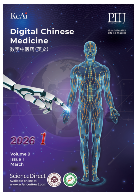
Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed



QingNangTCM: a parameter-efficient fine-tuning large language model for traditional Chinese medicine

Xuming Tong[a, b†], Liyan Liu[b†], Yanhong Yuan[c], Xiaozheng Ding[b], Huiru Jia[b], Xu Yang[a], Sio Kei Im[d], Mini Han Wang[e], Zhang Xiong[f], Yapeng Wang[a*]

a. Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China

b. School of Information Science and Engineering, Hebei North University, Zhangjiakou, Hebei 075000, China

c. Academic Affairs Office, Hebei North University, Zhangjiakou, Hebei 075000, China

d. Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence of Ministry of Education, Macao Polytechnic University, Macao 999078, China

e. Faculty of Medicine, Chinese University of Hong Kong, Hong Kong 999077, China

f. School of Computer Science and Engineering, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history
Received 21 June 2025
Accepted 05 January 2026
Available online 25 March 2026

Keywords
Large language model (LLM)
Traditional Chinese medicine (TCM)
Fine-tuning
P-Tuning v2
Clinical decision support

ABSTRACT

Objective To develop QingNangTCM, a specialized large language model (LLM) tailored for expert-level traditional Chinese medicine (TCM) question-answering and clinical reasoning, addressing the scarcity of domain-specific corpora and specialized alignment.

Methods We constructed QnTCM_Dataset, a corpus of 100 000 entries, by integrating data from ShenNong_TCM_Dataset and SymMap v2.0, and synthesizing additional samples via retrieval-augmented generation (RAG) and persona-driven generation. The dataset comprehensively covers diagnostic inquiries, prescriptions, and herbal knowledge. Utilizing P-Tuning v2, we fine-tuned the GLM-4-9B-Chat backbone to develop QingNangTCM. A multidimensional evaluation framework, assessing accuracy, coverage, consistency, safety, professionalism, and fluency, was established using metrics such as bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), metric for evaluation of translation with explicit ordering (METEOR), and LLM-as-a-Judge with expert review. Qualitative analysis was conducted across four simulated clinical scenarios: symptom analysis, disease treatment, herb inquiry, and failure cases. Baseline models included GLM-4-9B-Chat, DeepSeek-V2, HuatuoGPT-II (7B), and GLM-4-9B-Chat (freeze-tuning).

Results QingNangTCM achieved the highest scores in BLEU-1/2/3/4 (0.425/0.298/0.137/0.064), ROUGE-1/2 (0.368/0.157), and METEOR (0.218), demonstrating a balanced and superior normalized performance profile of 0.900 across the dimensions of accuracy, coverage, and consistency. Although its ROUGE-L score (0.299) was lower than that of HuatuoGPT-II (7B) (0.351), it significantly outperformed domain-specific models in expert-validated win rates for professionalism (86%) and safety (73%). Qualitative analysis confirmed that the model strictly adheres to the "symptom-syndrome-pathogenesis-treatment" reasoning chain, though occasional misclassifications and hallucinations persisted when dealing with rare medicinal materials and uncommon syndromes.

†The authors contributed equally.

*Corresponding author: Yapeng Wang, E-mail: yapengwang@mpu.edu.mo.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2026.02.002

Citation: TONG XM, LIU LY, YUAN YH, et al. QingNangTCM: a parameter-efficient fine-tuning large language model for traditional Chinese medicine. Digital Chinese Medicine, 2026, 9(1): 1-12.

Conclusion Combining domain-specific corpus construction with parameter-efficient prompt tuning enhances the reasoning behavior and domain adaptation of LLMs for TCM-related tasks. This work provides a technical framework for the digital organization and intelligent utilization of TCM knowledge, with potential value for supporting diagnostic reasoning and medical education.

1 Introduction

The rapid evolution of large language models (LLMs)[1] has fundamentally transformed medical informatics through enhanced dialogue management and knowledge-intensive reasoning[2]. However, for specialized domains like traditional Chinese medicine (TCM), the focus has shifted from general pretraining toward more standardized domain alignment and inference strategies to handle its unique semantic complexity. The maturation of Chinese-optimized open-source backbones such as Baichuan[3], Qwen[4], DeepSeek-V2[5], and ERNIE Bot[6] now provides a robust linguistic substrate for such adaptation. Given TCM's integration of classical canons and complex clinical reasoning, it has emerged as a strategically vital application domain for validating the professional utility and diagnostic accuracy of specialized artificial intelligence (AI)[7, 8].

TCM constitutes both a codified medical system and a body of cultural heritage, integrating classical canons, clinical case records, and materia medica. Its clinical practices and canonical texts, including acupuncture, moxibustion, the *Huangdi Neijing* (《黄帝内经》, *Inner Canon of Huangdi*), and *Bencao Gangmu* (《本草纲目》, *Compendium of Materia Medica*), are internationally recognized for their historical and epistemic value. Notwithstanding this foundation, the digital and intelligent transformation of TCM remains constrained by a lack of robust platforms and clinically reliable AI-assisted diagnostic tools, which limits improvements in diagnostic accuracy, accessibility, and standardization. While LLMs offer a promising pathway for structuring knowledge, performing representation learning on classical corpora, and supporting for clinical reasoning and decision-making, general-purpose models often fail to capture TCM's core ontological principles and semantic frameworks. This leads to superficial treatment of herbal entities, formula composition, and case-based syndrome differentiation, ultimately constraining domain fidelity and clinical utility[9]. These limitations have motivated the development of medical and TCM-specialized LLMs that incorporate domain knowledge and task-specific supervision[10].

Given the prohibitive cost of training from scratch, domain adaptation via parameter-efficient fine-tuning has become the prevailing strategy[11]. P-Tuning v2[12], for instance, achieves competitive performance by optimizing only 0.1% – 3% of parameters while significantly reducing computational overhead. Based on these techniques, several TCM-specialized models have been introduced and can be categorized by their technical focus. The first group, including DoctorGLM[13] and ShenNong-TCM[14], utilizes lightweight adapters to enhance consultation and entity recall, yet often struggles to achieve deep consistency with TCM pathological logic, particularly in the linkage between Bianzheng (辨证, syndrome differentiation) and formula prescription. The second group, such as MedChatZH[15] and HuatuoGPT[16], emphasizes dialogue fluency via curated medical corpora, but exhibits superficial specialization and insufficient grounding in canonical TCM texts. The third group, represented by Qibo[17] and Zhongjing[18], integrates comprehensive pre-training or feedback mechanisms; however, these models still face challenges like inconsistent citation alignment and limited robustness in authentic clinical narratives. Collectively, a significant research gap remains in ensuring that the model's reasoning path is deeply consistent with TCM pathological logic while maintaining reliable stability across diverse and real-world clinical scenarios.

