TCMBench: A Comprehensive Benchmark for Evaluating Large Language Models in Traditional Chinese Medicine

Wenjing Yue^{1,2}, Xiaoling Wang¹(⋈)Wei Zhu¹, Ming Guan¹, Huanran Zheng¹, Pengfei Wang¹, Changzhi Sun³, and Xin Ma⁴

- Shanghai Institute of AI for Education, East China Normal University, Shanghai, China
 - ² School of Computer Science and Technology , East China Normal University, Shanghai, China

School of Pharmacy, East China Normal University, Shanghai, China Shanghai Skin Disease Hospital, School of Medicine, Tongji University, Shanghai, China {wjyue, wzhu, hrzheng, pfwang}@stu.ecnu.edu.cn xlwang@stu.ecnu.edu.cn czsun.cs@gmail.com nicole.ma@me.com

Abstract. Large language models (LLMs) have performed remarkably well in various natural language processing tasks by benchmarking, including in the Western medical domain. However, the professional evaluation benchmarks for LLMs have yet to be covered in the traditional Chinese medicine(TCM) domain, which has a profound history and vast influence. To address this research gap, we introduce TCM-Bench, an comprehensive benchmark for evaluating LLM performance in TCM. It comprises the TCM-ED dataset, consisting of 5,473 questions sourced from the TCM Licensing Exam (TCMLE), including 1,300 questions with authoritative analysis. It covers the core components of TCMLE, including TCM basis and clinical practice. To evaluate LLMs beyond accuracy of question answering, we propose TCMScore, a metric tailored for evaluating the quality of answers generated by LLMs for TCM related questions. It comprehensively considers the consistency of TCM semantics and knowledge. After conducting comprehensive experimental analyses from diverse perspectives, we can obtain the following findings: (1) The unsatisfactory performance of LLMs on this benchmark underscores their significant room for improvement in TCM. (2) Introducing domain knowledge can enhance LLMs' performance. However, for indomain models like ZhongJing-TCM, the quality of generated analysis text has decreased, and we hypothesize that their fine-tuning process affects the basic LLM capabilities. (3) Traditional metrics for text generation quality like Rouge and BertScore are susceptible to text length and surface semantic ambiguity, while domain-specific metrics such as TCMScore can further supplement and explain their evaluation results. These findings highlight the capabilities and limitations of LLMs in the TCM and aim to provide a more profound assistance to medical research.

Keywords: Benchmark \cdot Traditional Chinese Medicine \cdot Large Language Model \cdot Healthcare and Medicine.



Fig. 1: The difference between TCM and Western Medicine.

1 Introduction

Recently, Large language models (LLMs) have been demonstrated to lead performance in enhancing the accuracy of natural language understanding and text generation quality. Emerging studies have designed various LLMs like ChatMed³, HuaTuo [13], and ZhongJing-TCM ⁴, highlights the growing demand for LLMs in the medical domain. Therefore, the standardized medical benchmark is essential for effectively developing and applying LLMs in medicine, providing reliable and authoritative assessments.

Popular medical benchmarks mainly focus on Western medicine, such as MedMCQA [11] and MultiMedQA [12]. Among this, MultiMedQA combines new online medical queries to alleviate data contamination issues in evaluations based on publicly easily accessible data sources [2]. Nonetheless, there are significant differences among medical systems, including variations in clinical standards, procedures, and languages [2], particularly evident in the differences between Traditional Chinese Medicine (TCM) and Western Medicine. TCM has a long and rich history, making profound contributions to healthcare [3]. Unlike Western evidence-based medicine, TCM emphasizes the clinical experience of physicians [20]. Moreover, significant differences exist between their regarding theoretical foundations, diagnostic methods, treatment modalities, preventive concepts, and holistic views, as illustrated in Figure 1. These differences highlight the unique aspects of TCM diagnosis, treatment, and knowledge in the medical field. While some benchmarking has been considered to evaluate the Chinese medical domain [2,9], they also primarily evaluate modern Chinese medicine knowledge based on Western medicine principles. Therefore, directly applying existing Western medical benchmarks to assess TCM may not comprehensively evaluate the potential and practical utility of LLMs in this domain. Recently, the TCM LLM, such as ZhongJing-TCM, relies solely on physicians' subjective evaluation model performance, which consumes valuable time and leads to low efficiency. It highlights the urgent need for a standardized benchmark in TCM to provide objective and reliable evaluations of LLMs' performance.

To address these gaps and accommodate the unique characteristics of TCM knowledge, we introduce a new comprehensive benchmark, **TCMBench**, to supplement prior medical benchmarks. It is sourced from the TCM Licensing Exam (TCMLE), which is specially tailored for the TCM domain. To prevent data con-

³ https://github.com/michael-wzhu/ChatMed

⁴ https://github.com/pariskang/CMLM-ZhongJing

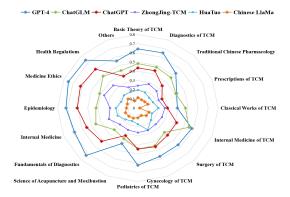


Fig. 2: The performance of different LLMs on different branches of TCM-Bench.

tamination, we construct a large-scale TCM evaluation dataset **TCM-ED** using actual TCMLE practice questions. It comprises 5,473 question-answer(Q&A) pairs, with 1,300 data pairs with standard analysis, ensuring the reliability of the data quality. We manually select questions from the original dataset to confirm comprehensive coverage of all TCM branches, ensuring a broad scope within the topic of TCM-ED. The accuracy of LLMs in all TCM branches is shown in Figure2. Significantly, due to the unique terminology of TCM, exemplified by phrases like 'catching a cold from the chilly wind' and 'external attack of wind cold,' which convey the same meaning despite having low word matching. Thus, using traditional text generation metrics based solely on word matching [9,2] or Chinese semantic representation models to evaluate the semantic consistency between two texts in TCM may not be appropriate. Based on these findings, we introduce an automatic metric called TCMScore to evaluate the consistency of TCM semantics and knowledge. It combines the matching of TCM terms and the semantic consistency between the generated and standard analyses.

