Instruction Fine-Tuning of Large Language Models for Traditional Chinese Medicine

Juntao Li, Ling Luo(✉), Tengxiao Lv, Chao Liu, Jiewei Qi, Zhihao Yang, Jian Wang, and Hongfei Lin

School of Computer Science and Technology, Dalian University of Technology, Dalian, China
lingluo@dlut.edu.cn

Abstract. Traditional Chinese Medicine (TCM) is a holistic healthcare system that encompasses a rich body of medical theories, extensive clinical experiences, and a vast pharmacopoeia, which plays a significant role in medical practices. To promote the development of large language models (LLMs) for TCM, the CCKS 2024 challenge organized the TCMBench track, including TCM knowledge comprehension evaluation and TCM natural language inference tasks. In this paper, we present our instruction fine-tuning method for this track. We first constructed a rich training set by collecting existing Chinese medical exam datasets and TCM-related internet sources. We then designed various instructional prompts tailored to different tasks, enabling the model to fully exploit the knowledge acquired during fine-tuning to answer questions accurately and contextually. We also conducted extensive tests on different LLMs. Experimental results demonstrate the effectiveness and robustness of our method, achieving the first place in this challenge.

Keywords: Large Language Model · Traditional Chinese Medicine · Instruction Fine-Tuning

1 Introduction

Traditional Chinese Medicine (TCM) [1], with a history spanning thousands of years, remains integral to global healthcare, grounded in the theories of Yin-Yang balance [2], the Five Elements [3], and Qi. These foundational concepts emphasize holistic and syndrome differentiation approaches, forming the basis of TCM's diagnostic and therapeutic practices. Although TCM is gaining international recognition for its effectiveness in treating chronic and neurological diseases [4, 5], its complexity and reliance on practitioner experience present challenges for standardization and modern research [6].

The advent of Large Language Models (LLMs) like ChatGPT and GPT-4 has opened new possibilities for the preservation and innovation of TCM [7]. However, significant differences between TCM and Western medicine in theory and practice mean that traditional Western benchmarks are inadequate for evaluating LLMs in the TCM domain. To promote the development and evaluation of LLMs for TCM, Yue et al. [8] developed the TCM evaluation benchmark and organized the TCMBench track at the 2024 China

B. Xu et al. (Eds.): CCKS-IJCKG 2024, CCIS 2229, pp. 419–430, 2025.
https://doi.org/10.1007/978-981-96-1809-5_34

Conference on Knowledge Graph and Semantic Computing (CCKS 2024). This track is divided into two subtracks: efficient fine-tuning and non-fine-tuning. We participated in the efficient fine-tuning track, where participants are restricted to fine-tuning LLMs with parameters not exceeding 1% of the model's total.

The efficient fine-tuning track consists of two core subtasks: the TCM knowledge comprehension evaluation and the TCM natural language inference. The first task includes an evaluation dataset, TCM-ED, composed of real practice questions from the TCM Licensing Exam (TCMLE), which reflect the fundamental medical knowledge and reasoning required to obtain a TCM license in China. However, no training data is provided. The second task is a TCM-specific natural language inference task to determine if there is semantic consistency between two sentences, using the TMNLI dataset. Unlike the first task, the TMNLI dataset includes a substantial amount of training data. Therefore, we focus more on the first subtask, which presents a more significant challenge.

In this paper, we present our instruction fine-tuning method of LLMs for this track. Our main contributions are as follows:

- We collected data from multiple sources, including available Chinese medical exam datasets, TCMLE-related web resources, and TCM LLMs training data. To enhance the dataset, we used advanced LLMs to generate answer analysis, particularly for questions lacking medical interpretations. After standardizing the format, we ultimately obtained over 100,000 TCM-related data instances.
- We explored various instructional prompt methods for generating instruction-tuning data. Experimental results indicate that the model achieves better performance by analyzing both correct and incorrect options, compared to analyzing only the correct answers. Furthermore, providing question option information in multi-round dialogues significantly improves the accuracy on B1-type questions.
- We conducted comprehensive tests on different LLM bases. The results show that InternLM series outperforms other models.

Owing to the above contributions, our method achieves the first place in the efficient fine-tuning track of the TCMBench Challenge at CCKS 2024, with the accuracy scores of 89.01 and 97.65% in the TCM-ED and TMNLI tasks, respectively.

