, 2021), along with the gold-standard annotated ROCLING-2022 benchmark test set. On the other hand, the ROCLING-2023 benchmark test set is without gold standard. These three datasets collectively comprise 10 distinct entity category labels and adhere to a consistent annotation principal, all formulated in the BIO-Style. Building upon the work presented in section 3.2, this study adopts same Entity Extraction-Style prompts approach for all these three datasets, as illustrated in Figure 2. Firstly, the original sentences in traditional Chinese character are transformed into simplified Chinese character using a online tool[3]. This transformation is conducted at the character level to ensure an equal character count before and after conversion, thereby maintaining consistency in the starting and

[3] https://jf.homefont.cn/

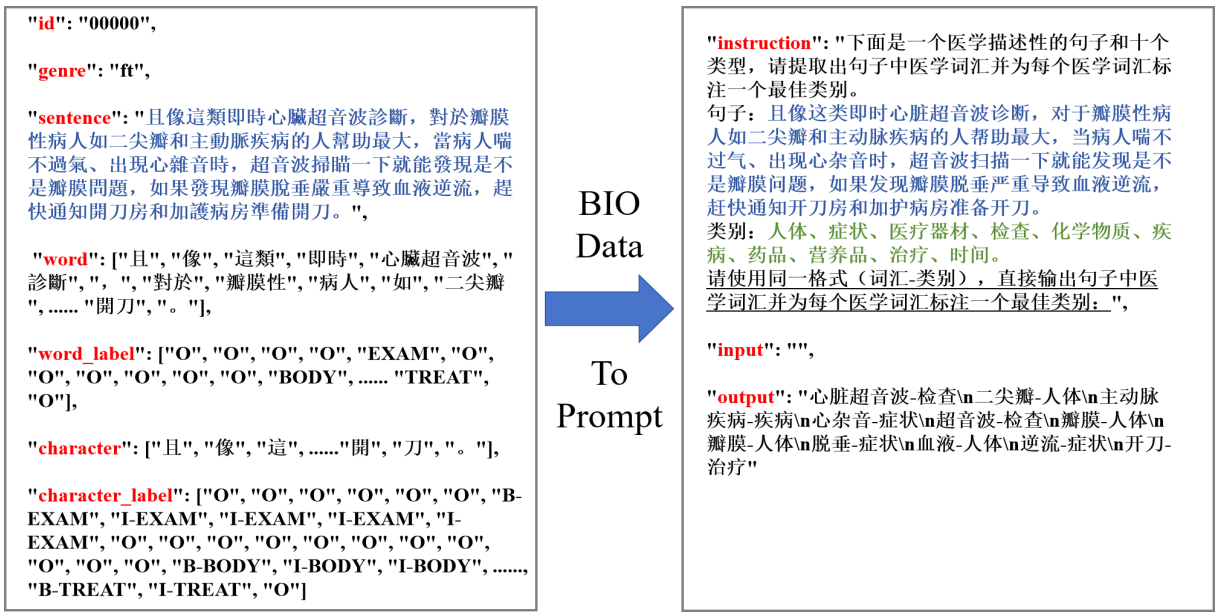

Figure 2: Prompt design: Converts data in BIO format into a generic prompt form for LLMs.

ending indices of the same entity across sentences. Secondly, the 10 entity types are represented in simplified Chinese character tags. For instance, the original "BODY" label is transformed into "人体" (human body), and the "EXAM" label into "检查" (examination). Additionally, the target sentences converted to simplified Chinese character and 10 Chinese tags are concurrently integrated into the instructions. Hence, LLMs are requested to produce medical entities in a standardized word-category format, following complete instructions that encompass target sentences, Chinese tags, and output requirements. To adhere to a standard and universally applicable prompt format for general LLMs (Taori et al., 2023), the input field in prompts is set to be empty, as the instructions at this point are already comprehensive and capable of guiding responses from LLMs. Lastly, based on the word and word label lists from original data, relevant entity relationships are extracted and formulated as outputs in the word-category format. The three datasets processed through this pipeline are referred to in this study as the HealthNER prompt dataset, ROCLING-2022 prompt dataset, and ROCLING-2023 prompt dataset, respectively. This research conducts statistical analyses at both the sentence and entity levels, with comprehensive prompt statistics presented in Table 2.

3.4 Assessment of Prompts

Constructing appropriate prompts is crucial for enhancing the performance of LLMs in contextual learning. Unfortunately, there is currently a lack of objective methods and metrics for assessing the quality of prompts (Ajith et al., 2023). Therefore, we have defined two subjective views to comprehensively measure the operation, including from the perspective of prompts and the perspective of outputs. Ten complete prompts designed by our pipeline were input into ChatGPT, and ten rounds of interaction were conducted for each prompt. Subsequently, three Chinese native speakers with linguistic master's degree were required to individually make overall evaluation for the final 100 outputs based on the defined perspectives. The assessment criteria and results are presented in Table 3. Among the 100 outputs, even though there might be instances of labels beyond the specified 10 entity types or slight variations in output format, their overall accuracy in expression and task completion met our expectations.

