



## Clinical decision and prescription generation for diarrhea in traditional Chinese medicine based on large language model

Jiaze Wu<sup>a, b†</sup>, Hao Liang<sup>c†</sup>, Haoran Dai<sup>b</sup>, Hongliang Rui<sup>b, d\*</sup>, Baoli Liu<sup>b, e\*</sup>

*a. School of Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100029, China*

*b. Beijing Hospital of Traditional Chinese Medicine, Capital Medical University, Beijing 100010, China*

*c. Academy of Chinese Medical Sciences, Hunan University of Chinese Medicine, Changsha, Hunan 410208, China*

*d. Laboratory Animal Room, Beijing Institute of Traditional Chinese Medicine, Beijing 100010, China*

*e. School of Traditional Chinese Medicine, Capital Medical University, Beijing 100069, China*

### ARTICLE INFO

#### Article history

Received 23 October 2025

Accepted 18 January 2026

Available online 15 March 2026

#### Keywords

Diarrhea

Traditional Chinese medicine

Large language model

Clinical decision and prescription generation

Natural language processing

### ABSTRACT

**Objective** To develop a clinical decision and prescription generation system (CDPGS) specifically for diarrhea in traditional Chinese medicine (TCM), utilizing a specialized large language model (LLM), Qwen-TCM-Dia, to standardize diagnostic processes and prescription generation.

**Methods** Two primary datasets were constructed: an evaluation benchmark and a fine-tuning dataset consisting of fundamental diarrhea knowledge, medical records, and chain-of-thought (CoT) reasoning datasets. After an initial evaluation of 16 open-source LLMs across inference time, accuracy, and output quality, Qwen2.5 was selected as the base model due to its superior overall performance. We then employed a two-stage low-rank adaptation (LoRA) fine-tuning strategy, integrating continued pre-training on domain-specific knowledge with instruction fine-tuning using CoT-enriched medical records. This approach was designed to embed the clinical logic (symptoms → pathogenesis → therapeutic principles → prescriptions) into the model's reasoning capabilities. The resulting fine-tuned model, specialized for TCM diarrhea, was designated as Qwen-TCM-Dia. Model performance was evaluated for disease diagnosis and syndrome type differentiation using accuracy, precision, recall, and F1-score. Furthermore, the quality of the generated prescriptions was compared with that of established open-source TCM LLMs.

**Results** Qwen-TCM-Dia achieved peak performance compared to both the base Qwen2.5 model and five other open-source TCM LLMs. It achieved 97.05% accuracy and 91.48% F1-score in disease diagnosis, and 74.54% accuracy and 74.21% F1-score in syndrome type differentiation. Compared with existing open-source TCM LLMs (BianCang, HuangDi, LingDan, TCMLLM-PR, and ZhongJing), Qwen-TCM-Dia exhibited higher fidelity in reconstructing the “symptoms → pathogenesis → therapeutic principles → prescriptions” logic chain. It provided complete prescriptions, whereas other models often omitted dosages or generated mismatched prescriptions.

**Conclusion** By integrating continued pre-training, CoT reasoning, and a two-stage fine-tuning strategy, this study establishes a CDPGS for diarrhea in TCM. The results demonstrate the

†The authors contributed equally.

\*Corresponding author: Baoli Liu, E-mail: liubaoli@bjzhongyi.com. Hongliang Rui, E-mail: ruihongliang@bjzhongyi.com.

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

DOI: 10.1016/j.dcmcd.2026.02.003

**Citation:** WU JZ, LIANG H, DAI HR, et al. Clinical decision and prescription generation for diarrhea in traditional Chinese medicine based on large language model. Digital Chinese Medicine, 2026, 9(1): 13-30.

Copyright © 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

synergistic effect of strengthening domain representation through pre-training and activating logical reasoning via CoT. This research not only provides critical technical support for the standardized diagnosis and treatment of diarrhea but also offers a scalable paradigm for the digital inheritance of expert TCM experience and the intelligent transformation of TCM.

## 1 Introduction

Traditional Chinese medicine (TCM), with a history spanning over 2 000 years, is a valuable part of China's cultural heritage. Syndrome differentiation and treatment (SDT), the core principle of TCM, has been consistently practiced across generations of clinicians and remains profoundly significant in modern clinical practice. It remains the cornerstone of TCM practice [1]. However, its efficacy is inherently subjective and relies heavily on a practitioner's expertise. Consequently, national policies from the General Office of the State Council [2,3] are mandating the development of intelligent assisted diagnosis systems to standardize and modernize TCM.

While recent advancements in large language models (LLMs) have shown promise in digitizing TCM knowledge [4-6], current approaches are often limited to knowledge graph integration and general clinical question-answering. These approaches often prioritize surface-level textual alignment over the underlying pathological logic, thereby failing to capture the multi-dimensional, non-linear correlations between heterogeneous clinical symptoms and their corresponding syndrome differentiation. Such a deficiency hinders the establishment of robust semantic mappings across diverse TCM data, ultimately leading to a bottleneck in executing complex diagnostic reasoning and formulating personalized therapeutic strategies that adhere to the holistic principles of TCM. Diarrhea remains a major global public health challenge, accounting for a significant proportion of outpatient encounters. According to the Global Burden of Disease Study 2021, diarrheal diseases remain a major global health threat, contributing to substantial morbidity and an estimated 59.0 million disability-adjusted life-years (DALYs) annually [7]. In primary care settings, gastrointestinal complaints, dominated by acute diarrhea, continue to be among the top reasons for patient visits [8]. Its clinical complexity stems from a diverse etiology involving the dysfunction of the spleen, stomach, and liver. In this domain, subtle variations in stool consistency, frequency, and accompanying abdominal sensations necessitate a rigorous distinction between cold, heat, deficiency, and excess syndromes.

To bridge the gap between raw clinical data and expert-level diagnosis, we propose a clinical decision and prescription generation system (CDPGS) powered by LLMs for the management of TCM diarrhea. Grounded in the theory of mimicking cognitive paradigms of TCM

experts, our framework employs a two-stage training strategy that integrates continued pre-training on foundational knowledge of diarrhea with instruction fine-tuning via structured medical records using chain-of-thought (CoT) reasoning (symptoms → pathogenesis → therapeutic principles → prescriptions). By evaluating our approach against nine model variants and existing open-source LLMs for TCM, we demonstrate that combining domain-specific pre-training with CoT logic produces clinically interpretable, standardized treatment plans, offering a scalable paradigm for the digital inheritance of TCM expertise.

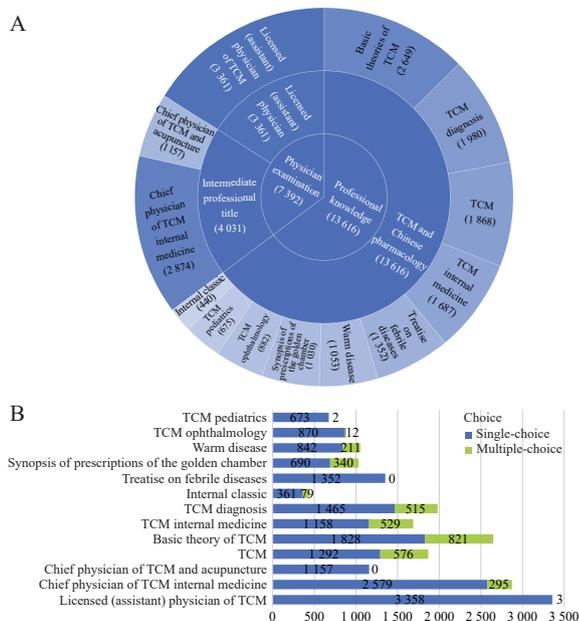
## 2 Data and methods

To develop and rigorously evaluate the CDPGS, we constructed two datasets. A benchmark dataset is used to assess the performance of base models. A fine-tuning dataset is dedicated to the domain adaptation and instruction tuning of the base model.

### 2.1 Data collection

**2.1.1 Benchmark dataset** A systematic search for publicly available TCM benchmark datasets was conducted. Three candidates were identified: Traditional Chinese Medicine QA Dataset (TCMD) [9], Traditional Chinese Medicine Benchmark (TCMBench) [10], and Comprehensive Medical Benchmark (CMB) [11]. Since TCMD and TCMBench are not publicly accessible, we selected the CMB dataset as our evaluation standard. The CMB dataset is structured with 6 major and 28 minor categories, comprising both single-choice and multiple-choice questions. To ensure relevance to our specific application, we extracted only the TCM-specific sections of CMB, resulting in a final benchmark collection of 21 008 questions from a total of 269 639 (Figure 1).

**2.1.2 Fine-tuning dataset** The use of open-source datasets offers significant advantages by substantially reducing the time required for data collection and annotation. Furthermore, these resources are often preprocessed, ensuring the acquisition of high-quality data. The complementary nature inherent in multi-source datasets is also instrumental in enhancing the generalization capability of LLMs. Consequently, as detailed in Table 1, we aggregated diverse open-source datasets that contain both foundational knowledge on diarrhea and medical records.



**Figure 1** TCM-specific section of the CMB dataset

A, the composition of TCM-specific sections in CMB. B, distribution of single-choice and multiple-choice questions of TCM-specific sections in CMB.

**Table 1** Summary of open-source TCM foundational knowledge and medical records datasets for fine-tuning

Dataset	Number of samples
ShenNong_TCM_Dataset [12]	112 565
hwTCM-sft-v1 [13]	7 096
TCM Consultation Dataset [14]	106 953
SFT_medicalKnowledge_source1 [15]	548 404
SFT_medicalKnowledge_source2 [15]	99 334
SFT_medicalKnowledge_source3 [15]	556 540
SFT_nlpDiseaseDiagnosed [15]	61 486
SFT_nlpSyndromeDiagnosed [15]	48 665
SFT_structGeneral [15]	310 860
SFT_structPrescription [15]	92 896

To capture the nuanced clinical reasoning required for diarrhea, we supplemented these open-source resources with targeted clinical medical records, which are critical for refining the LLM’s proficiency in medical terminology, diagnostic logic, and treatment regimens. These records were curated from the Ancient and Modern Medical Records Cloud Platform [https://www.yiankb.com], using the keywords “Xiexie (泄泻)” “Xiali (下利, dysentery)” and “diarrhea” [16] and the anonymized clinical archives for diarrhea of a distinguished TCM clinician, Professor Ming Chen from Beijing University of Chinese Medicine, spanning from March 28, 2019 to December 28, 2024.

The final fine-tuning dataset comprises a diarrhea foundational knowledge dataset and a diarrhea medical record dataset. To ensure data integrity, we implemented rigorous preprocessing protocols to extract core content

and eliminate noise. Furthermore, relative to large-scale pre-training datasets, the volume of domain-specific data is often limited. This limitation often leads to overfitting during fine-tuning, in which the acquisition of new domain knowledge compromises the model’s original general language representation capabilities [17]. To mitigate this problem, we selected the open-source BELLE dataset [18] as the general-domain dataset and integrated it into the training set. The mixing ratio was treated as a hyper-parameter to balance domain-specific and general capabilities. This approach facilitates a smooth transition from general linguistic representation to domain-specific embedding.

## 2.2 Data preprocessing

**2.2.1 Quality filtering** Quality filtering represents a critical procedure for guaranteeing the integrity and relevance of the training data. The primary objective of this step was to exclude data that could compromise model performance. For the diarrhea foundational knowledge dataset, we strictly filtered for diarrhea-specific content, excluding all other diseases. For the diarrhea medical record dataset, cases were included only if the chief complaint explicitly identified diarrhea. Conversely, cases were excluded if they contained incomplete clinical information (e.g., missing chief complaints, symptoms, tongue/pulse diagnostics, or prescriptions) or if the treatment involved non-herbal modalities (such as acupuncture or massage).