2 Related Work

Recently, LLMs (such as GPT-4 [9] and LLaMA [10]) have shown promising results in various natural language processing (NLP) tasks and have received widespread attention around the world. However, theseLLMs often perform sub-optimal and struggle to follow instructions when applied to specific domains like biomedicine. Training LLMs from scratch typically requires much resource. To address these challenges, Instruction Fine-tuning (IFT) methods [11–13] have been proposed to enhance the performance of LLMs, particularly in the medical field. The primary objective of IFT is to improve the model's ability to follow various human instructions and execute specific tasksIFT achieves this goal by utilizing specially constructed training datasets that typically contain instruction-input-output triplets. These approaches enable the development of LLMs that can more

accurately interpret and respond to medical queries, follow complex medical instructions, and perform specialized medical tasks, thereby potentially improving the application of LLMs in healthcare and medical research.

In the field of TCM question-answering (QA), the introduction of LLMs has brought new vitality to the integration of traditional medicine with modern technology. Models like ShenNong-TCM-LLM [14] and CMLM-ZhongJing [15] represent significant advancements in TCM-focused LLMs. ShenNong-TCM-LLM is built upon a TCM knowledge graph and utilizes a vast amount of TCM-related instructional data generated with the assistance of ChatGPT and similar models. Subsequently, it undergoes instruction fine-tuning. In clinical TCM diagnosis and treatment, ShenNong-TCM-LLM show exceptional capabilities in prescription recommendation. CMLM-ZhongJing employs specialized tables and specific prompt templates to achieve precise simulation and reasoning of TCM prescription data and complex diagnostic logic. It can also generate detailed diagnostic and treatment plans. This provides comprehensive intelligent assistance for TCM clinical diagnosis and treatment, from diagnosis to therapy, thus further advancing the modernization of TCM practice.

3 Method

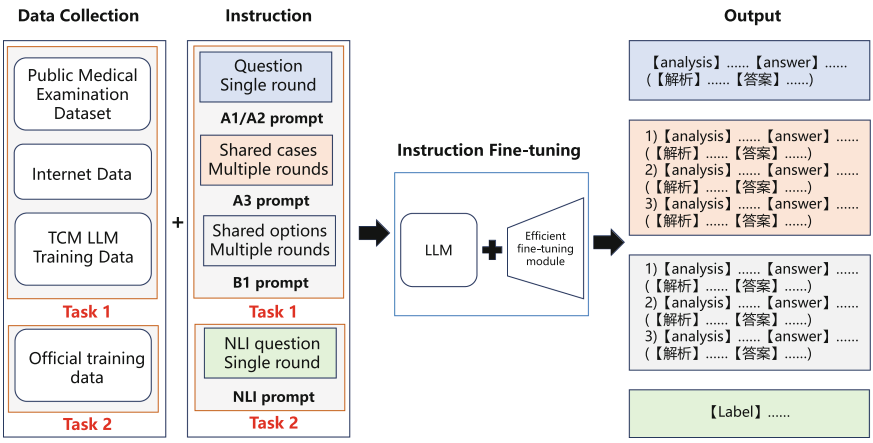

Fig. 1. Overview of our instruction fine-tuning method

To enhance the model's performance on TCMBench, we focused on two main aspects: data quality and instructional prompts design. An overview of our fine-tuning method for TCMBench is shown in Fig. 1. We first collected relevant data about TCM licensing exams from multiple sources for the TCM-ED task. To improve data quality, we employed GLM-4 [16] to generate medical interpretations for questions that lacked explanations. Next, we developed and tested specialized instructional prompts for different question types. We continuously optimized and adjusted our prompting strategies based on the structural characteristics of each question type. Through these methods,

we aim to optimize training data and design targeted prompting strategies to enhance the model's overall performance on TCMBench.

3.1 Training Data Collection

Since official TCMBench only provides training data for the TMNLI task, data for the TCM-ED task must be collected and processed extensively. This section details our data collection and processing methods. The distribution of the final dataset across different question types is presented in Table 1. It is important to note that while some data for A1 and A2 question types lack explanations, the A3 and B1 question types include complete explanations.