We conduct extensive experiments, combining various metrics, to provide a detailed analysis from different perspectives on evaluating LLMs' ability to understand, analyze, and apply knowledge in TCM. The main findings of this benchmark are as follows:

- The current performance of LLMs on this benchmark is unsatisfactory, indicating considerable room for improving their application in TCM. But general LLMs with hundreds of billions of parameters demonstrate potential for better application in TCM.
- Figure 2 shows that general LLMs without specialized tuning are biased toward Western medicine. However, infusing professional TCM knowledge and related linguistic-cultural corpus can significantly improve LLMs' comprehension of context in TCM.
- From the express quality and human evaluation, fine-tuning LLMs with domain knowledge in TCM weakens their fundamental abilities in logical rea-

- soning, knowledge analysis, and semantic expression. Thus, preserving these core abilities during pre-training is crucial.
- The expression quality metrics that rely on word matching or semantic similarity are easily affected by factors such as text length and surface semantic ambiguity. The TCMScore introduced in our work effectively addresses this limitation and can better supplement and explain the performance of LLMs in TCM semantic and knowledge consistency under the above metrics.

In conclusion, we propose an comprehensive benchmark aligned with the requirements of TCM, aiming to showcase LLMs' capabilities and limitations in the TCM domain and improve further developments in medical research.

2 Related Work

As LLMs advance rapidly, benchmarking is crucial for driving progress in natural language processing (NLP), particularly in professional domains like medicine. Various medical Q&A benchmarks tailored to different regions and medical systems have been instrumental in evaluating the efficacy of LLMs. For example, Kuang et al. [7] utilize the United States Medical Licensing Examination(USMLE) questions to evaluate ChatGPT, while MedMCQA [11] develops a benchmark using Indian medical data. CBLUE [19] and PromptCBLUE [22] focus on evaluation based on Chinese bio-medical information. Additionally, MultiMedQA [12] combines six existing medical benchmark datasets, like MedQA [5] and MedMCQA, to mitigate data contamination through new online data sets. It also conducts a human evaluation through instruction fine-tuning based on the limited number of doctor-labeled samples. However, existing benchmarks focus on Western medical systems and lack TCM content. TCM, recognized by the World Health Organization as an effective complementary and alternative medicine system, differs significantly from Western medicine in its theoretical structure, diagnosis, and treatment standards [20]. Therefore, Western medicinebased benchmarks cannot adequately evaluate the performance of LLMs in TCM. Although benchmarks like CMExam [9] and MedBench [2] have been proposed for CNMLE, they also focus on modern Chinese medicine questions based on Western medicine principles. Moreover, they only offer rough statistics on TCM and briefly use metrics like Rouge for analysis. However, due to the unique term expression characteristics of TCM, these benchmarks cannot fully evaluate LLMs' performance in answering TCM questions. There remains a lack of comprehensive benchmarks in the academic community for evaluating LLMs in the TCM domain. In the evaluation of ZhongJing-TCM, only subjective evaluation by doctors was used to evaluate model performance. However, manual evaluation is time-consuming and labor-intensive, making it difficult to achieve large-scale applications. Overall, it is urgent to develop objective and systematic LLM evaluation benchmarks that adapt to the characteristics of TCM to fill the gap in evaluation standards in this domain.

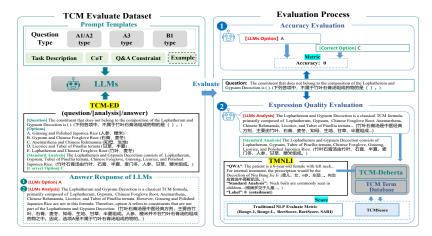


Fig. 3: The overview of TCM-Bench. It consists of two parts: (1) On the left is the construction process of the TCM-ED dataset. (2) On the right is the evaluation process of TCM-Bench. The bottom section showcases the TMNLI dataset and the TCM-Deberta model, as well as the TCMScore metric.

3 The Proposed Benchmark

3.1 Overview

We propose an comprehensive benchmark, TCMBench, for evaluating the effectiveness of LLMs in TCM, as depicted in Figure 3. It includes an evaluation dataset, TCM-ED, comprising 5,473 actual practice questions from TCMLE, which reflect the fundamental medical knowledge and reasoning logic required to obtain a TCM license in China. To create an automatic evaluation metric aligned with expert cognition, TCMScore, we first collect 9,788 recent real questions with analysis from TCMLE to build the first TCM natural language inference (NLI) dataset, TMNLI. We then introduce TCM-Deberta, a more stable semantic inference model, to effectively evaluate TCM semantic consistency. Additionally, we incorporate a task to evaluate TCM terminology matching in calculating TCMScore to reveal the knowledge consistency. In the end, we utilize multiple metrics to evaluate the ability and quality of LLMs to express TCM knowledge.