To address these gaps, this study introduces QingNangTCM, a specialized LLM tailored for expert-level TCM question-answering and clinical reasoning. We proposed a systematic research framework that began with the curation of a domain-specific corpus, designed to align model representations with TCM pathophysiological theories. Subsequently, we leveraged the parameter-efficient fine-tuning method (P-Tuning v2) to specialize a general-purpose backbone for advanced medical tasks. Finally, we established a comprehensive, multidimensional evaluation framework that integrates automated metrics with expert verification to assess the model's capabilities. This study seeks to validate the efficacy of the proposed model in enhancing diagnostic and treatment support, thereby contributing a practical pathway for the digital and intelligent modernization of TCM.

2 Data and methods

This section discusses the construction process of the QingNangTCM model, including the development of the dataset, the fine-tuning of the model, as well as the evaluation of its capability.

2.1 Construction and processing of the dataset

2.1.1 Dataset construction The QnTCM_Dataset, an instruction dataset for fine-tuning LLMs in TCM domain, was constructed in the study. It was assembled by integrating two primary data sources and applying a multi-stage preprocessing pipeline to ensure task relevance and data integrity.

The first source is ShenNong_TCM_Dataset [14], which contains over 110 000 instruction instances covering fundamental TCM theories, syndrome differentiation-oriented reasoning, herb compatibility, prescription inference, and clinical applications. According to the dataset documentation, part of the ShenNong_TCM_Dataset is derived from classical TCM literature, including *Huangdi Neijing*, *Shanghan Lun* (《伤寒论》, *Treatise on Febrile and Miscellaneous Diseases*), and *Shennong Bencao Jing* (《神农本草经》, *Divine Farmer's Materia Medica*). To ensure quality of the dataset, we applied a multi-stage curation pipeline. First, rule-based cleaning using regular expressions removed non-informative symbols, Chinese-English mixed sentences, and duplicate entries. Next, semantic-level screening filtered out samples with incomplete context or reasoning structures inconsistent with TCM clinical logic. Finally, the retained data underwent manual proofreading and verification. This process yielded 80 000 entries, spanning core TCM tasks such as clinical question answering, formula prescription guidance, and diagnostic reasoning.

The second source is the SymMap v2.0 TCM syndrome association database [19], which was integrated to strengthen the dataset's structured knowledge of the relationship between TCM symptoms, syndromes, herbal medicines, and their connections to modern medical concepts. SymMap v2.0 comprises 1 717 TCM symptoms, 499 herbs, 961 modern medical symptoms, 5 235 diseases, and 19 595 herbal constituents.

After format standardization, 703 instances were incorporated into QnTCM_Dataset, thereby improving the coverage of herbal attributes, clinical use scenarios, and dosage-related knowledge within the training data.

2.1.2 Data augmentation via retrieval-augmented generation (RAG) and persona-driven generation To address data scarcity and enhance the quality of conversational data, we implemented two targeted augmentation strategies. First, adhering to RAG principles [20], we utilized the LangChain framework to retrieve contextually relevant evidence from the foundational ShenNong_TCM_Dataset and SymMap v2.0 datasets. This approach grounded the generation process in verified domain knowledge, effectively mitigating hallucinations and ensuring medical factual accuracy. Second, we employed a persona-driven generation strategy using GLM-4-9B-Chat [21, 22]. Guided by a carefully designed system prompt (Supplementary Table S1), this strategy established a "senior TCM practitioner" persona. The prompt imposed strict logical constraints, instructing the model to generate responses that strictly adhere to the "symptom-syndrome-pathogenesis-treatment" TCM reasoning chain. This method effectively refined semantic clarity and ensured professional consistency in the generated medical responses. Through these strategies, a total of 21 000 raw candidate entries were generated for further validation.

2.1.3 Data validation and safety verification To ensure high data quality, we established a comprehensive quality control framework focused on medical accuracy and data safety.

First, regarding medical accuracy, a hybrid validation pipeline was applied to all augmented samples. (i) Manual expert review. Three licensed TCM practitioners (averaging five years of clinical experience) conducted a blinded assessment of a randomly selected 10% sample. Each entry was evaluated on a five-point Likert scale across factual accuracy, semantic clarity, and clinical safety, with the anchors defined as 1 (theory violation/incoherent reasoning/critical risk), 2 (major errors/semantic confusion/potential hazard), 3 (minor deviations/ambiguous phrasing/unverified advice), 4 (theory consistent/clear expression/safe guidance), and 5 (canonical alignment/professional terminology/fully compliant). Entries scoring < 4 in any dimension were subject to mandatory revision or rejection. The process demonstrated high interrater reliability (Cohen's kappa = 0.82). (ii) Automated cross-verification. All generated entries were systematically cross-referenced against authoritative knowledge bases (SymMap v2.0 and ShenNong_TCM_Dataset). This involved calculating Sentence-BERT embedding similarity [23], with a cosine similarity threshold of $\geqslant 0.85$ defined for semantic alignment. Additionally, inspired by the LLM-as-a-Judge framework [24], a GPT-4o-based screening strategy was implemented, requiring a minimum score of 4 out of 5 for logical consistency with the TCM diagnostic framework. This stage achieved an initial pass rate of 92%. The remaining 8% of flagged cases underwent secondary manual review, with unrecoverable errors being permanently discarded. Through this multi-stage process, a final set of 19 297 validated instances were retained for inclusion in the dataset.

Second, regarding data safety, although QnTCM_Dataset was constructed exclusively from publicly accessible sources, we implemented a dedicated screening and normalization protocol to eliminate potential privacy risks and ensure compliance. This involved applying regular expression-based rules to remove any personal identifiers, followed by natural language processing (NLP)-driven terminology normalization to reduce semantic ambiguity. Manual spot-checks were conducted to verify the absence of any residual sensitive information.

In summary, through a process of data cleaning, integration, validation, and screening for safety reasons, we curated the final QnTCM_Dataset, comprising 100 000 entries. To support domain research and transparency, this dataset will be released on GitHub under the CC BY-SA 4.0 license within three months after publication. As detailed in Table 1, this refined corpus provides a compliant and task-specific foundation for subsequent model training and evaluation.