Table 1. Information of dataset for fine-tuning

Type	Task1				Task2	Other
	A1/A2 (analysis)	A1/A2 (no analysis)	A3	B1	NLI	QA
Official training data	–	–	–	–	25,581	–
CMExam	44,199	7,912	–	–	–	–
CMB	–	31,561	–	–	–	–
Internet Data	7,940	–	2,779	4,641	–	–
TCM LLM Training Data	–	–	–	–	–	10,000
Total	52,139	39,473	2,779	4,641	25,581	10,000

To enhance the model's performance on Task1 (i.e., the TCM-ED task), we aggregated comprehensive sets of related data from three sources: public medical examination datasets, internet data and TCM LLM training data.

(1) **Public Medical Examination Dataset**

CMExam [17]. The questions are sourced from China's National Medical Licensing Examination. CMExam consists of over 60,000 multiple-choice questions designed for standardized and objective assessment, with 85.24% of the questions including medical analysises.

CMB [18]. This is a comprehensive, multi-level Chinese medical evaluation dataset. It contains 280,839 multiple-choice questions and 74 complex case consultation questions, covering all clinical medical specialties and different professional levels. However, CMB's answers do not include any medical analysis. It aims to comprehensively assess models' medical knowledge and clinical consultation abilities. We selected TCM-related single-choice questions, cleaned the data, and used them to improve the model's reasoning accuracy during fine-tuning.

(2) **Internet Data**

The datasets mentioned above mainly cover the A1/A2 type questions but lack other question types present in TCMBench. Therefore, we collected a large number of TCM licensing exam simulation questions and practice problems from the internet, such as Weipu[1] and Huanqiuwangxiao[2]. These questions are carefully screened and organized, primarily covering A3 and B1 type questions, which are relatively scarce in public datasets. We employed OCR technology to extract the question content and used regular expressions to standardize the data, ensuring all data conforms to a unified format. For some questions lacking standard analyses, we employed an advanced LLM, GLM-4, to generate relevant medical analyses. This data collection and processing not only addressed the scarcity of A3 and B1 type questions but also provided more diverse training materials, enabling the model to perform more stably and accurately on these scarce question types.

(3) **Traditional Chinese Medicine LLM Training Data**

ShenNong_TCM_Dataset[3]. The dataset serves as the training corpus for the ShenNong-TCM-LLM, aiming to enhance the LLM's knowledge in TCM and its ability to respond to medical inquiries. However, the dataset also contains some queries that cannot be answered with certainty, as well as questions unrelated to TCM.. Through a rigorous keyword matching and filtering strategy, we eliminated question-answer pairs that lack direct, clear answers or are not directly relevant to TCM topics, ensuring the dataset's purity and high relevance. Ultimately, we randomly extracted 10,000 dialogue entries closely from the filtered dataset, s. These entries serve as a fine-tuning training corpus, supporting the model's specialized optimization.

For Task2 (i.e., the TMNLI task), we used the training dataset provided by CCKS2024 TCMBench. In this dataset, each piece of data includes a premise, hypothesis, and a label indicating the entailment relationship. The data is constructed using questions, answers, and standard explanations from the TCM licensing exam. To eliminate bias caused by significant length differences between question-answer generated sentences and standard explanations, the premise and hypothesis were swapped for half of the data pairs, resulting in 25,581 instances of training data.

Through the collection and processing of these data, we have constructed a comprehensive and optimized training dataset, providing strong support for the model's excellent performance in the TCMBench task. This process has not only enhanced the model's ability to handle different types of questions but also promoted the digital inheritance and modernization research of TCM knowledge.

3.2 Instruction-Tuning Data Construction

To enable the LLMs to understand TCM task instructions for multitasking, we constructed and optimized instructional prompts for fine-tuning, covering the tasks described in the previous section. First, we utilized the Chain of Thought (CoT) prompt templates

[1] https://oldvers.cqvip.com/view/professional/index.aspx.

[2] https://www.hqwx.com/zyzyys-kaoshi/moniti/.