3.2 Construction and Statistics

TCM-ED The TCMLE assesses whether applicants possess the necessary professional knowledge and skills to practice as TCM physicians. Therefore, we collect 5,473 representative practice questions. Among them, the data we collect does not contain personal information but focuses on selecting data instances that can fully reflect and represent theoretical knowledge and practical skills

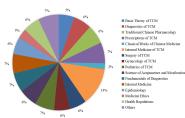


Fig. 4: Branches of TCM in TCM-ED.

Table 1: The statistical information of the TCM-ED (upper) and TMNLI(lower).

	Question Type	$\#\mathbf{Qustions}$	# Sub-Question # All		
TCM-ED	A1/A2	1,600	1,600		
	A3	198	642	5473	
	B1	1,481	3,231		
	DataSet	#Entailment	#Contradiction	#All	
TMNILI	Train	#Entailment 7,830	1 **	# All 25,581	
TMNLI			1 **		

in TCM. The multiple-choice questions in TCMLE are depicted on the left of Figure 3, including three types:

- The single-sentence best-choice questions(A1) and the case summary best-choice questions(A2) type: It consists of a question stem and five options with correct one, as shown in Figure 1 of the appendix file.
- The best choice questions for case group(A3) type: The stem presents a patient-centered case, followed by multiple sub-questions, each offering five options with one correct answer. It primarily centers on clinical applications, as shown in Figure 2 of the appendix file.
- The standard compatibility questions(B1) type: Multiple sub-questions share the same five options, where each option may be chosen zero, one, or multiple times. There is one correct answer among the five options for each sub-question, as shown in Figure 3 of the appendix file.

Specifically, we manually screen the original practice question bank from TCMLE with the advice of experts to ensure that TCM-ED encompasses all question types and branches of TCM found in TCMLE. First, we cleaned the original data in PDF format. Then, we extract the questions, options, correct answers, and standard analysis using rule templates. Then, we convert the information into a structured JSON format. Subsequently, we randomly select 100 questions from the original question bank in each specific medical branch under each question type based on expert guidance. If a particular branch has fewer than 100 questions, all are selected. The detailed statistics of the TCM-ED are provided in the upper of Table 1, indicating that each TCM branch in the A1/A2 type of questions comprises a complete 100 questions. Furthermore, Figure 4 illustrates that the distribution of all TCM-ED questions across different branches is relatively balanced, ensuring that evaluation results are not biased due to the branch distribution. This ensures fairness and comprehensiveness in evaluation.

TMNLI Due to the uniqueness of terminology in TCM, simple word matching or Chinese semantic similarity calculations cannot accurately measure semantic consistency in the TCM domain. In general, the NLI metrics are typically utilized to assess the fidelity of summaries. Previously work employ an entailment classifier trained on the MultiNLI(MNLI) dataset [15] to determine if summaries

Model Open Source Parameters Domain GPT-4 [10] 175B +General ChatGPT [1] 175BGeneral X ChatGLM [18] 130B General Chinese LlaMa [4] 7BGeneral HuaTuo[13] 7BChinese Medicine ZhongJing-TCM⁵ 7BTCM

Table 2: The statistical information of LLMs.

are consistent with the context. However, since the MNLI dataset is in English, significant differences exist between it and TCM terminology. Therefore, we construct a TCM-specific NLI dataset dataset, TMNLI. We select 9,788 recent examination questions with standard analysis from TCMLE, covering three question types and all TCM branches depicted in Figure 4. Following the setting of the MNLI dataset, TMNLI consists of three parts: premise, hypothesis, and label. We leverage rule templates to combine the question and its correct answer into a claim called QWA, which serves as the premise. We then consider the standard parsing as a hypothesis and set the relation label between the two as entailment. Additionally, we utilize the BM25 algorithm to rank other analyses based on the similarity with QWA, randomly selecting up to three analyses from the top 20 to 100 rankings as hypotheses, labeled as contradictions. We aim to increase the difficulty in identifying the relationship between the premise and hypothesis. Consequently, we generated 29,497 NLI data. Inspired by previous work [6], we consider that if there is stable semantic consistency between two sentences, it should be possible to derive the hypothesis from the premise and vice versa. Given the significant difference in length between QWA and the standard analyses in TMNLI, we permute the premises and hypotheses of half of the data pairs in TMNLI to eliminate bias caused by length discrepancies. The statistical information is detailed in the lower of Table 1.

4 Evaluation

4.1 Models

To assess the medical abilities in the TCM domain, we utilize TCMBench to evaluate various LLMs across general and medical domains. Specifically, we leverage the LLMs on an over 100 billion scale, such as the commercial (closed-source) GPT-4 and ChatGPT, as well as the open-source ChatGLM that supports Chinese. Additionally, we evaluate the general Chinese model Chinese-LlaMA and the Chinese medical-specific model HuaTuo that focuses on Western medicine, both fine-tuned from LlaMa-7B. Zhongjing-TCM is specialized in TCM gynecology Q&A tasks. The statistical information is presented in Table 2.