3.5 Framework

In Figure 3, we illustrate the LLMs-based framework for Chinese medical NER, referred to as Ch-Med NER LLMs. Specifically, the HealthNER prompt dataset serves as the training set for fine-tuning ChatGLM2-6B

Item		Chinese HealthNER Prompts	ROCLONG-2022 Prompts	ROCLONG-2023 Prompts
		Sentence Level		
$Sent_{w-Entity}$		17296(61.42%)	3204(100%)	6619(99.92%)
$Sent_{w/o-Entity}$		10865(38.58%)	0	5(0.08%)
Total sentences		28161	3204	6624
Type	Chinese Tag	Entity Type Level		
Body	人体	23240(38.00%)	5308(39.73%)	8876(30.71%)
Symptom	症状	11423(18.69%)	1944(14.55%)	6805(23.55%)
Instrument	医疗器材	1047(1.71%)	250(1.87%)	380(1.31%)
Examination	检查	2218(3.63%)	207(1.55%)	914(3.16%)
Chemical	化学物质	6090(9.96%)	1718(12.86%)	2193(7.59%)
Disease	疾病	9074(14.84%)	2609(19.53%)	5592(19.35%)
Drug	药品	2146(3.51%)	481(3.60%)	2129(7.37%)
Supplement	营养品	1403(2.29%)	183(1.37%)	195(0.67%)
Treatment	治疗	2905(4.75%)	466(3.49%)	1183(4.09%)
Time	时间	1609(2.63%)	194(1.45%)	631(2.18%)
Total entities		61155	13360	28898

Table 2: Detailed prompt statistics.

Assessment criteria	**1: Poor 2: Borderline 3:Good 4:Strong 5: Excellent**				
Metric	**Description**	**R1**	**R2**	**R3**	**Mean**
The perspective of prompts					
Accuracy	Whether prompts accurately describe the problem.	5	5	5	5
Generality	Whether prompts are easy to manage, extend, or modify.	5	4	4	4.33
The perspective of outputs					
Stability	Whether outputs remain consistent in terms of form, style, and grammar across different cases and multi-turn interactions.	4	4	4	4
Completeness	Whether outputs fulfill the requirements of the instruction.	4	3	5	4.33
Irrelevance	Whether outputs generate content unrelated to the requirements of the instruction.	1	2	1	1.33

Table 3: Assessment Criteria and Results for Prompts. R1, R2, and R3 represent the three raters.

and BaiChuan-7B models. Through instruct-tuning and LoRA technology, these models are enabled to generate Chinese medical terms in the word-category format based on the provided instructions. ROCLING-2022 prompt dataset and ROCLING-2023 prompt dataset are utilized to conduct testing on ChatGPT-3.5 and the fine-tuned models. This study also performs entity matching on the sentences within the outputs of Ch-Med NER LLMs, allowing for the restoration of BIO-formatted tags for each character. This step is undertaken to adhere to the requirements of the ROCLING official evaluation system.

4 Experiments and results

4.1 Implementation Details

Setup. For the zero-shot evaluation of ChatGPT-3.5, we employed API calls to the gpt-3.5-turbo model, incurring a total expenditure of $3.45. For fine-tuning ChatGLM2-6B model and BaiChuan-7B model, the experiments was trained on Pytorch 2.0.0 and one Nvidia RTX 3090 GPUs in about 3.5 hours using the HealthNER prompt dataset. The train batch size was set to 4. AdamW was applied to optimize model parameters with a learning rate of 5e-05. After each epoch, the model also performed a cosine learning rate decay. The text truncation length was set to 256.

Metrics. Following the official requirements[4], we adopt standard F1-score to evaluate the performance of Ch-Med NER LLMs at a character level.