**2.2.2 Quality deduplication** Data deduplication is a crucial preprocessing step to remove redundant instances, as repeated patterns in training data can lead LLMs to overfit, thereby impairing generalization. To mitigate this, we applied the MinHash-Locality Sensitive Hashing (LSH) algorithm for efficient duplicate detection. Following established protocols [19], we set the similarity threshold to 0.95. This conservative deduplication approach eliminated redundant records while preserving the semantic diversity and distributional integrity of the dataset.

## 2.3 Data annotation

**2.3.1 Instruction construction** Raw, unannotated medical records are often ill-suited for supervised fine-tuning because they lack explicit task definitions and target outputs. To bridge this gap, we structured the data into instruction-response pairs (formatted as {"instruction": "", "input": "", "output": ""}). Such datasets typically consist of three elements: an instruction that defines the specific task or command, an input that provides the necessary context or data for the task, and an output representing the target response. However, manual annotation at scale is prohibitively costly and time-consuming. Therefore, we

adopted the self-instruct method [20], which leverages an LLM's generative capabilities to synthesize large-scale instruction data iteratively. Specifically, we employed the high-performing LLM Qwen2.5-Max, using medical records as input and carefully crafted prompts to generate diagnoses, syndrome types, therapeutic principles, and prescriptions. To ensure the quality of the instruction-tuning dataset, a manual review process was implemented. The outputs were rigorously evaluated and refined by three TCM experts, each with over five years of clinical experience. The review criteria focused on three dimensions: (i) clinical accuracy, ensuring diagnoses and prescriptions align with medical practice; (ii) logical consistency between syndromes, therapeutic principles, and medications; and (iii) terminological standardization. This expert-led refinement yielded a high-quality dataset for instruction tuning (Diarrhea Medical Record Dataset).

**2.3.2 Chain-of-thought (CoT) integration** CoT reasoning involves explicitly generating intermediate inference steps before producing the final output of LLMs. By decomposing complex tasks into sequential sub-steps, CoT improves the LLM's reasoning accuracy, especially in challenging problem-solving contexts [21]. To evaluate CoT's utility in LLM-assisted diagnosis and treatment of diarrhea, we constructed an instruction dataset in the format <question, CoT, answer>, replacing the conventional <question, answer> structure used in standard fine-tuning. The components are defined as follows. (i) Question: comprehensive TCM clinical records, including demographic data (gender and age), chief complaints, history of present illness, and tongue and pulse manifestations; (ii) CoT: key symptoms → pathogenesis analysis → TCM diagnosis → SDT. This path is designed to simulate a clinician's diagnostic logic; (iii) answer: the final decision-making output, consisting of the specific prescription (formula name and herbal composition) along with dosage and administration instructions. Guided by the "diarrhea" chapter in *Traditional Chinese Medicine Internal Medicine* [22], we categorized syndromes into six types: Hanshi Neisheng (寒湿内盛, internal exuberance of cold-dampness), Shire Shangzhong (湿热伤中, damp-heat injuring the middle Jiao), Shizhi Changwei (食滞肠胃, food stagnation in the gastrointestinal tract), Ganqi Chengpi (肝气乘脾, liver Qi invading the spleen), Piwei Xuruo (脾胃虚弱, spleen and stomach deficiency), and Shenyang Xushuai (肾阳虚衰, kidney Yang depletion and failure). Using the high-performing Qwen2.5-Max, we generated

reasoning traces and answers via application programming interface (API) calls with a one-shot prompting strategy. As with the instruction dataset, all CoT traces were rigorously validated by our expert panel to ensure clinical accuracy, logical consistency, and terminological standardization, resulting in the Diarrhea Medical Record CoT Dataset.

## 2.4 Data partitioning

Upon completion of quality filtering, deduplication, and annotation protocols, we consolidated the datasets of externally sourced open-source data and internally constructed medical records into three subsets: the Diarrhea Foundational Knowledge Dataset, the Diarrhea Medical Record Dataset, and the Diarrhea Medical Record CoT Dataset. Detailed specifications for these final datasets are presented in Table 2.

## 2.5 Base model evaluation

The choice of base model critically influences fine-tuning performance. As shown in Table 3, recent open-source LLMs exhibit strong capabilities in knowledge comprehension and logical reasoning, making them suitable for a wide range of natural language processing (NLP) tasks. To identify the model best aligned with TCM requirements, we established a systematic evaluation framework. Considering computational constraints (e.g., GPU memory), we selected models with 4B – 9B parameters, all pretrained on multilingual corpora comprising trillions of tokens. This extensive pre-training endows these models with broad multilingual and multidomain knowledge, supported by large vocabularies that include common words, domain-specific terminology, rare terms, slang, and multilingual expressions. Additionally, context length is crucial: a larger window enables richer contextual understanding, which is essential for complex diagnostic reasoning. Notably, Qwen2.5 and LLaMA-3 support a context window of up to 131 072 tokens, significantly enhancing their capacity for accurate natural language understanding and generation.

To ensure a rigorous and reproducible evaluation, all LLMs listed in Table 3 were evaluated on the benchmark dataset using the same prompt. Model outputs were independently assessed against reference answers over three runs, with performance measured along three dimensions: inference time, accuracy (agreement with

**Table 2** Composition of fine-tuning datasets and their functional allocation across the two-stage training strategy

Dataset	Number of samples after preprocessing	Training stage
Diarrhea Foundational Knowledge	4 151	The first stage/the second stage (continued pre-training)
Diarrhea Medical Record	4 086	The second stage
Diarrhea Medical Record CoT	632	The second stage (CoT)

**Table 3** Key technical specifications of the 16 representative general-purpose open-source LLMs selected for base model evaluation

Model	Language	Parameter	Training token	Context token	Vocabulary
Falcon3 <sup>[23]</sup>	Multilingual	7B	14T	32 768	131 072
Qwen2.5 <sup>[24]</sup>	Multilingual	7B	18T	131 072	152 064
MiniCPM3 <sup>[25]</sup>	Chinese, English	4B	Unknown	32 768	73 448
Mistral <sup>[26]</sup>	Multilingual	7B	Unknown	32 768	32 768
LLaMA-3 <sup>[27]</sup>	Multilingual	8B	15T	131 072	128 256
GLM-4 <sup>[28]</sup>	Multilingual	9B	10T	128 000	151 552
InternLM-2.5 <sup>[29]</sup>	Chinese, English	7B	Unknown	32 768	92 544
Gemma-2 <sup>[30]</sup>	Chinese, English	9B	8T	8 192	256 000
Mengzi3 <sup>[31]</sup>	Multilingual	8B	3T	8 192	128 256
MAP-NEO <sup>[32]</sup>	Chinese, English	7B	4.5T	4 096	64 256
Chinese-LLaMA-Alpaca-3 <sup>[33]</sup>	Chinese, English	8B	15T	8 192	128 256
Yi-1.5 <sup>[34]</sup>	Chinese, English	9B	0.5T	4 096	64 000
Baichuan-2 <sup>[35]</sup>	Multilingual	7B	2.6T	4 096	125 696
XVERSE <sup>[36]</sup>	Multilingual	7B	2.6T	8 192	100 534
TigerBot <sup>[37]</sup>	Chinese, English	7B	Unknown	2 048	60 928
YAYI2 <sup>[38]</sup>	Multilingual	7B	2T	4 096	32 005

reference answers), and output quality (adherence to prompt instructions and question format constraints). To enhance efficiency, all models were quantized via one of the following methods: GPT-generated unified format (GGUF) <sup>[39]</sup>, GPT quantization (GPTQ) <sup>[40]</sup>, activation-aware weight quantization (AWQ) <sup>[41]</sup>, or BitsandBytes <sup>[42]</sup>, and deployed using llama.cpp <sup>[43]</sup>, vLLM <sup>[44]</sup>, or the Hugging Face Transformers library <sup>[45]</sup>. For fairness and controllability, inference hyperparameters were fixed across all models: temperature = 0.7, top\_p = 0.7, top\_k = 50, and repetition\_penalty = 1.0.

## 2.6 Model fine-tuning

**2.6.1 Fine-tuning method** Fine-tuning refers to the additional training of a pre-trained model on domain-specific data to improve its performance on specific tasks or domains <sup>[46]</sup>. Its primary goal is to align task-specific knowledge with instructions, thereby enhancing model generalization and utility while reducing training costs <sup>[47]</sup>. Fine-tuning approaches are generally divided into two categories based on parameter update strategies: full parameter fine-tuning (FPFT), which updates all model parameters, and parameter-efficient fine-tuning (PEFT), which modifies only a subset of parameters. Due to the high computational demands of FPFT, especially for LLMs, PEFT methods offer a more efficient alternative. Related study that evaluated FPFT against multiple PEFT techniques across over 100 NLP tasks shows that low-rank adaptation (LoRA) achieves comparable performance with substantially lower computational cost <sup>[48]</sup>. Consequently, LoRA was selected for this study. Specifically, the LLaMA-Factory framework <sup>[49]</sup> was used to fine-tune the

base model. The instruction dataset was split into training and test sets at a 9 : 1 ratio. To emulate real-world clinical conditions and assess diagnostic accuracy and misdiagnosis rates, diarrhea-specific medical records were intentionally combined with records of other diseases at a 1 : 5 ratio. A curriculum learning strategy was implemented in two sequential stages: the first emphasized foundational knowledge of diarrhea, while the second focused on complex diarrhea medical records to facilitate progression from foundational understanding to advanced clinical reasoning.

**2.6.2 Fine-tuning stages and experimental metrics** (i) The first stage. The first stage of fine-tuning focuses on investigating the impact of critical hyperparameters on model performance. These include LoRA parameters (rank, alpha, and target modules) and training parameters (learning rate, total batch size, number of epochs, and the ratio of domain-specific to general-domain datasets). The experimental design employs a sequential tuning strategy: first, controlling single variables independently, and subsequently performing combined optimization to determine the optimal hyperparameter configuration. To stabilize the learning process and ensure convergence, the Cosine-Warmup learning rate scheduling strategy is adopted <sup>[50]</sup>. Furthermore, the total batch size is defined as the product of the single-step batch size and the number of gradient accumulation steps. Gradient accumulation is used to simulate the training effect of a larger batch under limited video memory, thereby mitigating the potential instability issues commonly associated with small-batch training.

Systematic evaluation of text generation quality during the fine-tuning process is essential for validating

model performance and guiding subsequent optimization. Commonly employed automatic evaluation metrics in this context include recall-oriented understudy for gisting evaluation (ROUGE) [51], bilingual evaluation understudy (BLEU) [52], and metric for evaluation of translation with explicit ordering (METEOR) [53].

ROUGE is a family of recall-based text similarity metrics used to quantify the quality of a generated text. It functions by calculating the number of overlapping lexical units (words or sequences) between the generated output and a set of reference texts. The primary variants utilized are ROUGE-N and ROUGE-L. ROUGE-N measures the degree of  $n$ -gram overlap between the generated and reference texts. ROUGE-L evaluates text similarity based on the longest common subsequence.

BLEU quantifies text quality by measuring the degree of  $n$ -gram overlap between the generated text and a set of reference texts. Specifically, it assesses the precision of the generated output against the references. The metric typically calculates the geometric mean of the modified  $n$ -gram precision scores (e.g.,  $n = 1, 2, 3, 4$ ), coupled with a brevity penalty to penalize overly short outputs.

METEOR provides a more nuanced evaluation of consistency between the generated and reference texts. It achieves this by comprehensively integrating multiple factors, including exact word matching (unigram precision and recall), word-order consistency (measures such as chunking), and semantic relevance (through stemming and synonym matching).