Model	$\overline{ m TMNLI(Test)}$	TCM-ED(Test)	
DeBERTa-v3-base	33.2%	0.0%	
DeBERTa-v3-base-mnli	41.5%	4.08%	
TCM-Deberta	96.48%	95.38%	

Table 3: The accuracy of different NLI models on two testing datasets.

4.2 Experimental Settings

We conduct extensive experiments to evaluate the zero-shot performance of LLMs, ensuring their capability to respond in a multiple-choice format and provide corresponding analyses. Additionally, we partition the TCM-ED dataset based on medical branches and question types, conducting independent tests on each subset for comprehensive analysis. Depending on the question type, we design different prompt templates, including task descriptions, Chain-of-Through (CoT), and Q&A constraints. The task description clarifies the question types that LLMs need to answer. The CoT guides LLMs in giving the options and providing corresponding analyses simultaneously, which can evaluate the ability of LLMs to understand and express TCM knowledge comprehensively. The specific content of the prompt template is shown in Section B of the appendix file in additional materials. Particularly, due to the shared content among several questions of types A3 and B1, strong logical coherence exists between the questions. To evaluate the logical reasoning ability of LLMs in TCM, we adopt a multi-turn dialogue format, using answers from preceding questions as historical context for subsequent dialogues. Moreover, we observe that the A3 type of questions closely resemble real-world clinical diagnosis and treatment processes, yet requiring LLMs to answer in a fixed format poses considerable difficulty. Hence, in such questions, we introduce the few-shot to incorporate an A3-type answer example at the beginning of the question as a prompt for the answering format.

4.3 Evaluation Metrics

We evaluate both general and medical LLMs using TCM-Bench, following the evaluation process depicted on the right of Figure 3, which comprises two key steps. Firstly, we employ accuracy as the evaluation metric to automatically compare the options generated by LLMs with the correct options, evaluating their understanding and application ability for TCM knowledge. Secondly, we choose 1,300 questions with standard analysis from TCM-ED to automatically evaluate the quality of LLMs in expressing knowledge in TCM. We employ traditional metrics of text generation tasks, including word matching methods like ROUGE [8] and SARI [16], as well as deep learning methods like BertScore [21] and BartScore [17], to compare the semantic similarity between the analysis generated by LLMs and the standard analysis. Additionally, we introduce the expert-level supplement metric TCMScore, which reflects TCM semantics and knowledge consistency. The two evaluation processes complement each other,

providing a more comprehensive perspective on evaluating the medical performance of LLMs in TCM.

Now, we introduce TCMScore. Firstly, we fine-tune the NLI model, DeBERTa-v3-base-mnli ⁶, to create TCM-Deberta tailored for inferring TCM semantic consistency between two sentences. To further illustrate its effectiveness, we evaluate the inference accuracy of different NLI models on the TMNLI test set. Moreover, we evaluate the model accuracy of predicting the relationship between standard analysis and QWA in TCM-ED, whose data differ from TMNLI. Among this, we employ a more stable method in which set analysis and QWA are both the premise and hypothesis. When the two inference results are the entailment, we consider a stable semantic consistency between them. The inference results presented in Table 3 demonstrates that TCM-Derberta achieves stable and high accuracy on two testing datasets.

In addition, to evaluate the TCM knowledge consistency between texts, we design a metric, Term F1 Score($F1^*$), which quantitatively measures the matching score of TCM terms between two text. The core idea of the $F1^*$ is to comprehensively consider the redundancy (i.e., precision), matching degree (i.e., recall), and term diversity of TCM terminologies. The $F1^*$ is as follows.

$$Precision = \frac{|M(T(S_1), T(S_2))|}{T(S_1)},$$

$$Recall = \frac{|M(T(S_1), T(S_2))|}{T(S_2)},$$

$$Diversity = \frac{|M(T(S_1), D)|}{T(S_1)},$$

$$F1^* = 3 \times \frac{Precision \times Recall \times Diversity}{Precision + Recall + Diversity}.$$
(1)

Guided by experts, we standardize 61,987 TCM terminologies from official publications on TCM diagnosis and treatment, TCM disease and syndrome codes, and TCM-KB[14] to create the TCM terminology database, D. $T(S) = Counter(S) \cap Counter(D)$ represents the set of TCM terminologies and their quantities in sentence S, and $M(T_1, T_2) = T_1 \cap T_2$ denotes the matching quantities of TCM terminologies between two sets.

Finally, we combine $F1^*$ and the TCM-Deberta model to construct TCM-Score. Its essence lies in focusing more on the semantic consistency of the LLMs generated sentence when its TCM terms match numerous with the standard analysis. Conversely, if the term matching is low, even if the semantics of the sentence are similar to the analysis, it appears relatively unimportant. Since the evaluated text is lengthy, we employ a sentence-by-sentence analysis method. We calculate the $F1^*$ score between each standard analysis sentence and LLMs-generated sentence to measure the knowledge matching degree. This score is then normalized as the importance weight when evaluating the semantic consistency of the sentence in LLMs-generated response. Next, we use the TCM-Deberta to calculate the semantic consistency score between each pair of sentences, which