[4] https://rocling2023.github.io/

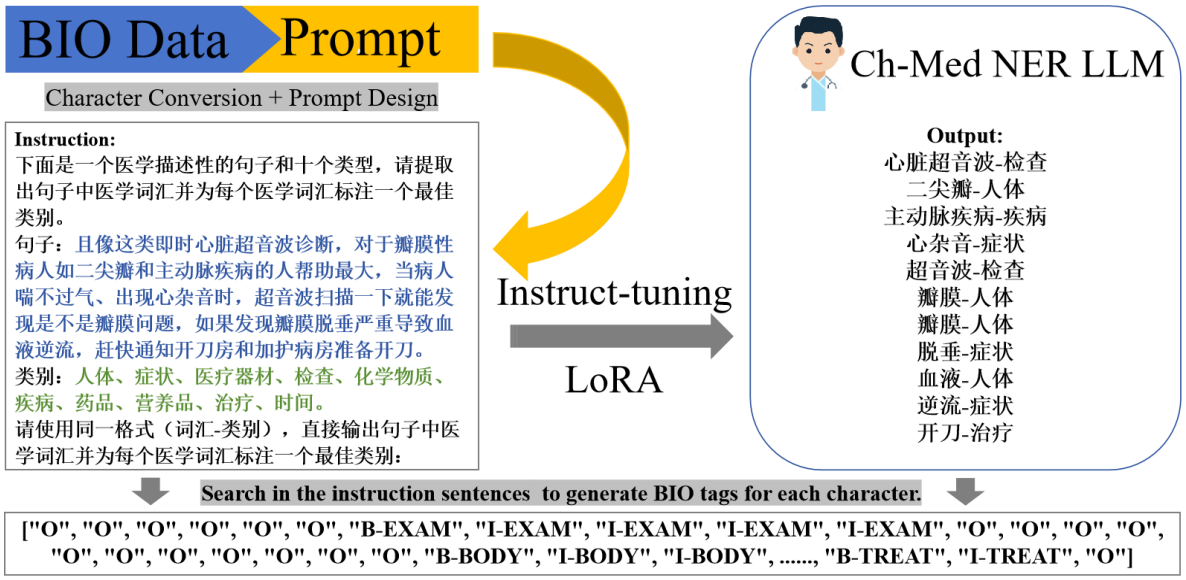

Figure 3: The framework of Ch-Med NER LLMs.

Team	F1 Score (%)			
	FT	SM	WA	Average
crowNER	**65.49**	69.54	**73.63**	**69.55**
YNU-HPCC	61.96	71.11	72.13	68.40
ISLab	62.52	**71.42**	71.19	68.38
SCU-MESCLab	62.51	71.33	70.57	68.14
YNU-ISE-ZXW	62.79	70.22	70.37	67.79
Official baseline BiLSTM+CRF (Word2vec)	60.99	67.16	67.91	65.35
Official baseline BiLSTM+CRF (BERT)	61.08	70.77	72.54	68.13
LingX (Run 1)	48.51	54.96	59.07	54.18
LingX (Run 2)	51.23	59.28	60.54	57.02
LingX (Run 3)	43.61	49.27	54.08	48.99

Table 4: Evaluation results on ROCLING-2023 benchmark test set.

4.2 Main Results

The benchmark test set released for ROCLING-2023 MultiNER-Health Task encompasses text samples from three distinct sources: Formal Texts (FT), Social Media (SM), and Wikipedia Articles (WA). Our team (LingX) made three submissions, all based on models fine-tuned on ChatGLM2-6B using the HealthNER prompt dataset. Run 1 represents the outcome of training for 5 epochs, Run 2 for 3 epochs, and Run 3 represents the results obtained by incorporating a knowledge graph distilled from the HealthNER prompt dataset on top of Run 2. Additionally, the official leaderboard and two models founded on BiLSTM-CRF serve as baseline models. The main evaluation results on the ROCLING-2023 benchmark test set are presented in Table 4. It is observed that while NER systems based on LLMs exhibit competence in extracting Chinese medical entities, they fall short in outperforming the recognition capabilities of traditional BiLSTM-CRF-based models. This discrepancy could stem from the inability of LLMs to precisely ascertain entity boundaries. This underscores the high level of intricacy in sequence labeling tasks, which might render LLMs less effective compared to baseline models. In addition, Run 2 achieved the best results among the three submissions. Specifically, training for 5 epochs in Run 1 may have led to overfitting in Ch-Med NER LLMs, resulting in a 2.84% decrease in the average F1 score compared to Run 2. Additionally, the Chinese medical knowledge graph constructed from the HealthNER prompt dataset may contain a substantial amount of noisy data for the ROCLING-2023 benchmark test set, leading to an 8.03% decrease in the average F1 score for Run 3 compared to Run 2.

4.3 Fine-tuning Results

We conduct a comparative analysis between the zero-shot performance of LLMs and

Ch-Med NER LLMs	Type	F1 Score (%)	
		ROCLING 2023	ROCLING 2022
ChatGPT-3.5	Zero-shot	39.32	50.83
ChatGLM2-6B	Zero-shot	30.68	44.69
BaiChuan-7B	Zero-shot	19.91	31.02
ChatGLM2-6B	Fine-tuned	57.02	65.23
BaiChuan-7B	Fine-tuned	**57.84**	**68.00**

Table 5: Comparison of LLMs' performance before and after fine-tuning.

the performance of fine-tuned Ch-Med NER LLMs for ROCLING-2023 and ROCLING-2022 benchmark test sets. As shown in Table 5, the results demonstrate that even ChatGPT-3.5, boasting a substantial parameter count of 175 billion, its performance remains moderate in Chinese medical NER tasks. It's worth noting that the performance of fine-tuned ChatGLM2-6B and BaiChuan-7B models exhibits significant breakthroughs when compared to their pre-finetuned performance. Specifically, on the ROCLING-2023 and ROCLING-2022 test sets, after fine-tuning, ChatGLM2-6B displayed enhancements of 26.34% and 20.54% respectively, while BaiChuan-7B showed improvements of 37.93% and 36.98% under similar conditions. These results provide evidence that our proposed approach demonstrates strong generalization capabilities.