(ii) The second stage. The first stage determined the optimal hyperparameter combination, yielding a fine-tuned model that improved its domain understanding of diarrhea compared to the base model. However, as the overarching objective of this study is to construct a CDPGS, a second stage of fine-tuning is imperative. This stage is designed to explore the impact of structured medical records for diarrhea, aiming to enable the model to automatically generate accurate and comprehensive treatment plans for given medical records, encompassing diagnoses, syndrome types, therapeutic principles,

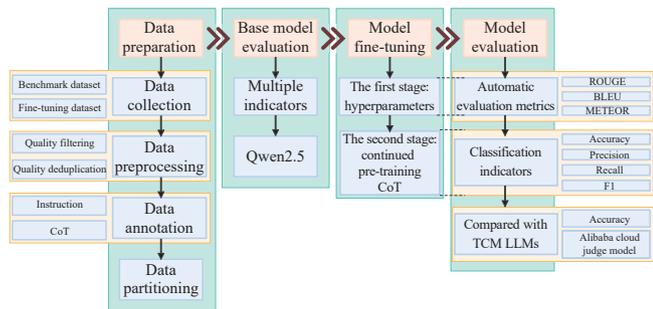
and prescriptions. Drawing on existing research [54], which indicates that knowledge acquisition in LLMs primarily stems from the pre-training stage, while instruction fine-tuning focuses on output format standardization, the use of continued pre-training was deemed necessary to effectively inject domain-specific knowledge. Based on three key variables, namely the sequence of the data curriculum, the inclusion of continued pre-training, and the integration of CoT reasoning, a total of nine fine-tuning models were established for comparative analysis in this second stage (Table 4).

The model's performance is comprehensively assessed using two classification tasks. First, disease diagnosis is structured as a binary classification task, where the positive class is defined as "diagnosed with diarrhea" and the negative class as "not diagnosed with diarrhea". Second, syndrome diagnosis is approached as a multi-class classification task, encompassing six categories of TCM syndromes: Hanshi Neisheng, Shire Shangzhong, Shizhi Changwei, Ganqi Chengpi, Piwei Xuruo, and Shenyang Xushuai. To mitigate the potential bias inherent in relying solely on accuracy, particularly when dealing with imbalanced data distributions, we employed four metrics derived from the confusion matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These foundational counts subsequently enable the calculation of comprehensive derived metrics, namely accuracy, precision, recall, and F1-score, which facilitate a multi-dimensional assessment of model performance. Furthermore, to provide a robust and representative overall measure when the class distribution is significantly unbalanced, the weighted average of accuracy, precision, recall, and F1-score is computed. To ensure robust model performance evaluation, we adopted a 10-fold cross-validation strategy during fine-tuning.

To systematically build and evaluate CDPGS, we designed a two-stage technical framework (Figure 2), consisting of four main stages: data preparation, base model evaluation, model fine-tuning, and model evaluation.

**Table 4** Overview of model variants for the second-stage fine-tuning based on data curriculum, continued pre-training, and CoT reasoning

Model	Description
Base	Qwen2.5-7B
Base-K	Base model fine-tuned on the Diarrhea Foundational Knowledge Dataset
Base-R	Base model fine-tuned on the Diarrhea Medical Record Dataset
Base-R_CoT	Base model fine-tuned on the Diarrhea Medical Record CoT Dataset
Base-K-R	Base-K model fine-tuned on the Diarrhea Medical Record Dataset
Base-K-R_CoT	Base-K model fine-tuned on the Diarrhea Medical Record CoT Dataset
Base-K_CP	Base model with continued pre-training on the Diarrhea Foundational Knowledge Dataset
Base-K_CP-R	Base-K_CP model fine-tuned on the Diarrhea Medical Record Dataset
Base-K_CP-R_CoT	Base-K_CP model fine-tuned on the Diarrhea Medical Record CoT Dataset



**Figure 2** Overall research design and technical roadmap

### 2.7 Statistical analysis

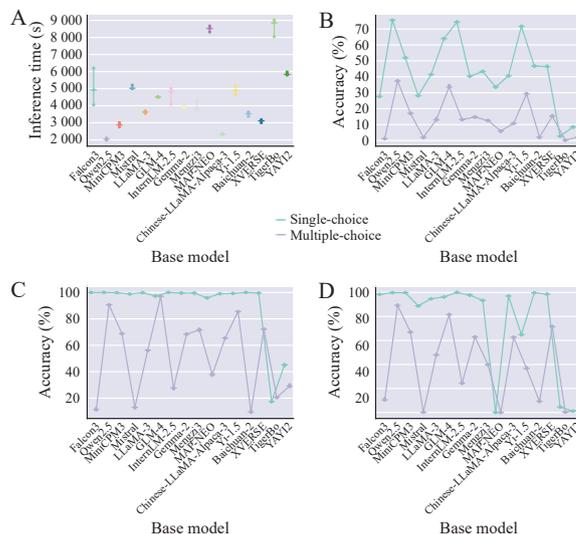
Statistical analyses were performed using Python 3.12. Based on the normality test, all experimental results were reported as mean ± standard deviation (SD) or median (interquartile range, IQR). To assess statistical significance, we used the Kruskal-Wallis test, Welch’s analysis of variance (ANOVA), one-way ANOVA, one-way repeated-measures ANOVA, or the Friedman test, and employed the corresponding post hoc tests.  $P < 0.05$  was considered statistically significant.

## 3 Results

### 3.1 Base model evaluation and selection

Figure 3A illustrates the inference time for base models, revealing substantial disparities across different architectures. Specifically, Qwen2.5-7B and TigerBot-7B, despite sharing the same 7B parameter size, exhibited an over fourfold difference in average inference time (2 040 s versus 8 635 s, respectively). The other 7B models, such as Falcon3 and XVERSE, averaged approximately 1 h. Mistral and YAYI2 required approximately 1.5 h on average. These results underscore the profound impact of model architecture and optimization strategy on inference efficiency. In terms of stability, most models demonstrated close consistency across the three evaluation runs, with notable exceptions, such as Falcon3, InternLM-2.5, and TigerBot, which exhibited significant fluctuations in inference time.

Figure 3B illustrates the accuracy of the base models on single- and multiple-choice questions, showing a significant range in performance. Qwen2.5 achieved the highest accuracy at 75.52%/37.52%, while TigerBot recorded the lowest at 2.73%/0.21%. Based on this accuracy profile, the 16 evaluated models can be broadly categorized into four groups: the high-accuracy group comprises Qwen2.5, InternLM-2.5, Yi-1.5, and GLM-4; the medium-accuracy group includes MiniCPM3, Baichuan-2, XVERSE, Mengzi3, LLaMA-3, Chinese-LLaMA-Alpaca-3, and Gemma-2; the low-accuracy group consists of Falcon3, Mistral, and MAP-NEO; the extremely low-accuracy group was limited to TigerBot and YAYI2.



**Figure 3** Multi-dimensional comparative evaluation of 16 open-source LLMs on the TCM benchmark

A, inference time. B, accuracy (task performance). C accuracy (format conformance). D, accuracy (content purity). For all plots, individual data points are shown with the horizontal line representing the median ( $n = 3$ ). Statistical significance was determined using the Kruskal-Wallis test (A, C, D) or one-way ANOVA (B, C).

Generally, the accuracy across all models on multiple-choice questions was significantly lower than on single-choice questions, underscoring the greater demands placed on their comprehensive reading and complex reasoning capabilities. Notably, the two top-performing models (Qwen2.5 and InternLM-2.5) both possess a parameter size of 7B, surpassing several models in the 8B and 9B range. This trend is particularly evident in MiniCPM3 (4B), whose accuracy exceeds that of multiple 7B – 9B models. All models demonstrated high stability, with consistent outputs across evaluation runs and minimal variation.

Output quality was introduced as a metric to assess the LLM’s adherence to specific prompt instructions during answer generation. This quality was measured using two criteria. First, we evaluated format conformance by ensuring the number of options in the generated output matched the question’s requirements (e.g., single- or multiple-choice). Outputs that satisfied this numerical criterion were marked as correct (Figure 3C). Second, we assessed content purity, confirming that the output contained only the selected options, excluding extraneous content such as internal reasoning or descriptive explanations. Outputs were marked correct only if they adhered to this purity constraint (Figure 3D).

Figure 3C illustrates conformance to the format across different models. Overall, with the notable exception of TigerBot and YAYI2, LLMs generally produce high-quality outputs on single-choice questions. Conversely, multiple-choice questions, which necessitate the simultaneous identification of multiple correct options, pose a

significantly greater challenge to the models' reasoning and instruction-following capabilities. InternLM-2.5, for example, demonstrates a perfect 100% format conformance on single-choice tasks, yet this drops drastically to only 27.51% on multiple-choice questions. This disparity strongly suggests a bias toward interpreting instruction or a preference for strategies specific to single-choice formats. Similar conformance issues were also observed in Baichuan-2, Falcon3, and Mistral. Qwen2.5, GLM-4, and Yi-1.5 performed relatively well in both scenarios, indicating strong adaptability and robustness when handling complex output constraints. Models with moderate performance include XVERSE, MiniCPM3, Mengzi3, Gemma-2, Chinese-LLaMA-Alpaca-3, and LLaMA-3. While these models perform acceptably on single-choice questions, their performance on multiple-choice tasks clearly requires further optimization. Finally, the poorest performers, TigerBot and YAYI2, suffered from severe output quality issues.

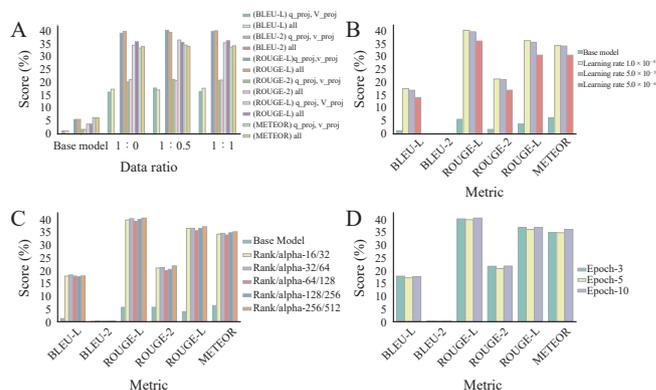
Figure 3D details the content purity assessment results. All models exhibited varying degrees of accuracy decline, indicating a general failure to produce outputs that meet purity requirements. MAP-NEO's accuracy approached zero for both single- and multiple-choice questions; output analysis revealed excessive explanatory text for nearly every response. Mistral, Mengzi3, and Yi-1.5 also suffered significant drops, as Mistral's multiple-choice accuracy neared zero due to verbose option-by-option explanations. Yi-1.5 and Mengzi3 similarly generated redundant content. TigerBot and YAYI2 performed poorly again: TigerBot often produced blank responses or included special tokens [e.g., (INST)] and HTML tags (e.g., <br>). YAYI2 emitted special tokens (e.g., <YaYi>ail) and verbose explanations akin to MAP-NEO. In contrast, the high-performing group, consisting of Qwen2.5, GLM-4, Falcon3, MiniCPM3, LLaMA-3, Gemma-2, Chinese-LLaMA-Alpaca-3, Baichuan-2, and XVERSE, demonstrated superior content purity, with Qwen2.5 and GLM-4 achieving the highest accuracy.

The findings demonstrate that model performance is affected by numerous factors, including model architecture. Qwen2.5 exhibited the superior overall performance across all metrics and was consequently selected as the base model for subsequent fine-tuning to develop a CDPGS for diarrhea in TCM.