⁶ https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli

Algorithm 1: The calculation process of TCMScore.

```
Input: The LLMs generated sentences S, The standard analysis sentence A Output: TCMScore

for s_a \in A do

| for s_{LLM} \in S do
| F1^* \leftarrow calculate Term F1 score between s_{LLM} and s_a;
| nli_{score} \leftarrow (TCM-Deberta(s_{LLM}, s_a) + TCM-Deberta(s_a, s_{LLM})) / 2;
| add F1^* to a list F1^*_{LLM}, and add NLI_{score} to a list Senmentic_{LLM};
end
| for i \leftarrow 1 to length(F1^*_{LLM}) do
| W_{senmentic} \leftarrow F1^*_{LLMs}[i]/\sum F1^*_{LLM} \times Senmentic_{LLM}[i];
| add W_{senmentic} to a list L_{TCMScore};
end
end

w_{length} = e^{1-log(max(|S|,|A|)/log(min(|S|,|A|))};
TCMScore \leftarrow \sum L_{TCMScore}/length(L_{TCMScore}) \times w_{length};
return TCMScore
```

Table 4: The accuracy on LLMs in three question types of TCM-Bench. The best performing model is bold, while the strongest models are underlined.

LLMs	A1/A2	A3 (zero-shot)	A3(few-shot)	B1	Total
Chinese LlaMa	0.0969	0.1075	0.1620	0.1151	0.1089
HuaTuo	0.1944	0.1981	0.1402	0.1876	0.184
ZhongJing-TCM	0.3537	0.3364	0.3178	0.2182	0.2695
ChatGLM	0.3613	0.4595	0.6168	0.4568	0.4477
ChatGPT	0.4510	0.4657	0.4782	0.4444	0.4398
GPT-4	0.5819	0.6231	0.6277	0.6011	0.5986

is then multiplied by weight to obtain the weighted semantic consistency score. Finally, we summarize the scores of each sentence to determine the difference between the analysis generated by LLMs and the standard analysis, resulting in overall TCM semantic and knowledge consistency scores. Moreover, we introduce a length penalty term w_{length} to balance the impact of text length differences. It imposes a more significant penalty on short text than on long text. The calculation process is outlined in Algorithm 1.

4.4 Main Results

Analysis of LLMs Accuracy. From Table 4, we comprehensively analyze the accuracy for different LLMs in TCMLE. The main findings are as follows.

(1) None of the tested LLMs passed the TCMLE. A notable observation is that accuracy improves with increasing model parameters. GPT-4 consistently outperforms other LLMs despite some being trained on extensive Chinese or medical corpora. Even so, the total accuracy of GPT-4 does not exceed 60%,

which is the minimum requirement for passing TCMLE. This also indicates that there is still significant room for improvement in the medical performance of LLMs in the TCM domain.

- (2) In the pre-training stage of LLMs, incorporating domain-specific knowledge becomes more crucial in the same magnitude order of parameters. Despite having over 100 billion parameters, the overall accuracy of ChatGPT is lower than ChatGLM. It is due to ChatGLM using a more extensive Chinese corpus during the pre-training stage that its ability to understand Chinese-based TCM questions is enhanced. However, the slight difference in overall accuracy between the two models highlights the gap between general Chinese and TCM semantics. In the 7 billion parameter level LLMs, incorporating professional knowledge of medicine, especially TCM, during pre-training can significantly improve model performance. For example, ZhongJing-TCM achieves an accuracy of 35.37% on the A1/A2 type of questions, which is only 2% lower than ChatGPT despite a 25-fold difference in the parameter numbers. The comparison powerfully illustrates that merely increasing the parameter numbers is not optimal when dealing with specific domains like TCM, which has profound cultural backgrounds and professional terminology. Instead, carefully designing and integrating high-quality professional TCM data during the pre-training stage is an effective way to enhance the performance of LLMs in TCM applications.
- (3) Adding examples to prompts can enhance the ability of LLMs to handle complex logical reasoning. We can observe the unsatisfactory zero-shot performance on ChatGLM, ChatGPT, and Chinese LlaMa, as shown in Table 4. However, the performance of these models significantly improved once examples are introduced, with ChatGLM showing a performance increase of 34.23%. This demonstrates that designing compelling TCM examples can enhance LLMs' understanding of complex reasoning logic in the TCM field.
- (4) Adding examples in prompts to guide LLMs through complex logical reasoning may not necessarily be effective. Incorporating domain knowledge may damage the model's original logical reasoning ability in the fine-tuning process of LLMs. Their performance decreases when HuaTuo and Zhongjing-TCM engage in few-shot learning through examples. The possible reason is that adding examples would create a longer prompt, which exceeds the capability of LLMs to handle long texts. Therefore, in the future, while enhancing the domain adaptability of LLMs, it is crucial to maintain and optimize their ability to handle complex and lengthy text logical tasks.
- (5) LLMs perform differently across various medical branches. We further evaluate their accuracy in addressing questions within different medical domains. Firstly, based on the scope of the TCMLE exam, we categorize medical branches in TCM-Bench into three groups: TCM Basis, TCM Clinical Medicine, as well as Western Medicine and Clinical Medicine. The accuracy of different models in each category is illustrated in Figure 7, with each category comprising five medical branches corresponding to subplots of a row in Figure 6. Figure 6 details the model's accuracy across each question type. If LLMs perform well on A3 questions simulating clinical scenarios, they are better suited for complex

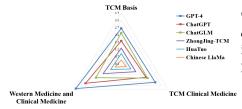


Table 5: The comparative evaluation of GPT-4 and HuaTuo in answering to $\rm A1/A2$ questions under various prompts.