Table 1 Statistics and composition of the QnTCM_dataset

Dataset	Scale	Attribute description	Validation method
ShenNong_TCM_Dataset	80 000	TCM consultations, prescriptions, and treatment methods for syndromes	Rule-based cleaning + manual review
SymMap v2.0	703	Types of Chinese herbal medicines, usage, and dosage	Rule-based cleaning + manual review
RAG + GLM-4-9B-Chat	19 297	TCM data generated by enhancement and optimization algorithm	Expert review + automated cross-verification

2.2 Foundational model

We selected GLM-4-9B-Chat as the backbone for QingNangTCM due to its strong Chinese language capabilities and its practical suitability for domain adaptation [25]. As a recent open-source model in the GLM series, GLM-4-9B-Chat supports long-context inputs, which was beneficial for incorporating evidence from TCM classics, clinical narratives, and materia medica descriptions. In this work, we primarily leveraged two properties of this backbone: (i) stable instruction following, which facilitated parameter-efficient adaptation with P-Tuning v2 to align the model closely with TCM clinical reasoning; (ii) tool-use compatibility, including support for function/tool calling interfaces when available, simplifying integration with our retrieval-augmented data construction pipeline. In addition, the model's 9 billion parameter scale offers a practical trade-off between performance and computational cost, making it suitable for resource-constrained fine-tuning and evaluation.

2.3 Model fine-tuning with P-Tuning v2 for TCM knowledge integration

To adapt GLM-4-9B-Chat for the TCM domain under resource constraints, we employed P-Tuning v2, a parameter-efficient fine-tuning strategy that specializes in a frozen backbone by learning a small set of continuous (soft) prompt parameters on the QnTCM_Dataset. This design enables effective domain adaptation with minimal parameter updates, while preserving the model's general language competence.

Formally, given an instruction-response pair (x, y), we optimized the conditional likelihood $p(y|x)$ while keeping all backbone weights frozen. We introduced layer-wise prefix prompts that were injected into the self-attention modules across all Transformer layers as key-value prefixes. These continuous prompts steered the model's intermediate representations toward TCM-specific semantics and reasoning patterns, learned end-to-end from data without reliance on handcrafted rules. This deep prompt-injection mechanism enabled the model to internalize domain knowledge across multiple levels of abstraction. In practice, prompts in the lower Transformer layers tend to capture lexical and surface-level features, such as symptom descriptions, while prompts in the upper layers guide higher-order reasoning structures, such as Bianzheng, Bingji (病机, pathogenesis) inference, and formula-herb associations. As illustrated conceptually in Figure 1, this hierarchical encoding ensures that TCM semantics are embedded throughout the entire generation process.

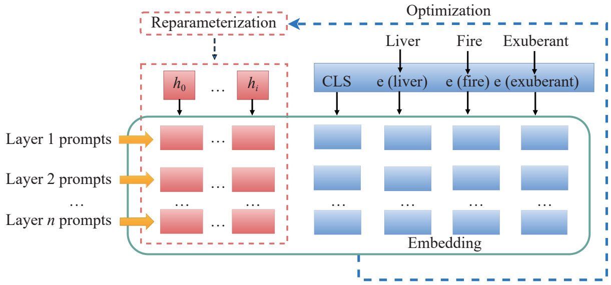


Figure 1 Architecture of the P-Tuning v2 deep prompt-based adaptation strategy used in QingNangTCM

The diagram highlights two primary components in the input embedding stack: the learnable prompt block (orange), which is optimized during training, and a frozen token embedding block (blue), which remains unchanged. For illustration, the phrase "liver fire excess" is tokenized and represented as word embeddings e (liver), e (fire), e (excess). The classification (CLS) token embedding is used to form a global sentence representation. Crucially, the vectors $\{h_0, h_1, \cdots, h_i\}$ denote continuous soft prompts learned by P-Tuning v2. These vectors are systematically injected into the attention computation of each Transformer layer to serve as prefix conditioning.

Compared with full supervised fine-tuning (SFT), which updates the entire parameter space and risks instability or overfitting in specialized domains, P-Tuning v2 limits updates to a lightweight set of prompt parameters. This approach enhances training efficiency and minimizes interference with the pre-trained backbone. Relative to methods like low-rank adaptation (LoRA), which inserts low-rank adapters into selected linear layers, layer-wise prefix prompting provides a simpler mechanism to modulate attention behavior throughout the entire network, supporting fine-grained domain alignment with minimal architectural changes.

2.4 Baseline models

To evaluate the performance of QingNangTCM, we selected four baseline models with distinct architectures and design objectives. These are categorized as follows. (i) General-purpose models: we employed GLM-4-9B-Chat (the foundational backbone of our model) [25] and DeepSeek-V2 (a mixture-of-experts model) [26]. The two models were applied to quantify the performance gains achieved by fine-tuning and to provide a reference for general reasoning capabilities in non-domain-specific contexts. (ii) Domain-specific model: we selected HuatuoGPT-II (7B) [16], a model developed via a single-stage adaptation strategy that integrates pre-training and fine-tuning, which has demonstrated competitive performance across established TCM benchmarks. (iii) Tuning-variant model: to validate the efficacy of our model, we implemented a freeze-tuning variant of GLM-4-9B-Chat [27, 28]. Under identical dataset conditions, only the last three Transformer layers were updated. This comparison highlights the specific advantages of P-Tuning v2 over partial parameter freezing.

2.5 Evaluation metrics

To systematically assess model performance, we established a multi-dimensional evaluation framework spanning accuracy, coverage, consistency, safety, professionalism, and fluency. Quantitative evaluation was first conducted using three complementary standardized metrics. All evaluations were performed on an independent test set constructed from QnTCM_Dataset and excluded from model training and fine-tuning. The test samples were designed to reflect representative TCM diagnostic and treatment-oriented question-answering scenarios, supporting the assessment of model performance within the proposed multi-dimensional evaluation framework.

(i) Bilingual evaluation understudy (BLEU)-1/2/3/4 [29], which measures n-gram precision, was used to evaluate accuracy by quantifying the model’s fidelity in reproducing standardized TCM clinical terminology.

(ii) Recall-oriented understudy for gisting evaluation (ROUGE)-1/2/L [30], a recall-oriented metric, was employed to assess coverage, evaluating whether the generated responses contain all key elements of the diagnostic reasoning chain.

(iii) Metric for evaluation of translation with explicit ordering (METEOR) [31], which incorporates synonym matching and stemming-aware alignment to mitigate the limitations of exact string matching, was adopted as a consistency metric to measure the semantic coherence and linguistic naturalness of TCM explanations. Collectively, these metrics provided a multi-faceted quantitative view of performance, measuring factual term matching, information coverage, and semantic alignment, respectively.