[3] https://huggingface.co/datasets/michaelwzhu/ShenNong_TCM_Dataset.

provided by the CCKS-2024 TCMBench organizers, which included A1/A2, A3, and B1 question types. For the TMNLI task that do not have predefined prompts, we referred to prompt templates from other question types and created instructional prompts that matched the characteristics of the TMNLI task. To maximize the use of A1/A2 question type data that lack explanations, we designed new instruction templates modeled after the prompt templates for questions with explanations. These customized instructional prompts not only provide the model with clearer problem-solving guidance but also enable it to demonstrate strong understanding and reasoning abilities even when faced with unannotated data. For ShenNong dialogue data, we developed specific instruction templates. These templates help the model distinguish between different tasks and significantly enhance its performance on TCM tasks, particularly by aiding the model in learning a large amount of TCM knowledge.

Notably, due to the unique structure of B1 question types—where options are extracted as shared content—we designed two different forms of instructional prompts. As shown in Fig. 2, Prompt1 places the shared options only before the first sub-question, with subsequent questions providing only the question content. This format requires the model to have strong memory capabilities to effectively utilize the previously provided information in subsequent answers. In contrast, Prompt2 places the options after each sub-question, making the structure of each sub-question more similar to A1/A2 types. This design leverages the model's learned TCM knowledge from other question types, thereby improving performance on B1 questions.

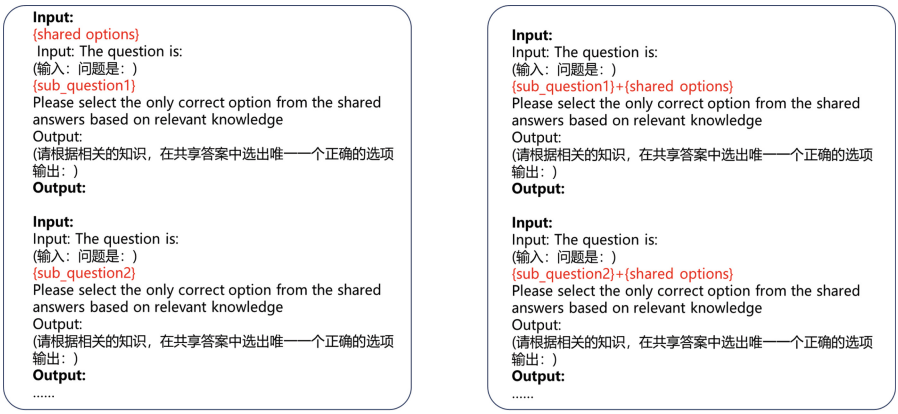

Fig. 2. Two methods of prompts for B1 type questions

Furthermore, we explored the effect of two different types of answer analyses for the models. The first is the original Long Analysis provided by the CMExam dataset. This analysis provides a comprehensive explanation of the answers, including explanations for the correct options as well as the reasons for all incorrect options. We also tested the second type, i.e., Shot Analysis, where the data was cleaned and processed to focus on key points, making the explanations shorter and more consistent with the length

and format of standard explanations in the TCMBench validation set.. An example of different analyses is shown in Fig. 3.