$\mathbf{Model} \mid \mathbf{CoT} \mid \mathbf{No} \ \mathbf{CoT} \mid \mathbf{Answer} \ 5 \ \mathbf{times}$				
GPT-4 0.	5819 0.5475 1944 0.185	0.5475		
HuoTuo 0.	1944 0.185	0.1856		

Fig. 5: The total accuracy results on category of TCMLE.

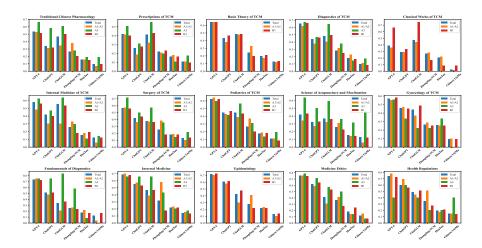


Fig. 6: The accuracy results on different branches of TCM-Bench.

clinical case analysis tasks. Additionally, the high accuracy of A3 and B1 questions through multiple rounds of dialogue indicates their effective understanding and correlation of various medical knowledge points. A notable finding is that GPT-4 and ChatGPT perform well in Western medicine. In contrast, ChatGLM excels in TCM Basis, especially in correlating and analyzing the knowledge of classic works of TCM. It highlights the importance of the Chinese corpus in comprehending theoretical TCM knowledge, but more professional knowledge is needed for clinical assistance. Although ZhongJing-TCM is trained based on the corpus generated from TCM gynecological medical records, it performs well in all branches, surpassing ChatGPT on the A1/A2 question of five branches, like Traditional Chinese Pharmacology. It also surpasses ChatGLM on four Western medical branches' A1/A2 questions. This demonstrates the model's performance in cross-domain knowledge transfer and comprehensive application.

(6) Chain-of-Thought prompting and model stability. In TCM-Bench, we set up CoT-based prompts for evaluation. In addition, we compare LLMs without using CoT prompts, and the accuracy is shown in Table 5. After removing CoT, the overall performance of LLMs declined, confirming the importance

performing model is bold, while the strongest models are underlined.						
LLMs	Rouge-1	Rouge-L	SARI	BertScore	BartScore	TCMScore
ZhongJing-TCM HuaTuo Chinese LlaMa	3.41% 2.52% 2.01%	3.38% 2.48% 1.93%	3.14% $2.95%$ $9.99%$	48.01% 52.01% 56.92%	-6.06 -5.45 -4.4	7.84 % 8.8 % 10.48%
ChatGLM ChatGPT GPT-4	$\begin{array}{c c} 4.85\% \\ 4.93\% \\ 4.8\% \end{array}$	$\begin{array}{ c c } \hline 4.7\% \\ 4.8\% \\ 4.6\% \\ \hline \end{array}$	$21.58\% \underline{23.93\%} \mathbf{25.34\%}$	$\overline{66.95\%}$	-4.17 -3.92 -3.91	43.49% 43.39% 4 5.71%

Table 6: The expression quality of the response of LLMs on TCM-Bench. The best performing model is bold, while the strongest models are underlined.

of CoT cues in enhancing model comprehension of TCM knowledge. To evaluate the stability of LLMs without CoT, we take LLMs by answering each question 5 times and calculate the average score. The results showed that LLMs can still maintain high stability in this scenario.

Analysis of expression quality of LLMs. We leverage three types of evaluation metrics: (a) word matching based methods like Rouge-1, Rouge-L, and SARI, (c) deep leaning based methods like BertScore, and BartScore, and (c) a hybrid method to professional term matching and deep learning, i.e., TCMScore to evaluate the LLMs' expression quality comprehensively. From Table 6, We find GPT-4 consistently demonstrates outstanding performance, but ZhongJing-TCM is the worst. Further findings are as follows.

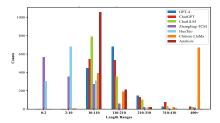
- (1) Metrics based on word matching are affected by the length of the generated text. Rouge favors LLMs with minimal differences in length compared to the standard analysis. Analyzing the Rouge score and the generated text length that depicted in Figure 8, ChatGPT and ChatGLM have an advantage due to their generated text length being closer to the standard analysis. Additionally, Rouge focuses on the recall metric, meaning that if the shorter text that LLMs generated, it may score higher. This explains why ZhongJing-TCM, generating less content yet maintaining higher accuracy, outperforms HuaTuo in Rouge score, while HuaTuo surpasses Chinese LlaMa. From the statistics in Figure 7 for the 0-2 to 2-10 length range, it is evident that HuaTuo and ShenNong-TCM struggle to generate analysis, emphasizing the need to maintain fundamental analysis and reasoning capabilities when fine-tuning domain knowledge again. We calculate the SARI score with retention rate due to without extra reference text. It considers word frequency to evaluate the information content of the generated text. For instance, GPT-4 excels in this metric. However, LLMs generating longer texts like Chinese LlaMa can achieve higher SARI scores despite potential errors due to containing more information. The above analysis confirms that relying solely on word matching to evaluate TCM text generation quality introduces bias due to text length.
- (2) Deep learning-based metrics like BertScore and BartScore can evaluate the semantic similarity of generated texts but do not directly evaluate or interpret the accuracy of expertise in the texts. GPT-4

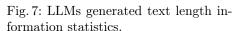
maintains its super in the two metrics, indicating high semantic similarity between its generated text and standard analysis. ChatGLM surpasses ChatGPT in BertScore, indicating a better ability to produce concise and focused responses. The analysis of the generated text length from Figure 8 further supports this. Moreover, Chinese LlaMa outperforms HuaTuo and ZhongJing-TCM in both metrics, indicating its strength in generating diverse and complex semantic content. The texts it produces are often longer and offer more detailed analysis and expansion. However, it is essential to note that the length and semantic coherence of the generated text do not guarantee the absolute correctness of the content. In summary, evaluating text in a knowledge-intensive domain like TCM requires evaluating not just language fluency and semantic consistency but also specific metrics related to professional knowledge accuracy and domain adaptability. This ensures a comprehensive quality evaluation of the text.