(iv) Hybrid evaluation strategy [32], which integrates LLM-as-a-Judge with expert review to transcend the limitations of surface-form overlap, was employed as a qualitative alignment protocol to evaluate higher-level attributes including safety, professionalism, and fluency. Safety prioritizes clinical reliability, harm prevention, and ethical compliance. Professionalism emphasizes the depth of comprehension, explanatory clarity, and clinical initiative. Finally, fluency evaluates discourse coherence, stylistic consistency, and professional empathy. Specifically, grounded in the methodological framework delineated by YANG et al. [18], the protocol first used GPT-4o as an auxiliary evaluator to produce preliminary scores based on predefined rubrics. These scores were then validated and calibrated by senior TCM practitioners to correct potential deviations from clinical and professional standards. This hybrid approach preserved domain fidelity and clinical reliability while substantially improving the scalability of safety-oriented evaluation.

2.6 Task-oriented clinical scenario simulation for qualitative assessment

To comprehensively assess the model’s performance under clinically motivated TCM consultation settings, such as adherence to the syndrome differentiation-based reasoning chain, the appropriateness of formula/herb recommendations, and potential safety risks, we designed a task-oriented clinical scenario simulation framework based on predefined, expert-curated prompts rather than real clinical cases. This qualitative assessment focuses on evaluating the model’s response behavior and reasoning coherence, rather than on conducting clinical validation. We designed four representative qualitative evaluation tasks.

(i) Symptom analysis. This scenario evaluates the model's ability to accurately diagnose TCM symptoms, including syndrome differentiation, identification of treatment principles, and generation of personalized recommendations while maintaining strict standards of safety and professionalism.

(ii) Disease treatment. This task presents prompts anchored by common disease names, requiring the model to perform TCM-style syndrome differentiation and propose appropriate therapeutic interventions. The evaluation focuses on whether the output remains strictly grounded in the TCM diagnostic-therapeutic framework, demonstrating coherent syndrome-treatment alignment rather than defaulting to recommendations for biomedical medications.

(iii) Herb inquiry. This task centers on knowledge-oriented questions about a single herb or medicinal materials. It evaluates the completeness and accuracy of provided herb attributes, as well as the adequacy of dosage guidance and contraindication/caution information.

(iv) Failure cases. Rare medicinal materials and uncommon syndromes are introduced for stress testing, aiming to identify typical failure modes in long-tail settings, including hallucinations, syndrome misclassification, unsafe recommendations, and breakdowns in the reasoning chain.

2.7 Experimental setup

2.7.1 Computational environment All fine-tuning and evaluation experiments were conducted in a high-performance computing environment running Ubuntu 20.04.5 long-term support (LTS). The software stack included Python 3.10.8, PyTorch 2.7.0, and CUDA 12.9, to ensure compatibility and optimized acceleration. Model fine-tuning was performed using two NVIDIA A800 GPUs (80 GB VRAM per card), providing the memory bandwidth required for efficient large-scale parameter optimization. To ensure reproducibility, the source code of QingNangTCM will be made available on GitHub under the Massachusetts Institute of Technology (MIT) license within the same timeframe as the dataset release.

2.7.2 P-Tuning v2 configuration This study adopted the P-Tuning v2 strategy for parameter-efficient fine-tuning. The configuration was as follows: 64 trainable prefix tokens were introduced at the input layer and mapped to the key and value matrices in every Transformer layer. For training, the maximum input and output sequence length was set to 512 tokens, a value chosen to fully represent clinical narratives while maintaining computational efficiency. A per-device batch size of 8 with 2 gradient accumulation steps was used, resulting in an effective global batch size of 32. This approach optimized memory usage while maintaining stable gradient updates. In the inference phase, the output length was also capped to 512 tokens to ensure the coherence and completeness of the generated responses.

2.7.3 Ablation studies and hyperparameter optimization As part of the experimental design, we conducted a systematic ablation study to empirically justify the selected hyperparameter configuration reported in the experimental setup and to assess model robustness, rather than to present additional performance results. The study was structured into three comparative groups. (i) Training duration: with the learning rate fixed at 1 × 10^{-4}, we varied the number of epochs (1 vs 3) to analyze convergence dynamics. (ii) Learning rate: under extended training, we compared learning rates of 1 × 10^{-4} and 1 × 10^{-5} to examine optimization stability and potential overfitting. (iii) Prefix length: using the optimal epoch and learning rate configuration, we evaluated prefix lengths of 32 versus 64 tokens to evaluate the trade-off between semantic alignment and feature coverage. The configuration of "epoch = 1, learning rate = 1 × 10^{-4}" served as the baseline for measuring incremental performance gains.

Figure 2 summarizes the end-to-end construction pipeline of QingNangTCM, including dataset curation, augmentation, validation, and parameter-efficient fine-tuning.

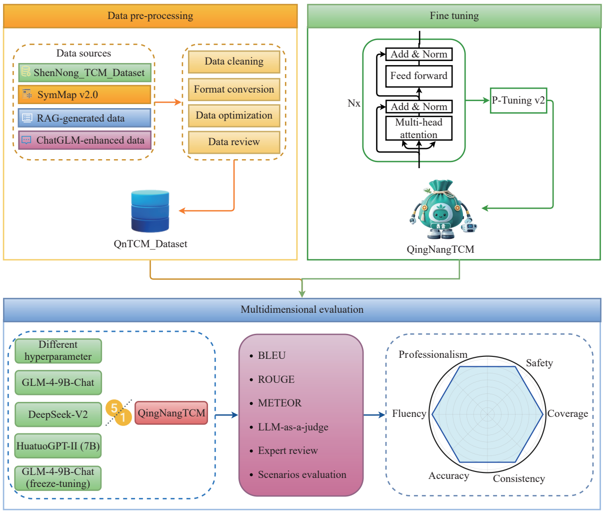


Figure 2 Overall workflow of QingNangTCM construction

3 Results

3.1 Quantitative evaluation via automatic metrics

A systematic quantitative evaluation was conducted using three complementary automated metrics, including BLEU, ROUGE, and METEOR, to assess performance across the dimensions of accuracy, coverage, and consistency. The evaluation compared our proposed model again general-purpose LLMs (DeepSeek-V2, GLM-4-9B-Chat) as well as domain-specific models tailored for TCM [HuatuoGPT-II (7B) and GLM-4-9B-Chat (freeze-tuning)].