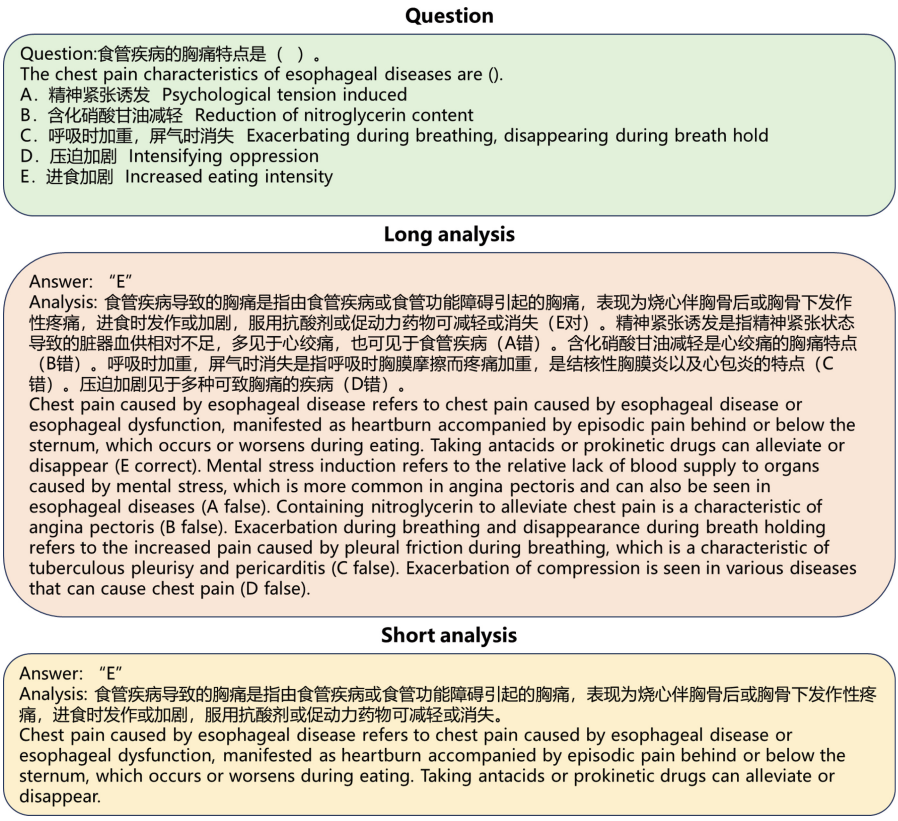

Fig. 3. An Example of different analyses for TCM questions

3.3 Model Fine-Tuning

After completing the preliminary data preparation, we constructed various training sets by combining different instructional prompts and dataset. Participants are required to choose one of the following open-source models in the challenge: Qwen series[4], InternLM series[5], ChatGLM3-6B [16] and ShenNong-TCM-LLM [14]. During the model fine-tuning process, we employed the Firefly[6] fine-tuning framework in conjunction with QLoRA [19] technology for efficient fine-tuning. This strategy allowed us to develop an efficient and accurate model fine-tuning process while ensuring that the

[4] https://huggingface.co/qwen.

[5] https://huggingface.co/internlm.

[6] https://github.com/yangjianxin1/Firefly.

model maintains a high level of understanding and reasoning ability with complex TCM tasks.

4 Experiments

4.1 Datasets and Evaluation Metrics

The official final test set includes TCM-ED and the TMNLI test sets. In the TCM-ED test set, there are 1,598 combined samples of types A1 and A2, and 198 samples of type A3, which include 642 sub-questions. For type B1, there are 1,481 test samples with 3,231 sub-questions. The TCM-NLI test set consists of 3,916 samples.

For the two evaluation tasks, CCKS2024 TCMBench uses different evaluation metrics as follows: For the TCM-ED task, the accuracy metric is first used to compare the correct options with the model's responses. The other evaluation metrics are for generating question analyses, using Rouge1 [20], RougeL [20], BertScore [21], and SARI [22] to comprehensively evaluate the quality of model-generated analyses. For the TMNLI task, the accuracy is used as the evaluation metric. The model's ability to judge the semantic similarity between two sentences is assessed by comparing the labels generated by the model with the correct labels. The average of all scores is used as the final score.

4.2 Implementation Details

We used QLoRA to efficiently fine-tune the base model. During the training process, we set the learning rate to 0.0001, the maximum sequence length to 1024 and the number of training epochs to 4. We adopted a constant learning rate strategy with warm-up, setting the warm-up steps to 100, and selected paged_adamw_32bit as the optimizer. In the QLoRA parameter configuration, we set the rank to 32, the alpha value to 32 and the dropout rate for LoRA layers to 0.05, without fine-tuning the bias terms.

During the efficient fine-tuning process, we used three Nvidia A5000 GPUs for training and two separate GPUs for inference testing. During inference, we set the temperature coefficient to 0.35, top_p to 0.9, and the maximum number of generated tokens to 500.

4.3 Performance of Different LLMs Without Fine-Tuning

To determine the base model for fine-tuning, we first tested different chat models using zero-shot prompt without fine-tuning. The accuracies of various models on different tasks are shown in Table 2. Compared to other models, ChatGLM3-6B shows lower accuracy across all tasks. For the Qwen series, we found that Qwen2-7B-Instruct performed better than Qwen1.5-14B-Chat on Task1 (the TCM-ED task), but its accuracy on Task2 (the TMNLI task) was significantly lower. This was because when Qwen2-7B-Instruct was instructed to judge whether there was an entailment relationship between the hypothesis and premise sentences, with label values of 0 and 2 (where 0 indicates an entailment relationship between the premise and hypothesis, and 2 indicates no relation or contradiction), it outputted a large number of cases with label value 1.