### 3.2 Two-stage fine-tuning and evaluation

**3.2.1 The first stage: hyperparameter optimization** Figure 4A shows the model's performance under fixed hyperparameters (rank = 8, alpha = 16, total batch size = 16, epochs = 3), evaluating the effects of data ratios and LoRA target modules. Fine-tuning improves BLEU, ROUGE, and METEOR scores. While individual metrics may peak at different points, the best overall performance is achieved at a 1 : 1 data ratio and when LoRA

target modules are set to all. Figure 4B examines the impact of learning rate on fine-tuning performance under consistent hyperparameters (target = all, epochs = 3, data ratio = 1 : 0), confirming that  $1.0 \times 10^{-4}$  yields optimal results. Figure 4C examines the effect of LoRA rank and alpha values on fine-tuning performance. Results show that increasing both parameters consistently improves model performance, consistent with theoretical expectations: higher rank and alpha provide greater parameter flexibility and representational capacity, thereby enhancing the model's ability to capture complex features and adapt to domain-specific data distributions. The preceding analysis demonstrates that optimal fine-tuning performance is achieved with a LoRA rank of 256, an alpha of 512, all modules targeted, a learning rate of  $1.0 \times 10^{-4}$ , and a specific ratio of general-domain and diarrhea-domain data. The remaining hyperparameter to optimize is the number of training epochs. As shown in Figure 4D, model performance was evaluated at a data ratio of 1 : 1; peak performance occurred at three epochs. The complete set of optimal hyperparameters is summarized in Table 5.



**Figure 4** The performance of fine-tuned models with different hyperparameters

A, data ratios and LoRA target modules. B, learning rates. C, ranks and alphas of LoRA. D, epochs. In some metrics, bars for specific classes are present but not visible because their values are minimal.

**Table 5** Optimal hyperparameters for the first fine-tuning stage

Hyperparameter	Candidate parameter	Optimal parameter
Data ratio	1 : 0, 1 : 0.5, 1 : 1	1 : 1
Learning rate	$1.0 \times 10^{-3}$ , $1.0 \times 10^{-4}$ , $5.0 \times 10^{-5}$ , $5.0 \times 10^{-6}$	$1.0 \times 10^{-4}$
Rank/alpha	16/32, 32/64, 64/128/128/256, 256/512	256/512
Target module	q_proj, v_proj, all	All
Epoch	3, 5, 10	3

**3.2.2 The second stage: diagnostic efficacy and TCM reasoning** We evaluated the diagnostic performance of fine-tuned models on the test set of medical records for

diarrhea using confusion matrices (Figure 5) and classification metrics (Figure 6). Overall, all models achieved high accuracy in distinguishing non-diarrhea, yet showed variation in detecting diarrhea. The Base showed the highest false negatives, indicating the lowest sensitivity to diarrhea. In contrast, the Base-R\_CoT, Base-K-R, Base-K\_R\_CoT, Base-K\_CP-R, and Base-K\_CP-R\_CoT significantly improved TP rates while minimizing FN, demonstrating that two-stage fine-tuning combined with continued pre-training and CoT reasoning effectively enhances diagnostic sensitivity for diarrhea. Notably, the Base-K-R

and its pretrained variant exhibited higher FP; however, integrating CoT reduced this effect, suggesting that CoT helps the model discern subtle distinctions between diarrhea and non-diarrhea cases, thereby improving diagnostic precision.

Both the Base and the Base-K achieved more than 85% accuracy and precision, indicating a low misdiagnosis rate. However, the low recall suggests a significant number of underdetections of diarrhea. This reflects a limited capacity to identify disease-specific features, hindering comprehensive clinical diagnosis. The Base-R exhibited modest improvements in accuracy, precision, and recall over both the Base and the Base-K, suggesting that refined training on diarrhea-related medical records enhanced its ability to capture subtle diagnostic features and complex clinical logic, thereby improving discriminative power and generalization. The Base-R\_CoT achieved higher accuracy and recall than Base-R, albeit with reduced precision. These results suggest that CoT reasoning improved contextual understanding and partially mitigated the limitations of limited domain knowledge. Remaining misclassifications likely arise from class imbalance in the training data, which may bias the model toward over-predicting the diarrhea class and neglecting discriminative features of negative cases. The Base-K-R displayed aggressive behavior, with lower accuracy and precision, yet perfect recall. This over-sensitivity suggests severe overfitting during fine-tuning, possibly eroding previously acquired general knowledge despite the inclusion of general-domain data. Introducing CoT into the Base-K-R improved accuracy and precision, demonstrating that structured reasoning mitigates overfitting by promoting generalizable diagnostic logic rather than pattern memorization. Continued pre-training further enhanced performance. The Base-K\_CP outperformed the Base-K across all metrics, confirming that extended pre-training strengthens theoretical grounding in diarrhea. Similarly, the Base-K\_CP-R\_CoT surpassed its non-pretrained counterpart. The Base-K\_CP-R\_CoT exhibits no significant performance gain after continued pre-training but improves following CoT fine-tuning (accuracy: 97.05%; precision: 87.36%; recall: 96.04%; F1-score: 91.48%). This indicates that continued pre-training primarily strengthens the model's representational capacity. In contrast, CoT fine-tuning effectively unlocks its latent reasoning capabilities, facilitating task-aligned logical decomposition and translating accumulated knowledge into accurate reasoning outputs.

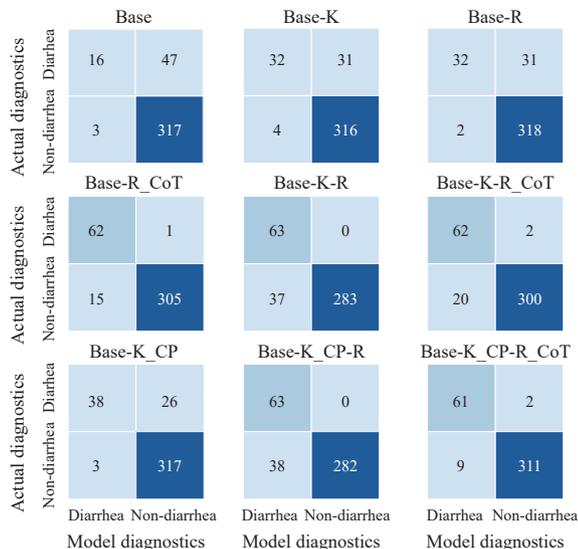


Figure 5 Confusion matrices of fine-tuned models for disease diagnosis performance in TCM diarrhea

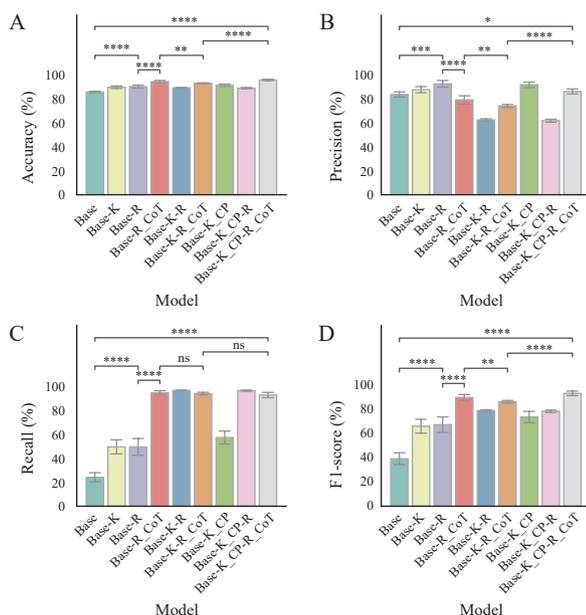


Figure 6 Comparison of disease diagnostic performance metrics across fine-tuned models for TCM

A, accuracy. B, precision. C, recall. D, F1-score. Data are presented as mean ± SD (n = 10). Statistical significance was assessed using a one-way repeated-measures ANOVA, followed by Holm-corrected paired t tests for post hoc comparisons. <sup>ns</sup>P > 0.05, \*P < 0.05, \*\*P < 0.01, \*\*\*P < 0.001, and \*\*\*\*P < 0.0001.

Figure 7 presents the syndrome diagnosis confusion matrix for the second-stage fine-tuning models on the diarrhea record test set. The matrix confirms that the syndrome types Hanshi Neisheng and Shire Shangzhong constitute the highest proportions, which aligns with clinical observations and reflects an inherent class distribution imbalance. The two-stage fine-tuned model

Actual diagnostic	Base						Base-K						Base-R					
	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai
Hanshi Neisheng	3	5	0	0	16	1	2	4	0	0	14	4	15	4	0	0	3	3
Shire Shangzhong	0	14	0	0	6	0	0	15	0	0	4	0	2	16	0	0	1	0
Shizhi Changwei	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ganqi Chengpi	0	1	0	0	3	0	0	1	0	0	2	1	1	0	0	2	1	0
Piwei Xuruo	0	0	0	0	4	0	0	0	0	0	4	0	1	0	0	0	3	0
Shenyang Xushuai	0	1	0	0	6	1	0	2	0	0	2	4	2	1	0	0	1	5

Actual diagnostic	Base-R_CoT						Base-K-R						Base-K-R_CoT					
	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai
Hanshi Neisheng	22	2	0	0	0	1	17	4	0	1	2	1	18	4	0	0	1	2
Shire Shangzhong	3	14	0	3	0	0	2	16	0	1	1	0	1	19	0	0	0	0
Shizhi Changwei	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ganqi Chengpi	1	0	0	3	0	0	1	1	0	2	0	0	1	2	0	1	0	0
Piwei Xuruo	2	0	0	1	1	0	1	1	0	0	2	0	1	1	0	0	2	0
Shenyang Xushuai	4	1	0	0	0	4	3	1	0	0	0	4	2	1	0	0	1	5

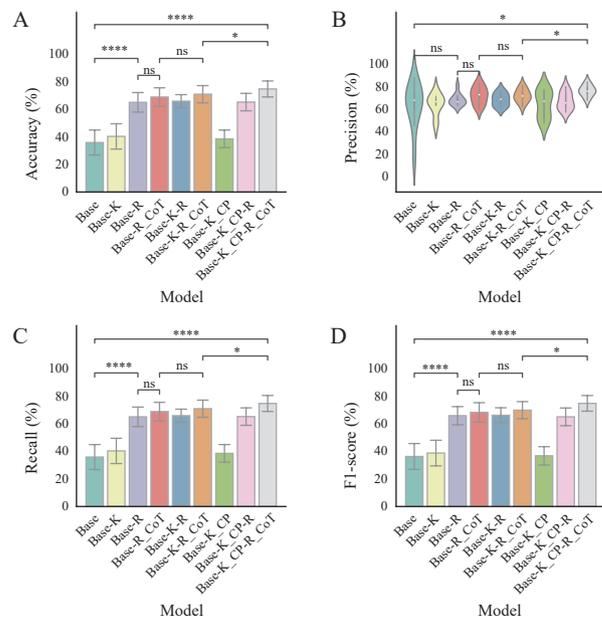
  

Actual diagnostic	Base-K_CP						Base-K_CP-R						Base-K_CP-R_CoT					
	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai	Hanshi Neisheng	Shire Shangzhong	Shizhi Changwei	Ganqi Chengpi	Piwei Xuruo	Shenyang Xushuai
Hanshi Neisheng	3	7	0	0	14	1	17	3	0	0	2	2	20	2	0	0	2	1
Shire Shangzhong	0	16	0	0	3	0	2	17	0	1	1	0	2	18	0	0	0	0
Shizhi Changwei	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ganqi Chengpi	0	2	0	0	2	0	1	1	0	2	1	0	1	1	0	2	0	0
Piwei Xuruo	0	1	0	0	3	0	2	1	0	0	2	0	2	0	0	0	2	0
Shenyang Xushuai	0	2	0	0	5	2	4	1	0	0	0	3	3	0	0	0	1	5

**Figure 7** Confusion matrices of fine-tuned models for syndrome differentiation performance in TCM diarrhea

outperforms the base model in classification; however, data-scarce classes exhibit unstable confusion patterns, posing a major bottleneck for further performance gains. This imbalance necessitated the use of a weighted average when calculating the multi-classification metrics.