Table 2 summarizes the comprehensive automated evaluation results, including ROUGE, METEOR, and BLEU metrics. QingNangTCM demonstrated superior

Table 2 Quantitative comparison of QingNangTCM and baseline models using BLEU, ROUGE, and METEOR metrics

LLM	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GLM-4-9B-Chat	0.312	0.129	0.244	0.127	0.385	0.262	0.108	0.037
DeepSeek-V2	0.345	0.120	0.276	0.135	0.368	0.273	0.114	0.051
HuatuoGPT-II (7B)	0.308	0.131	0.351	0.179	0.394	0.274	0.107	0.049
GLM-4-9B-Chat (freeze-tuning)	0.315	0.129	0.345	0.182	0.386	0.269	0.112	0.054
QingNangTCM	0.368	0.157	0.299	0.218	0.425	0.298	0.137	0.064

overall performance, achieving the highest scores across most evaluated indicators. In terms of coverage and consistency, QingNangTCM attained the highest scores for ROUGE-1 (0.368) and ROUGE-2 (0.157), as well as the highest METEOR score (0.218) among all models. Regarding lexical accuracy, it consistently outperformed all baseline models across all *n*-gram orders, with BLEU-1 to BLEU-4 scores of 0.425, 0.298, 0.137, and 0.064, respectively. Conversely, for the ROUGE-L metric, which emphasizes the longest common subsequence, the domain-specific baseline model HuatuoGPT-II (7B) achieved the highest score (0.351), followed by the GLM-4-9B-Chat (freeze-tuning) (0.345), while QingNangTCM recorded 0.299. Among the baseline models, the freeze-tuned variant exhibited competitive performance on METEOR (0.182) and BLEU-4 (0.054), whereas the general-purpose DeepSeek-V2 model achieved a relatively higher BLEU-4 (0.051) compared with GLM-4-9B-Chat (0.037).

To enable comparative analysis across different metrics, we applied a normalization strategy using the RobustScaler method. Figure 3 illustrates the normalized performance profile across consistency (METEOR), coverage (ROUGE), and accuracy (BLEU) dimensions. QingNangTCM exhibited a superior and balanced profile, achieving a uniformly high normalized score of 0.900 across all three dimensions, resulting in the largest enclosed area on the radar chart. In contrast, the domain-oriented models [HuatuoGPT-II (7B) and GLM-4-9B-Chat (freeze-tuning)] remained competitive in coverage (0.898 and 0.878, respectively) but demonstrated substantial deficits in consistency (0.478 and 0.504, respectively) and accuracy (0.700 and 0.600, respectively). General-purpose models recorded the lowest scores; DeepSeek-V2 achieved an accuracy of 0.400 but only 0.159 in consistency, while the base GLM-4-9B-Chat defined the lower performance bound, scoring approximately 0.100 across all metrics.

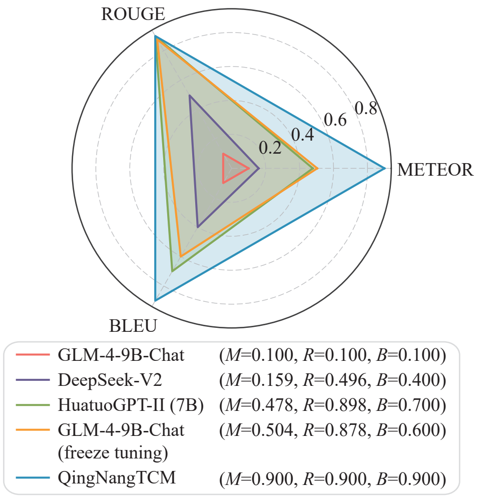


Figure 3 Radar chart comparing the normalized performance of QingNangTCM and baseline models

3.2 Assessment of professionalism, fluency, and safety

Figure 4 depicts the pairwise comparison of QingNangTCM with the baseline models across three dimensions: professionalism, fluency, and safety. The judgments shown were obtained using a hybrid evaluation strategy, wherein GPT-4o provided preliminary ratings that were subsequently validated by experts.

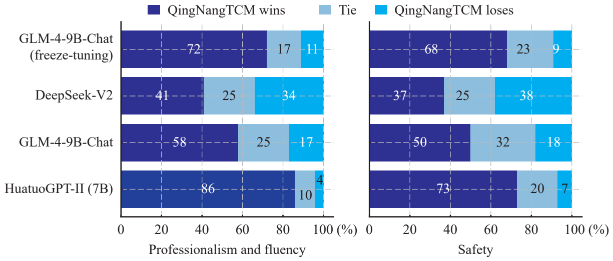


Figure 4 Pairwise evaluation of QingNangTCM against baseline models

(i) Professionalism and fluency. In these dimensions, QingNangTCM achieved higher win rates, particularly against domain-specific baseline models. It recorded a win rate of 86% versus HuatuoGPT-II (7B) and 72% versus the GLM-4-9B-Chat (freeze-tuning), with corresponding loss rates limited to 4% and 11%, respectively. In comparisons with general-purpose models, QingNangTCM achieved a win rate of 58% against the base GLM-4-9B-Chat and 41% against DeepSeek-V2.

(ii) Safety. Regarding the safety dimension, QingNangTCM exhibited a distinct advantage over domain-specific models, achieving a win rate of 73% against HuatuoGPT-II (7B) and 68% against the GLM-4-9B-Chat (freeze-tuning). Compared with the base GLM-4-9B-Chat, QingNangTCM recorded a 50% win rate and an 18% loss rate. The comparison with DeepSeek-V2 yielded a balanced outcome, with a win rate of 37% and a loss rate of 38%.

3.3 Ablation study results under different training configurations

To further evaluate the stability of QingNangTCM and verify the experimental design, we conducted an ablation study on the GLM-4-9B-Chat backbone. The results, summarized in Table 3, illustrate the performance variations across different training epochs, learning rates, and virtual token lengths.

(i) Training epochs. Compared with the baseline model (epoch = 1) with the epoch = 3 setting [both with learning rate (LR) = 1 × 10^{-4}], significant improvements were observed. The baseline model showed insufficient convergence (ROUGE-1 = 0.220), while three epochs of training increased ROUGE-1 to 0.331 and METEOR to 0.195. This indicates that a moderate number of iterations is essential for the model to adequately capture

Table 3 Ablation study results of QingNangTCM under different hyperparameter configurations

Hyperparameter setting	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Epoch = 1, LR = 1×10^{-4}	0.134	0.220	0.118	0.240	0.353	0.265	0.107	0.031
Epoch = 3, LR = 1×10^{-4}	0.195	0.331	0.143	0.267	0.398	0.289	0.117	0.056
Epoch = 4, LR = 1×10^{-5}	0.106	0.248	0.088	0.198	0.238	0.122	0.066	0.022
Epoch = 4, LR = 1×10^{-4}, tokens = 32	0.119	0.322	0.121	0.269	0.286	0.172	0.112	0.041
Epoch = 4, LR = 1×10^{-4}, tokens = 64	0.218	0.368	0.157	0.299	0.425	0.298	0.137	0.064

semantic patterns in the TCM corpora. However, extending training further to four epochs without appropriate hyperparameter tuning (e.g., reducing LR to 1×10^{-5}) resulted in degraded performance (ROUGE-1 dropped to 0.248), suggesting that continuing with a fixed learning rate may hinder optimization in later stages.