- (3) Further, the findings will be supplemented by using TCMScore. TCMScore serves as a comprehensive supplementary evaluation metric, which considers both semantic and knowledge consistency in TCM, along with the length of the generated text. With the introduction of TCMScore, we can draw the following conclusions to complement the findings above:
- (a) GPT-4 can generate highly accurate content with rich TCM characteristics and expand knowledge boundaries. Its significant advantages on TCMScore highlight its prowess. Despite potentially lower scores on Rouge, this underscores the advantage of GPT-4 in providing in-depth expansion and refinement while ensuring semantic consistency with the standard analysis. It offers a more comprehensive and detailed analysis related to TCM.
- (b) Wrong facts in the LLMs generated text diminish their advantage in semantic similarity. ChatGPT and Chinese LlaMa can expand TCM knowledge, but the accuracy of their generated content still needs to be improved. While ChatGPT outperforms ChatGLM in most metrics after integrating TCMScore, its advantage diminishes due to the redundancy and errors in the generated content. Chinese LlaMa faces a similar issue, as evidenced by its reduced advantage in TCMScore despite excelling in BartScore and BertScore. This indicates a higher likelihood of incorrect information in its generated content to decrease its semantic advantage. Overall, these LLMs have room to enhance the accuracy of their generated content and reduce misinformation.
- (c) The correctness of text content does not equal the effective application of TCM knowledge. For instance, while ZhongJing-TCM may provide accurate answers, it lacks essential skills like language fluency, knowledge coherence, and logical reasoning, resulting in poor performance on TCMScore. Therefore, to advance LLMs in TCM, enhancing their ability to apply knowledge is crucial while ensuring a deep understanding of domain-specific concepts.

4.5 Human evaluation

In addition to automatic evaluation, we invited a TCM expert and a medical doctoral student to conduct a manual evaluation of 18 questions related to TCM





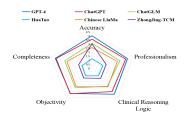


Fig. 8: Human evaluation result.

fundamentals and clinical knowledge to evaluate the performance of LLMs quantitatively. The evaluation dimensions include accuracy, professionalism, clinical reasoning logic, objectivity, and comprehensiveness, see Figure 8. For fairness, LLMs' names are anonymized during the evaluation process. The results revealed that GPT-4 excelled across all evaluation dimensions. ChatGPT showed slightly lower accuracy and clinical reasoning logic than GPT-4. Notably, Chinese LlaMa matched ChatGLM in objectivity and clinical reasoning logic despite its lower accuracy, which reflects that it can preserve its basic model's capabilities. ZhongJing-TCM exhibited a strong understanding of TCM knowledge on the accuracy dimension. However, it scored lower on other dimensions due to challenges in providing specific analyses (similar to HuaTuo).

5 Conclusion

In this paper, we introduce the TCMBench, a comprehensive benchmark for evaluating the performance of LLMs in TCM. The experiment shows that the performance of LLMs in this field is unsatisfactory. It also highlights the importance of maintaining the basic capabilities of LLMs while inducing domain expertise in their fine-tuning process. We also analyze the domain-specific metrics, like our TCMScore, which can further supplement and explain the evaluation results of traditional metrics for text generation.

Furthermore, experiments reveal that some LLMs produce wrong information (i.e., hallucination phenomenon) when generating content. This area will be a focus of our future in-depth research aimed at developing effective methods to identify and quantify such issues. Considering the central role of clinical practice in TCM, we plan to expand the data source, which includes actual TCM case data, covering the entire TCM diagnosis and treatment process. Notably, we will concentrate on accurately evaluating if LLMs can follow the unique clinical logic of TCM, which is syndrome differentiation and treatment, thereby enhancing and refining our benchmark.