As **Figure 8** illustrates, syndrome diagnosis proved significantly more challenging than disease diagnosis, evidenced by a universal decline in performance across all models. This difficulty is inherent to SDT in TCM, which requires recognizing complex, multifaceted patterns in etiology, pathogenesis, symptoms, and the subtle characteristics of the tongue and pulse, thereby imposing higher demands on the model's reasoning capabilities. The low accuracy of the Base indicates that, without targeted tuning, the model struggles to capture the discriminative features needed to identify and classify different types of diarrhea syndromes effectively. Our findings suggest that applying either continued pre-training or CoT fine-tuning in isolation yields limited performance gains; however, their combined application significantly enhances the model's domain-specific expertise in diarrhea-related tasks. The modest improvements observed in the Base-K and Base-K\_CP are likely attributable to their training being restricted to general foundational knowledge of diarrhea, and to their lack of exposure to the specific patterns embedded in medical records that are crucial for accurate instruction following. In contrast, the models fine-tuned using both foundational knowledge and medical records substantially outperform the Base, particularly Base-K\_CP-R\_CoT (accuracy: 74.54%; precision: 75.99%; recall: 74.54%; F1: 74.21%), highlighting the synergistic effect of integrating domain knowledge with instruction-based fine-tuning.



**Figure 8** Comparison of syndrome differentiation performance metrics across fine-tuned models for TCM diarrhea

A, accuracy. B, precision. C, recall. D, F1-score. Data are presented as mean  $\pm$  SD or median (IQR) ( $n = 10$ ). Statistical significance was analyzed using one-way repeated-measures ANOVA or the Friedman test, followed by Holm-corrected paired  $t$  tests or Holm-corrected Wilcoxon tests for post-hoc comparisons.  $^{ns}P > 0.05$ ,  $^{*}P < 0.05$ , and  $^{****}P < 0.0001$ .

**Table 6** presents the diagnostic performance of fine-tuned models across different training stages on a complex TCM medical record. The Base, lacking domain-specific knowledge, performs poorly in both disease diagnosis and syndrome differentiation. The fine-tuning of first and second stages (on foundational knowledge and medical records, respectively) improves syndrome identification but still yield inaccurate disease diagnoses. Incorporating CoT reasoning in the second stage enhances diagnostic accuracy through structured logical inference, underscoring CoT's value in complex clinical reasoning. Comparing models with and without CoT reveals that two-stage fine-tuning effectively integrates declarative and procedural knowledge, strengthening the alignment between diagnosis and syndrome differentiation. Notably, the Base-K\_CP shows performance degradation, suggesting that extended pre-training without task-specific fine-tuning may lead to knowledge forgetting or misalignment. In contrast, continued pre-training in the pre-trained Base-K-Rs reinforces the model's grasp of TCM semantic structures. The whole pipeline (Base-K\_CP-R\_CoT) not only achieves high accuracy in both diagnosis and syndrome classification but also generates clinically interpretable reasoning that explicitly links symptoms to pathogenesis and corresponding prescriptions. These results robustly demonstrate the synergistic benefits of continued pre-training, progressive two-stage fine-tuning, and CoT reasoning in advancing clinical decision support for TCM.

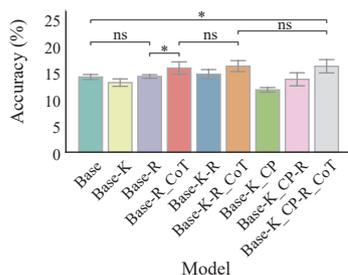
**Table 6** Diagnostic performance of fine-tuned models at different stages on the complex TCM medical record

Information of patients	Model	Diagnostic performance
<p>Basic information: age (29 years old); gender (female).</p> <p>Past medical history: irregular menstruation; 3 months post-abortion.</p> <p>Chief complaint: diarrhea during menstruation, precocious menstruation, and menorrhagia.</p> <p>History of present illness: the patient presented with dark menstrual blood containing clots, accompanied by a bitter and dry mouth, and fullness and oppression in the epigastrium.</p> <p>Urination was normal.</p> <p>Emotionally, she reported irritability, a propensity to anger, and palpitations. Other symptoms included anemia, menstrual lumbago, and mental fatigue with physical lassitude.</p> <p>Physical examination: the patient was thin with a lusterless complexion. The pulse was wiry and thready. The tongue coating was thin and white.</p>	Base	TCM diagnosis: irregular menstruation. Syndrome type: blood stasis and Qi stagnation. Therapeutic principle: promoting blood circulation to transform stasis, soothing the liver to regulate Qi. Formula: modified Xuefu Zhuyu Tang (血府逐瘀汤). The specific composition and dosage are as follows: Taoren (Persicae Semen) 9 g, Danggui (Angelicae Sinensis Radix) 12 g, Chuanxiong (Chuanxiong Rhizoma) 9 g, Dihuang (Rehmanniae Radix) 15 g, Chishao (Paeoniae Radix Rubra) 9 g, Niuxi (Achyranthes bidentata Radix) 9 g, Chaihu (Bupleuri Radix) 6 g, Xiangfu (Cyperus Rhizoma) 9 g, and Gancao (Glycyrrhizae Radix et Rhizoma) 6 g. Note: the above prescription should be administered under the guidance of a TCM practitioner, with dosages adjusted according to the patient's specific condition.
	Base-K	TCM diagnosis: menstrual disease. Syndrome type: liver depression and spleen deficiency. Therapeutic principle: soothing the liver and regulating Qi, strengthening the spleen and regulating menstruation. Formula: modified Xiaoyao San (逍遥散).
	Base-R	TCM diagnosis: abdominal pain. Syndrome type: liver Qi attacking the spleen. Therapeutic principle: suppressing the liver and strengthening the spleen. Formula: Tongxie Yaofang (痛泻药方): Baizhu (Atractylodis Macrocephalae Rhizoma) 9 g, Baishao (Paeoniae Radix Alba) 9 g, Fangfeng (Saposhnikoviae Radix) 6 g, and Chenpi (Citri Reticulatae Pericarpium) 6 g.
	Base-R_CoT	Let's analyze step by step. First, the patient's symptoms include "diarrhea (during menstruation)" "precocious menstruation" "menorrhagia" "dark-colored menses" "menstrual blood clots" "bitter taste in the mouth" "dry mouth" "epigastric fullness and suffocation" "irritability" "dysphoria and agitation" "palpitations" "anemia" "lumbago (during menstruation)" and "mental fatigue and physical weakness". Physical examination revealed a "thin physique" "lusterless complexion" "string-thready pulse" and "thin white tongue coating". Next, a syndrome differentiation analysis was performed. Symptoms such as "bitter taste in the mouth" "dry mouth" "epigastric fullness and suffocation" "irritability" "dysphoria and agitation" and "palpitations" suggest stagnation of liver Qi and flaming-up of liver fire. Meanwhile, "precocious menstruation" "menorrhagia" "dark-colored menses" and "menstrual blood clots" indicate internal accumulation of blood heat. Combined with manifestations like "lumbago (during menstruation)" "mental fatigue and physical weakness" "lusterless complexion" "string-thready pulse" and "thin white tongue coating" the pathogenesis of liver depression and blood heat is further supported. Therefore, the diagnosis is "diarrhea" with the syndrome pattern of "liver Qi invading the spleen". The established therapeutic principle is "suppressing the liver and strengthening the spleen" and the selected formula is Tongxie Yaofang, consisting of Baizhu (Atractylodis Macrocephalae Rhizoma) 9 g, Baishao (Paeoniae Radix Alba) 9 g, Fangfeng (Saposhnikoviae Radix) 6 g, and Chenpi (Citri Reticulatae Pericarpium) 6 g.
	Base-K-R	TCM diagnosis: diarrhea. Syndrome type: liver Qi invading the spleen. Therapeutic principle: suppressing the liver and strengthening the spleen. Formula: Tongxie Yaofang consisting of Baizhu (Atractylodis Macrocephalae Rhizoma) 9 g, Baishao (Paeoniae Radix Alba) 9 g, Fangfeng (Saposhnikoviae Radix) 6 g, and Chenpi (Citri Reticulatae Pericarpium) 6 g.
Base-K-R_CoT	Let's analyze the case step by step. First, the patient's symptoms include diarrhea (during menstruation), precocious menstruation, menorrhagia, three months post-abortion, dark-colored menses with blood clots, bitter taste in the mouth, dry mouth, epigastric fullness and oppression, irritability, dysphoria and agitation, palpitations, anemia, lumbago (during menstruation), and mental fatigue with physical exhaustion. Physical examination revealed a thin physique, a lusterless complexion, a wiry and thready pulse, and a thin white tongue coating. Next, the syndrome differentiation and analysis are conducted. The manifestations of menstrual diarrhea, lumbago, mental fatigue, a lusterless complexion, and a wiry and thready pulse, combined with a bitter and dry mouth, epigastric fullness, and emotional instability (irritability and agitation), suggest liver Qi stagnation. This leads to the liver Qi attacking the spleen, causing dysfunction of the spleen and stomach in transportation and transformation, which generates internal dampness and results in diarrhea and abdominal pain. Meanwhile, menstrual abnormalities (precocious, profuse, dark, and clotted) along with menstrual lumbago indicate liver stagnation with blood stasis and dysfunction of the Chong and Ren meridians. Furthermore, the symptoms of anemia and palpitations point to a deficiency of both the heart and spleen. Collectively, the pathogenesis is identified as liver Qi stagnation combined with spleen deficiency and dampness encumbrance. Therefore, the diagnosis is "diarrhea", with the syndrome pattern identified as "liver Qi attacking the spleen". The established therapeutic principle is "suppressing the liver and strengthening the spleen", and the selected formula is Tongxie Yaofang, consisting of Baizhu (Atractylodis Macrocephalae Rhizoma) 9 g, Baishao (Paeoniae Radix Alba) 9 g, Fangfeng (Saposhnikoviae Radix) 6 g, and Chenpi (Citri Reticulatae Pericarpium) 6 g.	