(ii) Virtual token length. Using a stable training configuration (epoch = 4, LR = 1×10^{-5}), we examined the impact of virtual token length. With num_virtual_tokens = 32, the model maintained decent structural matching but suffered a drop in METEOR score (0.119). Increasing the virtual tokens to 64 yielded the best overall performance (ROUGE-1 = 0.368, METEOR = 0.218, BLEU-4 = 0.064). These results suggest that a longer prefix length (64 tokens) provides greater capacity to encode task-specific TCM knowledge, thereby improving both precision and semantic coverage.

3.4 Qualitative results on simulated task-oriented clinical scenarios

We analyzed the generated responses of the four model categories across the clinical scenarios detailed in Supplementary Table S2 – S5.

(i) General-purpose models (DeepSeek-V2 and GLM-4-9B-Chat). In the symptom analysis (Supplementary Table S2) and disease treatment (Supplementary Table S3) tasks, these models generated lists of common herbs and general lifestyle advice. However, the outputs lacked specific TCM diagnostic terms such as Bianzheng or Bingji. When queried about Baizhi (Angelicae Dahuricae Radix) in the herb inquiry task (Supplementary Table S4), they provided basic indications but omitted critical details like Guijing (归经, meridian tropism) and contraindications specific to patient’s constitutions.

(ii) Domain-specific model [HuatuoGPT-II (7B)]. In the symptom analysis task, HuatuoGPT-II (7B) attributed the symptoms to “hormonal changes” and included Progesterone (a Western medication) in its recommendation list. Similarly, for the hypertension scenario (Supplementary Table S3), it suggested Losartan alongside TCM formulas. In the herb inquiry, the output included pharmacological descriptions related to anti-inflammatory effects based on modern medical terminology rather than TCM-specific attributes.

(iii) Comparative fine-tuning strategy (freeze-tuning). The GLM-4-9B-Chat (freeze-tuning) generated responses containing TCM-specific terminology. For instance, in the symptom analysis task, it recommended the formula Huanglian Jiedu Tang (黄连解毒汤). In the herb inquiry task, the model provided traditional efficacy descriptions but did not provide specific dosage modifications or tailored recommendations for different TCM syndromes.

(iv) QingNangTCM. In standard clinical tasks (Supplementary Table S2 – S4), QingNangTCM exhibited high medical accuracy and strict adherence to the TCM theoretical framework. For symptom analysis and disease treatment, the model accurately identified core TCM pathologies, such as “liver Qi stagnation transforming into fire” or “Yin deficiency with internal heat” for irritability, and “liver Yang rising” with “Yin deficiency with Yang excess” for hypertension. Correspondingly, it recommended targeted formulas, including Danzhi Xiaoyao San (丹栀逍遥散), Yangyin Qingxin Tang (养阴清心汤), Tianma Gouteng Yin (天麻钩藤饮), and Zhengan Xifeng Tang (镇肝熄风汤). Furthermore, the model provided personalized therapeutic principles and lifestyle guidance (e.g., dietary adjustments and emotional regulation), effectively embodying the clinical principles of Bianzheng Lunzhi (辨证论治, syndrome differentiation and treatment) and Zhiweibing (治未病, preventive care). Regarding the herb Baizhi (Angelicae Dahuricae Radix), QingNangTCM generated precise information covering properties, Guijing, and common uses. It also provided dosage instructions tailored to specific symptoms and detailed contraindication warnings, ensuring both completeness and safety in its clinical recommendations.

In the failure case analysis (Supplementary Table S5), the model exhibited specific limitations when handling rare entities and complex syndromes. For the inquiry of rare herb Tianmingjing (Carpesii Herba), the model incorrectly described its efficacy as “dispelling wind and relieving cough” (restricted to respiratory indications), which diverges from its documented TCM functions of clearing heat and cooling blood. Regarding Piyuezheng (脾约证, spleen bind syndrome), the model misclassified the pathology as general “spleen deficiency”, leading to recommendations focusing on “spleen-strengthening” that failed to address the syndrome’s core pathogenesis of fluid depletion and intestinal dryness.

4 Discussion

4.1 Bridging the TCM knowledge gap in foundation models via the construction of QnTCM_Dataset

Comparative analysis indicates that general-purpose large language models exhibit inherent limitations when applied to knowledge-intensive, vertical TCM tasks. This gap stems not from insufficient linguistic fluency, but from inadequate internalization of the structured knowledge schemas that underpin TCM reasoning. Consequently, general models tend to rely on surface-level language patterns, resulting in constrained semantic consistency and limited higher-order reasoning under domain-specific requirements. These limitations are particularly evident in tasks involving fine-grained TCM knowledge, such as Guijing and Peiwu Jinji (配伍禁忌, compatibility contraindications), where responses often remain generic and lack well-supported etiological analysis and therapeutic justification.

In contrast, the advantages observed in QingNangTCM can be attributed to targeted domain knowledge injection through the construction of QnTCM_Dataset rather than model scale or architectural differences. By introducing aligned concept–relation–expression signals, the dataset facilitates more effective organization of domain terminology and diagnostic reasoning pathways, thereby mitigating domain-specific knowledge sparsity. More broadly, this analysis suggests that adapting foundation models to TCM relies less on increasing model capacity and more on systematic corpus design and knowledge alignment, which are essential for bridging the gap between general-purpose language models and specialized medical reasoning tasks.

4.2 Reducing paradigm conflation and establishing clinical professionalism via hybrid validation

Comparison with HuatuoGPT-II (7B) demonstrates that surface-level text coverage alone cannot ensure clinical professionalism in TCM-oriented language models. Training on mixed biomedical corpora often blurs conceptual boundaries between traditional and modern medical systems, leading to paradigm conflation. This manifests as the inappropriate introduction of Western medications or modern pharmacological explanations into TCM contexts, thereby weakening syndrome-treatment coherence and reducing clinical specificity.

These limitations arise primarily from data composition and validation strategy rather than model capacity. When heterogeneous medical paradigms are jointly encoded without explicit constraints, models struggle to maintain a consistent diagnostic logic, particularly in safety-critical settings involving contraindications and therapeutic boundaries. This suggests that professionalism in TCM applications depends less on linguistic adequacy than on the integrity of domain-specific paradigms.

In contrast, QingNangTCM benefits from a hybrid expert-automated validation framework that enforces paradigm purity during data construction. By filtering paradigm-mismatched content and risk signals at the source, the model preserves syndrome-treatment correspondence and integrates safety awareness within a coherent TCM framework. More broadly, this analysis indicates that establishing clinical professionalism in medical LLMs requires explicit alignment between domain knowledge, validation criteria, and intended clinical paradigms, rather than post hoc correction at the output stage.