Table 2. Accuracy and final score of each model without fine-tuning on the final test set

Model	Task1	Task2	A1/A2	A3	B1	Score
ChatGLM3-6B	40.09	37.22	40.68	51.02	37.77	33.15
Qwen1.5-14B-Chat	70.48	73.77	74.84	75.17	67.49	48.61
Qwen2-7B-Instruct	71.66	56.35	**77.97**	75.17	67.93	47.32
InternLM2-20B-Chat	67.60	**84.85**	66.62	73.13	67.05	50.96
InternLM2_5-7B-Chat	**74.08**	76.45	**77.97**	**77.89**	**71.48**	**52.28**

For the InternLM series, we found that InternLM2_5-7B-Chat outperformed other models overall. Although its accuracy on Task2 is lower than that of InternLM2-20B-Chat from the same series, it achieves higher accuracy on other tasks compared to other models. Based on these findings, we conducted fine-tuning experiments using a small amount of data on Qwen2-7B-Instruct, InternLM2_5-7B-Chat, and InternLM2_5-7B. Ultimately, we decided to use InternLM2_5-7B as the base model for subsequent experiments.

4.4 Main Results

To evaluate the performance of different datasets and instructional prompts, we submitted multiple runs with various settings. After efficient fine-tuning, the model's accuracy on different tasks is shown in Table 3. Our best submission (i.e., Run5) achieves the highest final score, securing the first place in the efficient fine-tuning track of the challenge. The results of the top 2 and 3 teams are also shown for comparison. We further analyzed the results of different runs, and our main findings are as follows.

(1) **The TCM training data collected from various sources contributes to the model's performance in TCM knowledge comprehension.** The results from Run1, 4 and 5 show that incorporating internet data into the public medical examination dataset notably improves the accuracy in A1/A2 and A3 question types. This suggests that the internet data effectively help the model acquire the medical knowledge necessary for completing Task1. Furthermore, the results of fine-tuning with additional TCM LLM training data show an improvement in Task1 accuracy compared to Run4, suggesting that the dialogue data help the model supplement some TCM-related medical knowledge. By comparing the effects of different data sources, we found that the internet data contributed the most to improving the model's accuracy on Task1.

(2) **Providing question option information in multi-round dialogues significantly improves the model's performance on B1-type questions.** Comparing Run2 and Run3, we observed that prompt optimization significantly improves the model's accuracy on specific question types. The accuracy for type B1 increases from 83.26 to 90.03. This is because in Run2, we only provided shared answer options for the first sub-question, while in Run3, we added shared answer options after each sub-question's prompt, making the prompt for each B1 sub-question more similar to

Table 3. Accuracy and final score of different methods on the final test set

Method	Setting	Task1	Task2	A1/A2	A3	B1	Score
sf_cloud (Top2)	–	84.90	**98.42**	86.55	85.51	83.97	59.39
ZZUNLP (Top3)	–	85.98	96.99	**88.30**	**88.47**	84.34	58.97
Run1	Data: Exam; Prompt:1; Analysis: Long	77.98	97.20	83.44	81.29	74.71	57.67
Run2	Data: Exam + Internet; Prompt:1; Analysis: Short	85.19	97.27	88.05	87.85	83.26	58.53
Run3	Data: Exam + Internet; Prompt:2; Analysis: Short	88.83	97.14	87.30	86.60	90.03	59.15
Run4	Data: Exam + Internet; Prompt:2; Analysis: Long	88.52	97.45	86.36	87.69	89.76	59.42
Run5 (Top1)	Data: Exam + Internet + LLM; Prompt:2; Analysis: Long	**89.01**	97.65	86.98	85.98	**90.62**	**59.75**

the A1/A2 question types. This improvement allows the model to better leverage knowledge and experience gained from A1/A2 type training data during reasoning.