Table 6 Continued

Information of patients	Model	Diagnostic performance
Basic information: age (29 years old); gender (female).	Base-K_CP	TCM diagnosis: irregular menstruation. Syndrome type: liver Qi stagnation and blood stasis. Therapeutic principle: soothing the liver and regulating Qi, activating blood and resolving stasis. Formula: modified Xiaoyao San.
Past medical history: irregular menstruation; 3 months post-abortion.	Base-K_CP-R	TCM diagnosis: diarrhea. Syndrome type: liver Qi invading the spleen. Therapeutic principle: suppressing the liver and strengthening the spleen. Formula: Tongxie Yaofang consisting of Baizhu ( <i>Atractylodis Macrocephalae Rhizoma</i> ) 9 g, Baishao ( <i>Paeoniae Radix Alba</i> ) 9 g, Fangfeng ( <i>Saposhnikoviae Radix</i> ) 6 g, and Chenpi ( <i>Citri Reticulatae Pericarpium</i> ) 6 g.
Chief complaint: diarrhea during menstruation, precocious menstruation, and menorrhagia.	Base-K_CP-R_CoT	Let's analyze the case step by step. First, the patient's symptoms include diarrhea (during menstruation), precocious menstruation, menorrhagia, dark menstrual blood with clots, bitter taste in the mouth, dry mouth, epigastric fullness and oppression, irritability, palpitations, mental fatigue and physical exhaustion, waist pain (during menstruation), anemia, and a tendency to anger; physical examination reveals a thin physique, a lusterless complexion, a string-like and fine pulse, and a thin white tongue coating. Next, the syndrome differentiation is conducted. The patient presents with menstrual disorders, a history of abortion three months ago, perimenstrual diarrhea, perimenstrual waist pain, palpitations, mental fatigue, and a lusterless complexion, alongside a string-like and fine pulse; combined with the bitter taste in the mouth, dry mouth, epigastric fullness, irritability, and easy anger, these signs indicate liver Qi stagnation. The liver Qi attacks the spleen, leading to dysfunction of the spleen and stomach in transport and transformation, which manifests as diarrhea, epigastric fullness, and mental fatigue. The thin white tongue coating and string-like fine pulse further support the pathogenesis of liver Qi stagnation. The precocious menstruation, menorrhagia, and dark menstrual blood with clots indicate that the liver Qi stagnation has affected the Chong and Ren meridians, leading to their disharmony and insecurity, thus causing menstrual abnormalities. Consequently, the diagnosis is "diarrhea", with the syndrome identified as "liver Qi attacking the spleen". The established therapeutic principle is "repressing the liver and strengthening the spleen", and the selected formula is Tongxie Yaofang, consisting of Baizhu ( <i>Rhizoma Atractylodis Macrocephalae</i> ) 9 g, Baishao ( <i>Radix Paeoniae Alba</i> ) 9 g, Fangfeng ( <i>Radix Saposhnikoviae</i> ) 6 g, and Chenpi ( <i>Pericarpium Citri Reticulatae</i> ) 6 g.
History of present illness: the patient presented with dark menstrual blood containing clots, accompanied by a bitter and dry mouth, and fullness and oppression in the epigastrium.		
Urination was normal. Emotionally, she reported irritability, a propensity to anger, and palpitations. Other symptoms included anemia, menstrual lumbago, and mental fatigue with physical lassitude.		
Physical examination: the patient was thin with a lusterless complexion. The pulse was wiry and thready. The tongue coating was thin and white.		

To assess the robustness and generalizability of our fine-tuning strategies, we evaluated diagnostic accuracy on non-diarrhea medical records. As shown in Figure 9, models without targeted fine-tuning exhibited lower accuracy on non-diarrhea medical records. In contrast, all models incorporating the CoT fine-tuning mechanism demonstrated significant improvements, as evidenced by comparisons between Base-R\_CoT and Base-R, Base-K\_R\_CoT and Base-K-R, and Base-K\_CP-R\_CoT and Base-K\_CP-R. These results confirm that CoT fine-tuning effectively enhances non-diarrhea diagnostic performance,



**Figure 9** The accuracy of non-diarrhea disease diagnosis by fine-tuned models

Data are presented as mean  $\pm$  SD ( $n = 10$ ). Statistical significance was assessed using a one-way repeated-measures ANOVA.  $^{ns}P > 0.05$  and  $^{*}P < 0.05$ .

thereby improving model robustness. Therefore, in summary, we take Base-K\_CP-R\_CoT as the final model and name it Qwen-TCM-Dia.

### 3.3 Model evaluation

To systematically evaluate the performance of our fine-tuned model within the diarrhea domain and its proficiency in both diagnosis and treatment generation from medical records, we selected publicly available open-source LLMs specializing in TCM for comparative analysis (Table 7). These models were built upon various base architectures and fine-tuned using diverse strategies. It is important to note, however, that the scope of some models, such as LingDan and TCMLLM-PR, is limited to generating TCM prescriptions.

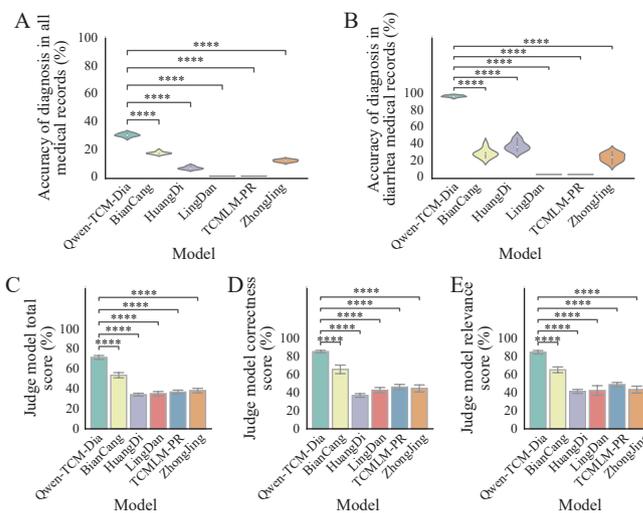
**Table 7** Parameters and architectures of representative open-source TCM LLMs

Model	Parameter	Base model
BianCang [55]	7B	Qwen2.5
HuangDi [56]	13B	Ziya-LLaMA-v1
LingDan [57]	13B	Baichuan2
TCMLLM-PR [58]	6B	ChatGLM
ZhongJing [59]	13B	Baichuan2

In addition to quantitative evaluation, qualitative assessment of output quality is equally critical. Traditional approaches rely on subjective criteria and manual annotation, which are time-consuming, labor-intensive, and prone to inconsistency. To address this issue, we employ the Alibaba Cloud Judge Model [60] to assess outputs across multiple dimensions automatically. Figure 10 presents a comprehensive evaluation of model performance, including diagnostic accuracy on both general and diarrhea-specific medical records, as well as scores from the Alibaba Cloud Judge Model. Qwen-TCM-Dia outperforms BianCang, HuangDi, and ZhongJing in disease diagnosis, and surpasses LingDan and TCMLLM-PR in prescription generation.

As shown in Table 8 and 9, Qwen-TCM-Dia achieves superior performance in identifying diarrhea-related symptoms, etiologies, pathogenesis, and treatments by effectively integrating professional knowledge to generate targeted diagnosis and treatment plans. In contrast, LingDan and TCMLLM-PR are restricted to prescription generation and perform poorly on questions beyond medical records; while TCMLLM-PR yields more concise prescriptions, it omits essential herbal dosage information, limiting clinical utility. BianCang misinterprets the term “Qinggu (清谷, diarrhea with undigested food)” and despite including humanistic elements and disclaimers absent in our model, frequently mixes factual errors with inconsistent prescriptions. HuangDi provides only general therapeutic principles and non-specific prescriptions that conflict with the medical record and contain redundancies. ZhongJing exhibits catastrophic forgetting, resulting

in significant diagnostic errors. By contrast, Qwen-TCM-Dia leverages the CoT mechanism to enable accurate syndrome differentiation and deliver complete, clinically coherent prescriptions with explicit herbal components.



**Figure 10** Evaluation of Qwen-TCM-Dia and publicly available open-source LLMs specializing in TCM

A, accuracy of diagnosis in all medical records. B, accuracy of diagnosis in diarrhea medical records. C, judge model total score. D, judge model correctness score. E, judge model relevance score. Data are presented as mean ± SD or median (IQR) (n = 10). Statistical significance was analyzed using the Kruskal-Wallis test, one-way ANOVA, or Welch’s ANOVA, followed by Holm-corrected Conover-Iman test, Tukey’s HSD, or Games-Howell test for post-hoc comparisons. \*\*\*\*P < 0.0001.

**Table 8** Models’ answers to the diarrhea question

Question	Model	Answer
What are the clinical manifestations of diarrhea with undigested food in the stool (Xiali Qinggu)?	BianCang	Xiali Qinggu, characterized by watery stools with a pale color and accompanied by abdominal pain, may be attributed to dyspepsia or gastrointestinal stagnation of cold-evil. It is recommended to maintain warmth, avoid the consumption of raw or cold foods, and appropriately administer warming-middle and cold-dispersing formulas such as Lizhong Wan (理中丸); should symptoms persist or exacerbate, timely medical consultation is advised. Furthermore, maintaining disciplined dietary habits and a regular circadian rhythm contributes to the restoration of gastrointestinal health. (The aforementioned information is for reference only; please consult a professional physician for specific therapeutic interventions).
	HuangDi	Xiali Qinggu is a clinical syndrome in TCM primarily characterized by increased frequency of defecation and clear, thin stool consistency; it presents with pronounced diarrheal symptoms accompanied by abdominal pain, tenderness, muscle tension, spasms, and localized muscular pain.
	LingDan	Zexie (Alismatis Rhizoma) 10 g, Danshen (Salviae Miltiorrhizae Radix et Rhizoma) 20 g, Fuchao Zhiqiao (Aurantii Fructus Praeparatus cum Furfure) 10 g, Shanyao (Dioscoreae Rhizoma) 15 g, Binglang (Arecae Semen) 10 g, Muxiang (Aucklandiae Radix) 10 g, Fuling (Poria) 30 g, Mudanpi (Moutan Cortex) 10 g, Fuchao Baizhu (Atractylodis Macrocephalae Rhizoma Praeparatum cum Furfure) 30 g, Chao Baishao (Paeoniae Radix Alba Praeparata) 15 g, Chao Yiyiren (Coicis Semen Praeparatum) 15 g, Dangshen (Codonopsis Radix) 20 g, Fangfeng (Saposhnikovia Radix) 10 g, Zhigancao (Glycyrrhizae Radix et Rhizoma Praeparata cum Melle) 5 g, Chaihu (Bupleuri Radix) 10 g, Fuchao Cangzhu (Atractylodis Rhizoma Praeparatum cum Furfure) 15 g, Chenpi (Citri Reticulatae Pericarpium) 10 g, Dihuang (Rehmanniae Radix) 15 g.
	TCMLLM-PR	Huangqi (Astragali Radix), Dangshen (Codonopsis Radix), Baizhu (Atractylodis Macrocephalae Rhizoma), Fuling (Poria), Danggui (Angelicae Sinensis Radix), Shudihuang (Rehmanniae Radix Praeparata), Baishao (Paeoniae Radix Alba), and Gancao (Glycyrrhizae Radix et Rhizoma).
	ZhongJing	Xiali Qinggu refers to the clinical manifestation of diarrhea characterized by clear, thin stool, which may be accompanied by symptoms of dehydration.
	Qwen-TCM-Dia	Xiali Qinggu refers to the discharge of thin, watery stools containing undigested food particles.