References

- Brockman, G., Eleti, A., Georges, E., Jang, J., Kilpatrick, L., Lim, R., Miller, L., Pokrass, M.: Introducing chatgpt and whisper apis. https://openai.com/blog/ introducing-chatgpt-and-whisper-apis (2023)
- Cai, Y., Wang, L., Wang, Y., de Melo, G., Zhang, Y., Wang, Y., He, L.: Medbench: A large-scale chinese benchmark for evaluating medical large language models. arXiv preprint arXiv:2312.12806 (2023)
- 3. Cheung, F.: Tcm: made in china. Nature 480(7378), S82-S83 (2011)
- 4. Cui, Y., Yang, Z., Yao, X.: Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177 (2023)
- Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences 11(14), 6421 (2021)
- Kuhn, L., Gal, Y., Farquhar, S.: Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv preprint arXiv:2302.09664 (2023)
- Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., et al.: Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. PLoS digital health 2(2), e0000198 (2023)
- 8. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
- 9. Liu, J., Zhou, P., Hua, Y., Chong, D., Tian, Z., Liu, A., Wang, H., You, C., Guo, Z., Zhu, L., et al.: Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. Advances in Neural Information Processing Systems 36 (2024)
- 10. OpenAI: Gpt-4 technical report (2023)
- 11. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on Health, Inference, and Learning. pp. 248–260. PMLR (2022)
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature 620(7972), 172–180 (2023)
- Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., Liu, T.: Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975 (2023)
- 14. Wang, X., Zhang, Y., Wang, X., Chen, J.: A knowledge graph enhanced topic modeling approach for herb recommendation. In: International Conference on Database Systems for Advanced Applications. pp. 709–724. Springer (2019)
- 15. Williams, A., Nangia, N., Bowman, S.R.: A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426 (2017)
- Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C.: Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics 4, 401–415 (2016)
- 17. Yuan, W., Neubig, G., Liu, P.: Bartscore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems 34, 27263–27277 (2021)
- Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022)

- Zhang, N., Chen, M., Bi, Z., Liang, X., Li, L., Shang, X., Yin, K., Tan, C., Xu, J., Huang, F., et al.: Cblue: A chinese biomedical language understanding evaluation benchmark. arXiv preprint arXiv:2106.08087 (2021)
- Zhang, Q., Zhou, J., Zhang, B.: Computational traditional chinese medicine diagnosis: a literature survey. Computers in Biology and Medicine 133, 104358 (2021)
- 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. Zhu, W., Wang, X., Zheng, H., Chen, M., Tang, B.: Promptcblue: A chinese prompt tuning benchmark for the medical domain. arXiv preprint arXiv:2310.14151 (2023)

A The Question Type of TCM-EB

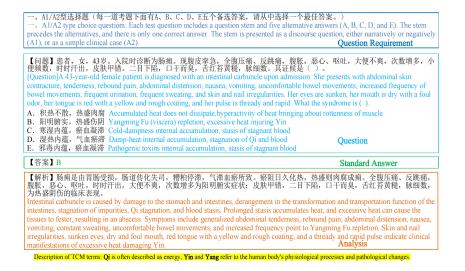


Fig. 9: The example of A1/A2 type of questions. The question requirement is indicated in dark blue text, the question along with the five options is in light blue text, the standard answer is in green text, and the standard analysis is in orange text. The related TCM terms are explained in the yellow highlight.



Fig. 10: The example of A3 type of questions. The question requirement is indicated in dark blue text, the patient-centered case is in light blue text, the first sub question along with the five options, standard answer and analysis is in green text, the second sub question is in orange text, and the second sub question is in purple text.

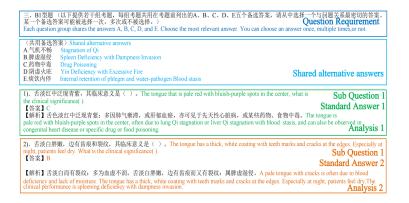


Fig. 11: The example of B1 type of questions. The question requirement is indicated in dark blue text, five options is in light blue text, the first sub question along with the five options, standard answer and analysis is in green text, and the second sub question is in orange text.

B Prompt and target output formats for Evaluation

```
(Task Description) Please do a multiple-choice question of type A1 or A2 in a TCM exam. (清你像一道中医测试中A1或者A2类型的选择题。) (CoT and Q&A Constraint) Please think step by step and write your thinking process between [Analysis] and<coc> You will choose the most correct answer from A, B, C, D, and E and write it between [Answer] and<coc> (请你一步一步思考并将思考过程写在【解析】和<coc>之间。你将从A, B, C, D, 戶中选出一个最正确的答案,并写在【答案】和<coc>之间。(请你一步一步思考并将思考过程写在【解析】和<coc>之间。你将从A, B, C, D, 戶中选出一个最正确的答案,并写在【答案】和<coc>之间。(请你一步一步思考并将思考过程写在【解析】和<coc>之间。你将从A, B, C, D, 中选出一个最正确的答案,并写在【答案】和<coc>问题:【答案】:A<coc>】
[Analysis]: —<coc〉【解析】 —<coc〉】
[Analysis]: —<coc〉【解析】 —<coc〉】
[Analysis]: —<coc〉【修作】 —<coc〉】
[Analysis]: —<coc〉【修作】 —<coc〉】
[Analysis]: —<coc〉【答案】 —<coco
```

Fig. 12: The zero-shot prompt template target output formats for evaluating A1/A2 type of questions.

Fig. 13: The zero-shot prompt template target output formats for evaluating A3 type of questions.

Fig. 14: The few-shot prompt template target output formats for evaluating A3 type of questions.

```
| Ritype | Clask Description | Please do a multiple-choice question of type B1 in a TCM exam. Each set of questions contains several multiple-choice questions, with five options listed as A, B, C, D, and E. Please choose the answer that is most closely related to the question. Each option may be chosen one, multiple times, or not at all. (情俗物: 当世 Westingthering Westingstang 1988] 希望國籍自有者 干脆速图器。共和列出的A, B, C, D, ET-设置 经营业 (情俗物: 当世 Westingthering Westingstang 1989) | Please choose the most contains a contains a contained by the part of the part of the contained by the part of t
```

Fig. 15: The zero-shot prompt template target output formats for evaluating B1 type of questions.