(3) **The models trained with long analysises achieve higher analysis scores than those trained with short analysises.** As shown in Table 4, a comparison between Run3 and Run4 reveals that using longer explanations resulted in slightly lower RougeL and BertScore scores but notably improve Rouge1 and SARI scores. Overall, the average score when using longer analysises is slightly higher than that of using shorter analysises. We also found that Rouge1 and SARI were positively correlated, as were RougeL and BertScore, while there was a negative correlation between these two groups of metrics. Additionally, the incorporation of these data sources effectively improved the model's average scores in generating analyses.

Table 4. The analysis scores of the different methods on the final test set

Method	Setting	Rouge1	RougeL	BertScore	SARI	*AVE*
sf_cloud (Top2)	–	42.35	28.54	**73.54**	28.60	**43.26**
ZZUNLP (Top3)	–	**45.65**	23.68	71.14	30.40	42.72
Run3	Data: Exam + Internet; Prompt:2; Analysis: Short	39.15	**28.99**	73.33	27.10	42.14
Run4	Data: Exam + Internet; Prompt:2; Analysis: Long	42.47	26.33	72.35	29.42	42.64
Run5 (Top1)	Data: Exam + Internet + LLM; Prompt:2; Analysis: Long	43.48	25.78	72.05	**30.54**	42.96

5 Conclusion

In this paper, we present our instruction fine-tuning method for LLMs in the TCM-Bench track. We successfully constructed training data using web scraping techniques and advanced GLM-4, while fully utilizing open-source TCM-related datasets. Experimental results show that incorporating long analysis data and additional TCM dialogue data can enhance models' reasoning performance during the fine-tuning process. Our method achieves the state-of-the-art performances on TCMBench test sets, demonstrating its effectiveness for TCM. In future work, we plan to further explore additional TCM capabilities of LLMs, such as assisting in disease diagnosis and treatment.

Acknowledgments. This research was supported by the CIPSC-SMP-Zhipu.AI Large Model Cross-Disciplinary Fund (NO. ZPCG2024010204), the National Natural Science Foundation of China (No. 62302076), and the Fundamental Research Funds for the Central Universities (No. DUT23RC (3)014).

References

1. Nestler, G.: Traditional chinese medicine. Med. Clin. **86**(1), 63–73 (2002)
2. Li, P.P.: The unique value of yin-yang balancing: a critical response. Manag. Organ. Rev. **10**(2), 321–332 (2014)
3. Pachuta, D.M.: Chinese medicine: the law of five elements. India Int. Centre Q. **18**(2/3), 41–68 (1991)
4. Jiang, M., Zhang, C., Cao, H., Chan, K., Lu, A.: The role of Chinese medicine in the treatment of chronic diseases in china. Planta Med. **77**(09), 873–881 (2011)
5. Ren, Z.L., Zuo, P.P.: Neural regeneration: role of traditional Chinese medicine in neurological diseases treatment. J. Pharmacol. Sci. **120**(3), 139–145 (2012)
6. Fung, F.Y., Linn, Y.C.: Developing traditional chinese medicine in the era of evidence-based medicine: current evidences and challenges. Evid.-Based Complement. Altern. Med. **2015**(1), 425037 (2015)

7. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nat. Med. **29**(8), 1930–1940 (2023)
8. Yue, W., et al.: Tcmbench: a comprehensive benchmark for evaluating large language models in traditional chinese medicine. arXiv preprint arXiv:2406.01126 (2024)
9. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
10. Touvron, H., et al.: Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
11. Luo, L., et al.: Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. J. Am. Med. Inf. Assoc. ocae037 (2024)
12. Tran, H., Yang, Z., Yao, Z., Yu, H.: Bioinstruct: instruction tuning of large language models for biomedical natural language processing. J. Am. Med. Inf. Assoc. ocae122 (2024)
13. Singhal, K., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
14. Wei Zhu, W., Wang, X.: Shennong-tcm: A traditional chinese medicine large language model (2023)
15. Yang, S., et al.: Zhongjing: Enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 19368–19376 (2024)
16. Glm, T., et al.: Chatglm: a family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793 (2024)
17. Liu, J., et al.: Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
18. Wang, X., et al.: Cmb: a comprehensive medical benchmark in Chinese. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6184–6205 (2024)
19. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: efficient finetuning of quantized llms. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
20. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
21. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
22. Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C.: Optimizing statistical machine translation for text simplification. Trans. Assoc. Comput. Linguist. **4**, 401–415 (2016)