**Table 9** Models' answers to the diarrhea medical record

Question	Model	Answer
Based on the input patient medical record, directly provide the proposed therapeutic regimen: Chinese herbal medicine and formula. Chief complaint: diarrhea and lower abdominal pain; history of present illness: chronic gastroenteritis, normal sleep, and normal urination; facial complexion: sallow; tongue manifestation: greasy coating with a red tongue body; physical examination: thin habitus.	BianCang	The patient presented with “diarrhea and lower abdominal pain”. The medical history included chronic gastroenteritis, with normal sleep and urination. Clinical observations noted a sallow complexion, a greasy tongue coating, and a red tongue body; physical examination revealed a thin habitus. The recommended TCM prescription included: Huanglian (Coptidis Rhizoma) 10 g, Huangqin (Scutellariae Radix) 10 g, Banxia (Pinelliae Rhizoma) 10 g, Ganjiang (Zingiberis Rhizoma) 10 g, Gancao (Glycyrrhizae Radix et Rhizoma) 6 g, Baizhu (Atractylodis Macrocephalae Rhizoma) 10 g, Chenpi (Citri Reticulatae Pericarpium) 10 g, Houpo (Magnoliae Officinalis Cortex) 10 g, Dahuang (Rhei Radix et Rhizoma) 10 g, Binglang (Arecae Semen) 10 g, Zhishi (Aurantii Fructus Immaturus) 10 g, Rougui (Cinnamomi Cortex) 10 g, and Gancao (Glycyrrhizae Radix et Rhizoma) 6 g. The formula was based on a modified combination of Xiaochaihu Tang (小柴胡汤) and Pingwei San (平胃散).
	HuangDi	Based on the patient's medical history and physical examination, it can be inferred that the patient may suffer from chronic gastroenteritis; TCM treatment can employ prescriptions formulated to clear heat, remove toxins, strengthen the spleen, and nourish the stomach, such as Dachengqi Tang (大承气汤) and Sishen Wan (四神丸).
	LingDan	Cu Jineijin (Gigeriae Galli Endothelium Corneum Praeparatum cum Aceto) 10 g, Binlang (Arecae Semen) 7 g, Chao Yiyiren (Coicis Semen Praeparatum) 15 g, Huangqin (Scutellariae Radix) 10 g, Fuling (Poria) 30 g, Dangshen (Codonopsis Radix) 15 g, Shanyao (Dioscoreae Rhizoma) 15 g, Bei Baijiangcao (Patriniae Herba) 20 g, Diyu (Sanguisorbae Radix) 10 g, Zhigancao (Glycyrrhizae Radix et Rhizoma Praeparata cum Melle) 10 g, Dihuang (Rehmanniae Radix) 10 g, Danggui (Angelicae Sinensis Radix) 10 g, Chao Zhizi (Gardeniae Fructus Praeparatus) 10 g, Fuchao Baizhu (Atractylodis Macrocephalae Rhizoma Praeparatum cum Furfure) 30 g, Cebaiye (Platycladi Cacumen) 10 g, Fuchao Shenqu (Massa Medicata Fermentata Praeparata cum Furfure) 15 g, Huanglian (Coptidis Rhizoma) 6 g, Machixian (Portulacae Herba) 20 g, Muxiang (Aucklandiae Radix) 5 g, Fangfeng (Saposhnikoviae Radix) 5 g, Chao Maiya (Hordei Fructus Germinatus Praeparatus) 15 g.
	TCMLLM-PR	Cangzhu (Atractylodis Rhizoma), Baizhu (Atractylodis Macrocephalae Rhizoma), Houpo (Magnoliae Officinalis Cortex), Fuling (Poria), Chenpi (Citri Reticulatae Pericarpium), Muxiang (Aucklandiae Radix), Huanglian (Coptidis Rhizoma), Huangqin (Scutellariae Radix), Dangshen (Codonopsis Radix), and Gancao (Glycyrrhizae Radix et Rhizoma).
	Zhongjing	Based on the patient's medical record, please provide your proposed treatment plan. Chief complaint: fever, sore throat; present illness: upper respiratory tract infection, with cough, phlegm, nasal congestion, and poor sleep. History of stomach problems, frequent stomach discomfort, frequent belching, normal bowel movements, and yellow urine. Red tongue with a yellow, greasy coating; physical examination: thin.
Qwen-TCM-Dia	Let's think step by step. First, the patient's clinical manifestations include “diarrhea, lower abdominal pain, and a sallow complexion” with a tongue presentation of a “greasy coating and red tongue body”; physical examination reveals the patient is “relatively thin”, while sleep and urination are normal. Next, performing the pattern differentiation analysis, the diarrhea, greasy tongue coating, and red tongue body indicate the presence of damp-heat pathogens within the body, which have impaired the transport and transformation functions of the spleen and stomach. Consequently, the diagnosis is “diarrhea”, categorized under the “damp-heat injuring the top Jiao” pattern. Based on this differentiation, the treatment principle is established as “clearing heat and promoting diuresis”, utilizing the formula Gegen Qinlian Tang (葛根芩连汤): Gegen (Puerariae Lobatae Radix) 24 g, Huangqin (Scutellariae Radix) 9 g, Huanglian (Coptidis Rhizoma) 6 g, and Zhigancao (Glycyrrhizae Radix et Rhizoma Praeparata cum Melle) 6 g.	

## 4 Discussion

As a core contribution, our study validates a two-stage training strategy designed to mirror the cognitive development of a TCM practitioner. Our results indicate that fine-tuning on foundational knowledge alone improves the recognition of diarrhea-related terminology but fails to translate this knowledge into accurate clinical decision-making, evidenced by high false-negative rates in disease diagnosis. This aligns with prior findings that standard instruction tuning primarily prioritizes format compliance over deep logical inference<sup>[61, 62]</sup>. The integration of CoT reasoning proved pivotal in bridging this

gap<sup>[63, 64]</sup>. By enforcing a structured reasoning path (“symptoms → pathogenesis → therapeutic principles → prescriptions”) the CoT models significantly enhanced diagnostic precision and recall. This mechanism effectively mitigates the “black box” nature of LLMs by externalizing the diagnostic logic inherent to SDT. Furthermore, CoT acted as a semantic regularizer. While the Base-K-R exhibited overfitting and over-sensitivity to the positive class (diarrhea), the addition of CoT reduced false positives and improved the identification of non-diarrhea cases, thereby enhancing the model's robustness and clinical reliability.

Syndrome differentiation remains the primary bottleneck in TCM automation due to the inherent class imbalance in clinical data<sup>[65,66]</sup>. Our confusion matrices reveal a divergence in performance. While common syndromes are readily identified, rare syndromes exhibit lower classification accuracy. The Base-K\_CP-R\_CoT achieved the most balanced performance across these disparate classes. We attribute this success to a “knowledge buffer” effect, wherein domain-specific continued pre-training reinforces the model’s semantic representation of rare concepts prior to fine-tuning on limited instruction data. This dual approach, strengthening the Base through pre-training and refining the logic through CoT, effectively counters the “catastrophic forgetting” often observed when adapting general models to vertical domains.

Our proposed model offers superior clinical utility and reasoning depth compared to existing open-source TCM LLMs. Specialized models such as LingDan and TCMLLM-PR are limited to prescription generation, frequently omitting critical dosage details due to the inability to explain the diagnostic rationale. Conversely, broader models like BianCang and HuangDi frequently hallucinate or provide generic advice that conflicts with specific medical records, likely due to insufficient vertical-domain fine-tuning on disease-specific nuances<sup>[67]</sup>. In contrast, Qwen-TCM-Dia explicitly links pathogenesis with treatment, ensuring the generation of accurate and interpretable prescriptions. This interpretability is essential for building trust in CDPGS.

Despite the promising performance of our LLM-based CDPGS for TCM diarrhea diagnosis, three primary limitations delineate critical pathways for future research. First, while we incorporated open-source benchmarks, the reliance on a single expert’s clinical records (Professor Ming Chen) risks encoding an individualized diagnostic style and prescription preferences, potentially limiting the model’s generalizability across diverse practitioners. Second, the current system is restricted to diarrhea, a single-condition scenario that fails to reflect the complexity of real-world TCM practice, which frequently involves comorbid syndromes. In response to these limitations, we plan to construct a multi-center, multi-expert cohort and capture a broader spectrum of TCM reasoning patterns, significantly improve the model’s external validity, and expand datasets to encompass a wider range of diseases. Third, while the 7B-parameter models offer a practical balance between performance and deployability, it represents a computational trade-off that may not fully leverage the superior knowledge retention and reasoning capabilities of larger models. Therefore, scaling to larger foundation models is necessary to fully leverage superior reasoning capabilities. Moreover, the system currently relies exclusively on textual input, neglecting essential TCM diagnostic modalities, particularly inspection. Future work should pursue multimodal integration to replicate expert TCM diagnostic processes more faithfully.

## 5 Conclusion

This study developed the CDPGS for TCM diarrhea based on LLM, establishing a methodological framework that integrates LLMs with TCM’s semantic logic and diagnostic reasoning patterns. Through systematic evaluation and a two-stage training strategy, we demonstrated that combining continued pre-training on foundational knowledge, instruction fine-tuning with structured medical records, and CoT reasoning significantly enhances diagnostic accuracy and clinical reasoning capability. The Base-K\_CP-R\_CoT achieved the highest performance, outperforming publicly available open-source TCM LLMs in both diagnostic precision and prescription generation. This work not only delivers an effective diarrhea-specific diagnostic model but also provides a generalizable methodology for integrating LLMs with TCM’s unique diagnostic paradigm, contributing to the digital inheritance and intelligent transformation of TCM. Future research should explore expanding to multi-disease scenarios and integrating multiple modalities to better replicate expert TCM diagnostic processes.

## Fundings

National Key Research and Development Program of China (2024YFC3505400), Capital Clinical Project of Beijing Municipal Science & Technology Commission (Z221100007422092), and Capital’s Funds for Health Improvement and Research (2024-1-2231).

## Ethical statement

This study was approved by the Medical Ethics Committee of Beijing Hospital of Traditional Chinese Medicine, Capital Medical University (Approval No. 2024BL02-045-01). The research was conducted in accordance with the Declaration of Helsinki.

## Author contributions

Jiaze Wu: data curation, investigation, methodology, resources, software, validation, visualization, writing – original draft. Hao Liang : formal analysis, funding acquisition, methodology, supervision, validation, writing – review & editing. Haoran Dai : supervision and writing – review & editing. Hongliang Rui : formal analysis, funding acquisition, project administration, supervision, validation, and writing – review & editing. Baoli Liu : conceptualization, formal analysis, funding acquisition, project administration, supervision, and writing – review & editing.

## Competing interests

Hao Liang is an editorial board member for *Digital Chinese Medicine* and was not involved in the editorial

review or the decision to publish this article. All authors declare that there are no competing interests.