4.3 Achieving deeper semantic alignment via P-Tuning v2 deep prompt tuning

Comparison with the freeze-tuning variant highlights a fundamental limitation of shallow fine-tuning strategies in supporting deep semantic alignment for TCM reasoning. Updating only posterior layers may preserve surface-level textual relevance, but it fails to guide reasoning across the full representational hierarchy of the model. Consequently, such models tend to rely on term-level associations, resulting in limited consistency and weak syndrome-specific treatment adaptation, which constrains their ability to capture the logic of syndrome differentiation.

In contrast, QingNangTCM employs P-Tuning v2 to perform layer-wise continuous prompt optimization by injecting trainable prefixes into all Transformer layers. This design allows domain knowledge to influence the reasoning process throughout the model depth, rather than being appended at the output stage. As a result, the model can better internalize structured causal relationships in TCM reasoning, such as the progression from pathogenesis to treatment principles, instead of relying on shallow statistical co-occurrence.

Further analysis indicates that the effectiveness of P-Tuning v2 stems from sustained representational guidance rather than isolated hyperparameter choices. Stable deep semantic alignment requires sufficient prompt capacity combined with consistent optimization dynamics, underscoring deep prompt tuning as a more suitable paradigm for domain knowledge integration than shallow fine-tuning methods in complex medical reasoning tasks.

4.4 Limitations and future work

From a practical perspective, QingNangTCM is designed to support several common TCM-oriented tasks, including herbal medicine inquiry, symptom-oriented reasoning, and disease-oriented treatment support. In these settings, the model provides structured information on herbal properties and syndrome-related therapeutic principles, which may assist information retrieval and communication in routine use. The model's response patterns in the evaluated scenarios reflect an explicit syndrome differentiation-oriented reasoning structure and are relevant to educational and auxiliary decision-support contexts.

Several limitations of the present study should be noted. First, the qualitative analyses are conducted using simulated clinical scenarios rather than real-world patient cases, and thus the findings do not constitute clinical validation. Second, the failure case analysis indicates that the model can produce inaccurate or incomplete outputs when queried about rare medicinal materials or complex syndromes, highlighting challenges associated with long-tail knowledge coverage. Third, the model's outputs depend on user-provided inputs and should be interpreted as reference information rather than diagnostic or therapeutic decisions, underscoring the continued need for professional clinical judgment.

Future work will focus on expanding the diversity and coverage of training data, refining evaluation protocols under more varied and realistic settings, and investigating mechanisms to better constrain model behavior in rare or ambiguous cases. Additional directions include incorporating structured expert feedback, improving interpretability of reasoning traces, and exploring deployment scenarios that comply with relevant ethical and regulatory requirements.

5 Conclusion

This study presents QingNangTCM, a specialized LLM optimized for TCM through parameter-efficient fine-tuning. By synergizing the systematic construction of the high-fidelity QnTCM_Dataset with the deep semantic alignment enabled by P-Tuning v2, the model significantly surpasses both general-purpose baselines and existing domain-specific models across critical dimensions of accuracy, coverage, consistency, safety, professionalism, and fluency. Comprehensive evaluations, spanning quantitative metrics, hybrid expert-AI reviews, and simulated clinical scenarios, confirm QingNangTCM's capability to deliver accurate, logically coherent, and clinically reliable responses. It effectively empowers core clinical tasks, including syndrome differentiation, herbal inquiry, and treatment planning, serving as a trustworthy decision-making support tool. Ultimately, this work offers a robust theoretical and technical foundation for the digitalization, intelligent modernization, and international dissemination of TCM.

Fundings

Hebei Province Higher Education Scientific Research Project (QN2025367), Zhangjiakou City 2022 Municipal Science and Technology Plan Self-raised Fund Project (221105D), and Hebei Province Education Science "14th Five-Year Plan" Project (2404224).

Acknowledgements

We acknowledge the support of the Macao Polytechnic University under the submission code fca.33be.6385.b.

Author contributions

Xuming Tong: conceptualization, methodology, software, formal analysis, and writing – original draft. Liyan Liu: conceptualization, methodology, validation, and writing – original draft. Yanhong Yuan: data curation and investigation. Xiaozheng Ding: data curation and investigation. Huiru Jia: resources and validation. Yapeng Wang: supervision, project administration, and writing – review & editing. Xu Yang: formal analysis and visualization. Sio Kei Im: methodology and validation. Mini Han Wang: software and visualization. Zhang Xiong: writing – review & editing. All authors approved the submission and take responsibility for this manuscript.

Competing interests

The authors declare no conflict of interest.

References

[1] WANGSA K, KARIM S, GIDE E, et al. A systematic review and comprehensive analysis of pioneering AI chatbot models from education to healthcare: ChatGPT, Bard, Llama, Ernie and Grok. Future Internet, 2024, 16(7): 219.

[2] SUN TX, ZHANG XT, HE ZF, et al. MOSS: an open conversational large language model. Machine Intelligence Research, 2024, 21(5): 888–905.

[3] XIAO JF, CHEN YC, OU YM, et al. Baichuan2-sum: instruction finetune Baichuan2-7B model for dialogue summarization. 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024: 1-8.

[4] HUANG L, HU J, CAI Q, et al. The performance evaluation of artificial intelligence ERNIE bot in Chinese National Medical Licensing Examination. Postgraduate Medical Journal, 2024, 100(1190): 952–953.

[5] GIBNEY E. China's cheap, open AI model DeepSeek thrills scientists. Nature, 2025, 638(8049): 13–14.

[6] LIAN L, LUO X, CHIPUSU K, et al. Large language models evaluation of medical licensing examination using GPT-4.0, ERNIE Bot 4.0, and GPT-4o. Bioengineering, 2026, 13(1): 113.

[7] DAI YZ, SHAO X, ZHANG JL, et al. TCMChat: a generative large language model for traditional Chinese medicine. Pharmacological Research, 2024, 210: 107530.

[8] HUA R, DONG X, WEI Y, et al. Lingdan: enhancing encoding of

traditional Chinese medicine knowledge for clinical reasoning tasks with large language models. Journal of the American Medical Informatics Association, 2024, 31(9): 2019–2029.

[9] SHOOL S, ADIMI S, SABOORI AMLESHI R, et al. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Medical Informatics and Decision Making, 2025, 25(1): 117.

[10] LIU FL, ZHOU HJ, GU BY, et al. Application of large language models in medicine. Nature Reviews Bioengineering, 2025, 3(6): 445–464.