5 Conclusion

This paper extends the application of LLMs methods to the domain of medical NER research. However, it has been observed that several common and Chinese-adapted LLMs do not perform satisfactorily in directly generating BIO labels for sentences. To bridge this gap, this study introduces a new research approach: the transformation of sequence labeling into entity extraction. We have devised specific Entity Extraction-Style prompts to stimulate the intelligent capturing of Chinese medical entities by LLMs. The overall assessment of prompts from the two different perspectives demonstrates the effectiveness and soundness of our prompt design pipeline. The evaluation results on the benchmark test sets of ROCLING-2023 and ROCLING-2022 indicate that although NER systems based on LLMs do not surpass conventional mainstream NER methods, the fine-tuned Ch-Med NER LLMs exhibit superior performance compared to the zero-shot performance of ChatGPT-3.5. Furthermore, we have also demonstrated that significant breakthroughs and strong generalization capabilities can be achieved for Ch-Med NER LLMs through instruct-tuning with specific prompts. In future, we intend to explore strategies that guide LLMs in BIO-Style fine-tuning, focusing on training data and prompts.

References

Anirudh Ajith, Chris Pan, Mengzhou Xia, Ameet Deshpande, and Karthik Narasimhan. 2023. Instructeval: Systematic evaluation of instruction selection methods. *arXiv preprint arXiv:2307.00259*.

Aitao Chen, Fuchun Peng, Roy Shan, and Gordon Sun. 2006. Chinese named entity recognition with conditional probabilistic models. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 173–176.

Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. 2016. Character-based lstm-crf with radical-level features for chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*, pages 239–250. Springer.

Guohong Fu and Kang-Kwong Luke. 2005. Chinese named entity recognition using lexicalized hmms. *ACM SIGKDD Explorations Newsletter*, 7(1):19–25.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

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*.

Bin Ji, Shasha Li, Jie Yu, Jun Ma, Jintao Tang, Qingbo Wu, Yusong Tan, Huijun Liu, and Yun Ji. 2020. Research on chinese medical named entity recognition based on collaborative cooperation of multiple neural network models. *Journal of biomedical informatics*, 104:103395.

Bin Ji, Rui Liu, Shasha Li, Jie Yu, Qingbo Wu, Yusong Tan, and Jiaju Wu. 2019. A hybrid approach for named entity recognition in chinese electronic medical record. *BMC medical informatics and decision making*, 19(2):149–158.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022a. Overview of the ROCLING 2022 shared task for Chinese healthcare named entity recognition. In *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)*, pages 363–368, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Lung-Hao Lee, Tzu-Mi Lin, and Chao-Yi Chen. 2023. Overview of the rocling 2023 shared task for chinese multi-genre named entity recognition in the healthcare domain. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*.

Lung-Hao Lee, Chien-Huan Lu, and Tzu-Mi Lin. 2022b. NCUEE-NLP at SemEval-2022 task 11: Chinese named entity recognition using the BERT-BiLSTM-CRF model. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1597–1602, Seattle, United States. Association for Computational Linguistics.

Lung-Hao Lee and Yi Lu. 2021. Multiple embeddings enhanced multi-graph neural networks for chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2801–2810.

Pan Liu, Yanming Guo, Fenglei Wang, and Guohui Li. 2022. Chinese named entity recognition: The state of the art. *Neurocomputing*, 473:37–53.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Caiyu Wang, Hong Wang, Hui Zhuang, Wei Li, Shu Han, Hui Zhang, and Luhe Zhuang. 2020. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree. *Journal of biomedical informatics*, 111:103583.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

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*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Fangzhao Wu, Junxin Liu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2019. Neural chinese named entity recognition via cnn-lstm-crf and joint training with word segmentation. In *The World Wide Web Conference*, pages 3342–3348.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Canwen Xu, Feiyang Wang, Jialong Han, and Chenliang Li. 2019. Exploiting multiple embeddings for chinese named entity recognition. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2269–2272.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Suxiang Zhang, Ying Qin, Wen-Juan Hou, and Xiaojie Wang. 2006. Word segmentation and named entity recognition for sighan bakeoff3. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 158–161.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.