[11] XU L, XIE H, QIN SJ, et al. Parameter-efficient fine-tuning methods for pretrained language models: a critical review and assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026. doi: 10.1109/TPAMI.2026.3657354.

[12] WANG L, CHEN S, JIANG L, et al. Parameter-efficient fine-tuning in large language models: a survey of methodologies. Artificial Intelligence Review, 2025, 58(8): 227.

[13] XIONG H, WANG S, ZHU Y, et al. DoctorGLM: fine-tuning your Chinese doctor is not a herculean task. arXiv, 2023. doi: 10.48550/arXiv.2304.01097.

[14] ShenNong-TCM: a traditional Chinese medicine large language model. GitHub Repository, 2023. Avaiable from: https://github.com/michael-wzhu/ShenNong-TCM-LLM.

[15] TAN Y, ZHANG ZX, LI MC, et al. MedChatZH: a tuning LLM for traditional Chinese medicine consultations. Computers in Biology and Medicine, 2024, 172: 108290.

[16] CHEN J, WANG X, JI K, et al. HuatuoGPT-II: one-stage training for medical adaptation of large language models. arXiv, 2023. doi: 10.48550/arXiv.2311.09774.

[17] JIA YZ, JI XY, WANG X, et al. Qibo: a large language model for traditional Chinese medicine. Expert Systems with Applications, 2025, 284: 127672.

[18] YANG SH, ZHAO HJ, ZHU SB, et al. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 19368–19376.

[19] WU Y, ZHANG FL, YANG K, et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. Nucleic Acids Research, 2019, 47(D1): D1110–D1117.

[20] ARSLAN M, GHANEM H, MUNAWAR S, et al. A survey on RAG with LLMs. Procedia Computer Science, 2024, 246: 3781–3790.

[21] HU L, ZHANG X, SONG D, et al. Efficient and effective role player: a compact knowledge-grounded persona-based dialogue model enhanced by LLM distillation. ACM Transactions on Information Systems, 2025, 43(3): 1–29.

[22] LIU C, SUN KJ, ZHOU QQ, et al. CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. Scientific Reports, 2024, 14: 6403.

[23] CHU YH, CAO HL, DIAO YF, et al. Refined SBERT: representing sentence BERT in manifold space. Neurocomputing, 2023, 555: 126453.

[24] ZHENG L, CHIANG WL, SHENG Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems, 2023, 36: 46595–46623.

[25] GLM T, ZENG A, XU B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 with all tools. arXiv, 2024. doi: 10.48550/arXiv.2406.12793.

[26] DEEPSEEK-AI, LIU A, FENG B, et al. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model. arXiv, 2024. doi: 10.48550/arXiv.2405.04434.

[27] ZHENG Y, ZHANG R, ZHANG J, et al. LLaMAFactory: unified efficient fine-tuning of over 100 language models. arXiv, 2024. doi: 10.48550/arXiv.2403.13372.

[28] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP. International Conference on Machine Learning, 2019: 2790-2799.

[29] GHASSEMIAZGHANDI M. An evaluation of ChatGPT's translation accuracy using BLEU score. Theory and Practice in Language Studies, 2024, 14(4): 985–994.

[30] BRIMAN MKH, YILDIZ B. Beyond ROUGE: a comprehensive evaluation metric for abstractive summarization leveraging similarity, entailment, and acceptability. International Journal on Artificial Intelligence Tools, 2024, 33(5): 2450017.

[31] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 65–72.

[32] GU J, JIANG X, SHI Z, et al. A survey on LLM-as-a-Judge. The Innovation, 2025. doi: 10.1016/j.xinn.2025.101253.

(Editor-in-Charge Jie Deng)

QingNangTCM：一种面向中医领域的参数高效微调大语言模型

通旭明[a, b†], 刘利岩[b†], 袁艳红[c], 丁晓征[b], 贾慧茹[b], 杨旭[a], 严肇基[d], 王涵[e], 熊璋[f], 王雅鹏[a*]

a. 澳门理工大学应用科学学院, 澳门 999078, 中国
b. 河北北方学院信息科学与工程学院, 河北 张家口 075000, 中国
c. 河北北方学院教务处, 河北 张家口 075000, 中国
d. 澳门理工大学机器翻译暨人工智能应用技术教育部工程研究中心, 澳门 999078, 中国
e. 香港中文大学医学院, 香港 999077, 中国
f. 北京航空航天大学计算机学院, 北京 100191, 中国

【摘要】目的 针对通用大语言模型在中医专业问答与临床推理中存在领域知识、专业对齐程度有限等问题，构建一种面向中医应用场景的专用大语言模型 QingNangTCM。**方法** 构建了一个包含 10 万条样本的中医领域语料库 QnTCM_Dataset，该语料库在整合 ShenNong_TCM_Dataset 和 SymMap v2.0 的基础上，引入检索增强生成与角色驱动生成策略进行数据扩展，覆盖中医诊断问答、处方建议及中药知识等核心任务。以 GLM-4-9B-Chat 为基座模型，采用 P-Tuning v2 方法进行参数高效微调，得到 QingNangTCM 模型。本研究建立了多维评测体系，从准确性、覆盖性、一致性、安全性、专业性与流畅性等方面进行综合评估，采用双语评估替补（BLEU）、面向召回的摘要评估研究（ROUGE）、机器翻译评估的度量（METEOR）等自动指标，并结合基于专家校验的 LLM-as-a-Judge 评测方法。同时设计症状分析、疾病诊疗、中药查询和失败案例四类模拟临床场景开展定性分析，并与 GLM-4-9B-Chat、DeepSeek-V2、HuatuoGPT-II（7B）及 GLM-4-9B-Chat（freeze-tuning）模型进行对比。**结果** QingNangTCM 在 BLEU-1/2/3/4（0.425/0.298/0.137/0.064）、ROUGE-1/2（0.368/0.157）及 METEOR（0.218）指标上均取得最优表现，在准确性、覆盖性与一致性维度上的归一化综合性能达到 0.900。尽管其 ROUGE-L 指标（0.299）略低于 HuatuoGPT-II（7B）（0.351），但在专家验证的专业性与安全性胜率评估中分别达到 86% 和 73%。定性分析显示，该模型能够较好遵循“症状–证候–病机–治法”的中医诊疗推理链条，但在处理罕见中药及复杂证候组合时仍存在一定误判与幻觉现象。**结论** 通过将中医领域语料构建与参数高效的提示微调方法相结合，可增强大语言模型在中医相关任务中的推理与领域适配能力。相关工作为中医知识的数字化与智能化提供了一种技术框架，对辅助中医诊疗与教育具有一定的应用价值。

【关键词】大语言模型；中医；微调；P-Tuning v2；临床决策支持