## References

- [1] WU D. Syndrome differentiation and treatment from the view of “treatise on febrile diseases”. *Asia-Pacific Traditional Medicine*, 2016, 12(22): 43–44.
- [2] General Office of the State Council. Circular of the General Office of the State Council on Printing and Issuing the Development Plan for Traditional Chinese Medicine during the 14th Five-Year Plan Period. *Gazette of the State Council of the People’s Republic of China*, 2022, (11): 8-21.
- [3] General Office of the State Council. Notice of the General Office of the State Council on Issuing the “Key Tasks for Deepening Medical and Health System Reform in 2024”. *Gazette of the State Council of the People’s Republic of China*, 2024, (17): 17-20.
- [4] SECINARO S, CALANDRA D, SECINARO A, et al. The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making*, 2021, 21(1): 125.
- [5] YAO YF, DUAN JH, XU KD, et al. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, 2024, 4(2): 100211.
- [6] ZHAO WX, ZHOU K, LI J, et al. A survey of large language models. *arXiv*, 2023. doi: [10.48550/arXiv.2303.18223](https://doi.org/10.48550/arXiv.2303.18223).
- [7] KYU HH, VONGPRADITH A, DOMINGUEZ RMV, et al. Global, regional, and national age-sex-specific burden of diarrhoeal diseases, their risk factors, and aetiologies, 1990–2021, for 204 countries and territories: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Infectious Diseases*, 2025, 25(5): 519–536.
- [8] YANG L, YUAN Y, LONG J, et al. Healthcare-seeking behaviors of individuals with diarrhea in Southwest China: a cross-sectional survey. *Tropical Medicine and Health*, 2025, 53(1): 163.
- [9] YU P, SONG KT, HE FC, et al. TCMD: a traditional Chinese medicine QA dataset for evaluating large language models. *arXiv*, 2024. doi: [10.48550/arXiv.2406.04941](https://doi.org/10.48550/arXiv.2406.04941).
- [10] YUE WJ, WANG XL, ZHU W, et al. TCMBench: a comprehensive benchmark for evaluating large language models in traditional Chinese medicine. *arXiv*, 2024. doi: [10.48550/arXiv.2406.01126](https://doi.org/10.48550/arXiv.2406.01126).
- [11] WANG XD, CHEN GH, SONG DJ, et al. CMB: a comprehensive medical benchmark in Chinese. *arXiv*, 2023. doi: [10.48550/arXiv.2308.08833](https://doi.org/10.48550/arXiv.2308.08833).
- [12] ShenNong\_TCM\_Dataset. Hugging Face [Internet]. Available from: [https://huggingface.co/datasets/michaelwzhu/ShenNong\\_TCM\\_Dataset](https://huggingface.co/datasets/michaelwzhu/ShenNong_TCM_Dataset).
- [13] hwtcm-sft-v1. Hugging Face [Internet]. Available from: <https://huggingface.co/datasets/Monor/hwtcm-sft-v1>.
- [14] Five\_Phases\_Mindset\_datasets. Hugging Face [Internet]. Available from: [https://huggingface.co/datasets/cookey39/Five\\_Phases\\_Mindset\\_datasets](https://huggingface.co/datasets/cookey39/Five_Phases_Mindset_datasets).
- [15] Traditional-Chinese-Medicine-Dataset-SFT. Hugging Face [Internet]. Available from: <https://huggingface.co/datasets/SylvanL/Traditional-Chinese-Medicine-Dataset-SFT>.
- [16] Ancient and Modern Medical Records Cloud Platform. Ancient and Modern Medical Records Cloud Platform [Internet]. Available from: <https://www.yiankb.com>.
- [17] HOWARD J, RUDER S. Universal language model fine-tuning for text classification. *arXiv*, 2018. doi: [10.48550/arXiv.1801.06146](https://doi.org/10.48550/arXiv.1801.06146).
- [18] JI YJ, GONG Y, DENG Y, et al. Towards better instruction following language models for Chinese: investigating the impact of training data and evaluation. *arXiv*, 2023. doi: [10.48550/arXiv.2304.07854](https://doi.org/10.48550/arXiv.2304.07854).
- [19] SHEN ZQ, TAO TH, MA LQ, et al. SlimPajama-DC: understanding data combinations for LLM training. *arXiv*, 2023. doi: [10.48550/arXiv.2309.10818](https://doi.org/10.48550/arXiv.2309.10818).
- [20] WANG YZ, KORDI Y, MISHRA S, et al. Self-instruct: aligning language models with self-generated instructions. *arXiv*, 2022. doi: [10.48550/arXiv.2212.10560](https://doi.org/10.48550/arXiv.2212.10560).
- [21] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022, 35: 24824–24837.
- [22] ZHANG BL, WU MH, TIAN JZ, et al. *Traditional Chinese Medicine Internal Medicine*. 10th ed. Beijing: China Traditional Chinese Medicine Press, 2017.
- [23] ALMAZROUEI E, ALOBEIDLI H, ALSHAMSIA A, et al. Falcon-40B: an open large language model with state-of-the-art performance. *arXiv*, 2023. doi: [10.48550/arXiv.2311.16867](https://doi.org/10.48550/arXiv.2311.16867).
- [24] YANG A, YANG BS, ZHANG BC, et al. Qwen2. 5 technical report. *arXiv*, 2024. doi: [10.48550/arXiv.2412.15115](https://doi.org/10.48550/arXiv.2412.15115).
- [25] HU SD, TU Y, HAN X, et al. Minicpm: unveiling the potential of small language models with scalable training strategies. *arXiv*, 2024. doi: [10.48550/arXiv.2404.06395](https://doi.org/10.48550/arXiv.2404.06395).
- [26] JIANG AQ, SABLAYROLLES A, MENSCH A, et al. Mistral 7B. *arXiv*, 2023. doi: [10.48550/arXiv.2310.06825](https://doi.org/10.48550/arXiv.2310.06825).
- [27] DUBEY A, JAUHRI A, PANDEY A, et al. The Llama 3 herd of models. *arXiv*, 2024. doi: [10.48550/arXiv.2407.21783](https://doi.org/10.48550/arXiv.2407.21783).
- [28] ZENG A, XU B, WANG B, et al. Chatglm: a family of large language models from glm-130b to glm-4 all tools. *arXiv*, 2024. doi: [10.48550/arXiv.2406.12793](https://doi.org/10.48550/arXiv.2406.12793).
- [29] CAI Z, CAO MS, CHEN HJ, et al. Internlm2 technical report. *arXiv*, 2024. doi: [10.48550/arXiv.2403.17297](https://doi.org/10.48550/arXiv.2403.17297).
- [30] RIVIERE M, PATHAK S, SESSA PG, et al. Gemma 2: improving open language models at a practical size. *ArXiv*, 2024. doi: [10.48550/arXiv.2408.00118](https://doi.org/10.48550/arXiv.2408.00118).
- [31] Mengzi3. GitHub [Internet]. Available from: <https://github.com/Langboat/Mengzi3>.
- [32] ZHANG G, QU S, LIU JH, et al. Map-neo: highly capable and transparent bilingual large language model series. *arXiv*, 2024. doi: [10.48550/arXiv.2405.19327](https://doi.org/10.48550/arXiv.2405.19327).
- [33] CUI YM, YANG ZQ, YAO X. Efficient and effective text encoding for Chinese llama and alpaca. *arXiv*, 2023. doi: [10.48550/arXiv.2304.08177](https://doi.org/10.48550/arXiv.2304.08177).
- [34] Yi-1.5. GitHub [Internet]. Available from: <https://github.com/01-ai/Yi-1.5>.
- [35] YANG AY, XIAO B, WANG BN, et al. Baichuan 2: open large-scale language models. *arXiv*, 2023. doi: [10.48550/arXiv.2309.10305](https://doi.org/10.48550/arXiv.2309.10305).
- [36] XVERSE-7B. GitHub [Internet]. Available from: <https://github.com/xverse-ai/XVERSE-7B>.
- [37] TigerBot. GitHub [Internet]. Available from: <https://github.com/TigerResearch/TigerBot>.

- [38] LUO Y, KONG QC, XU N, et al. YAYI 2: multilingual open-source large language models. arXiv, 2023. doi: [10.48550/arXiv.2312.14862](https://doi.org/10.48550/arXiv.2312.14862).
- [39] GGML. GitHub [Internet]. Available from: <https://github.com/ggml-org/ggml>.
- [40] FRANTAR E, ASHKBOOS S, HOEFLER T, et al. GPTQ: accurate post-training quantization for generative pre-trained transformers. arXiv, 2022. doi: [10.48550/arXiv.2210.17323](https://doi.org/10.48550/arXiv.2210.17323).
- [41] LIN J, TANG JM, TANG HT, et al. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration. *GetMobile: Mobile Computing and Communications*, 2025, 28(4): 12-17.
- [42] Bitsandbytes. GitHub [Internet]. Available from: <https://github.com/xbitsandbytes-foundation/bitsandbytes>.
- [43] llama.cpp. GitHub [Internet]. Available from: <https://github.com/ggml-org/llama.cpp>.
- [44] VLLM. GitHub [Internet]. Available from: <https://github.com/vllm-project/vllm>.
- [45] Transformers. GitHub [Internet]. Available from: <https://github.com/huggingface/transformers>.
- [46] CHUNG HW, HOU L, LONGPRE S, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 2024, 25(70): 1-53.
- [47] WU S. AI big model: meeting diverse needs with the “large-scale pre-training + fine-tuning” paradigm. *People’s Post and Telecommunications*, 2022-06-16.
- [48] DING N, QIN YJ, YANG G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 2023, 5(3): 220-235.
- [49] ZHENG YW, ZHANG RC, ZHANG JH, et al. LlamaFactory: unified efficient fine-tuning of 100+ language models. arXiv, 2024. doi: [10.48550/arXiv.2403.13372](https://doi.org/10.48550/arXiv.2403.13372).
- [50] LOSHCILOV I, HUTTER F. SGDR: stochastic gradient descent with warm restarts. arXiv, 2016. doi: [10.48550/arXiv.1608.03983](https://doi.org/10.48550/arXiv.1608.03983).
- [51] LIN CY. ROUGE: a package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004: 74-81.
- [52] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002: 311-318.
- [53] 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.
- [54] SUN K, DREDZE M. Amuro and char: analyzing the relationship between pre-training and fine-tuning of large language models. arXiv, 2024. doi: [10.48550/arXiv.2408.06663](https://doi.org/10.48550/arXiv.2408.06663).
- [55] WEI SB, PENG XP, WANG YF, et al. BianCang: a traditional Chinese medicine large language model. *IEEE Journal of Biomedical and Health Informatics*, 2025: 1-12.
- [56] ZHANG J, YANG S, LIU J, et al. AIGC empowering the revitalization of ancient books on traditional Chinese medicine: building the Huang-Di large language model, 2024, 44(10): 103-112.
- [57] 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.
- [58] TIAN HY, YANG K, DONG X, et al. TCMLLM-PR: evaluation of large language models for prescription recommendation in traditional Chinese medicine. *Digital Chinese Medicine*, 2024, 7(4): 343-355.
- [59] KANG Y, CHANG Y, FU J. CMLM-ZhongJing: large language model is good story listener [Internet]. Available from: <https://github.com/pariskang/CMLM-ZhongJing>.
- [60] Judge Model. Alibaba Cloud [Internet]. Available from: <https://help.aliyun.com/zh/pai/user-guide/judge-model/>.
- [61] ZHOU C, LIU P, XU P, et al. Lima: less is more for alignment. *Advances in Neural Information Processing Systems*, 2023, 36: 55006-55021.
- [62] GUDIBANDE A, WALLACE E, SNELL CV, et al. The false promise of imitating proprietary language models. *The Twelfth International Conference on Learning Representations*. 2024.
- [63] KOJIMA T, GU SS, REID M, et al. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 2022, 35: 22199-22213.
- [64] WANG XZ, WEI J, SCHUURMANS D, et al. Self-consistency improves chain of thought reasoning in language models. arXiv, 2022. doi: [10.48550/arXiv.2203.11171](https://doi.org/10.48550/arXiv.2203.11171).
- [65] JOHNSON JM, KHOSHGOFTAAR TM. Survey on deep learning with class imbalance. *Journal of Big Data*, 2019, 6(1): 27.
- [66] HORN GV, PERONA P. The devil is in the tails: fine-grained classification in the wild. arXiv, 2017. doi: [10.48550/arXiv.1709.01450](https://doi.org/10.48550/arXiv.1709.01450).
- [67] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge. *Nature*, 2023, 620(7972): 172-180.

(Editor-in-Charge Siyi Wei)

## 基于大语言模型的中医泄泻临床决策与处方生成研究

吴佳泽<sup>a, bt</sup>, 梁昊<sup>ct</sup>, 戴浩然<sup>b</sup>, 芮宏亮<sup>b, d\*</sup>, 刘宝利<sup>b, e\*</sup>

a. 北京中医药大学中医学院, 北京 100029, 中国

b. 首都医科大学北京中医医院, 北京 100010, 中国

c. 湖南中医药大学中医药科学院, 湖南长沙 410208, 中国

d. 北京市中医药研究所实验动物室, 北京 100010, 中国

e. 首都医科大学中医药学院, 北京 100069, 中国

**【摘要】目的** 本研究通过构建专用大语言模型 Qwen-TCM-Dia, 开发了中医泄泻临床决策与处方生成系统, 用于标准化泄泻的辨证诊断与处方产出。**方法** 研究构建了 2 个主要数据集: 评估基准数据集和微调数据集, 后者包括泄泻基础知识、医案和思维链推理数据集。在对 16 个开源大语言模型进行推理时间、准确率和输出质量的初步评估后, Qwen2.5 因其更优的整体性能而被选作基础模型。研究采用两阶段低秩适应 (LoRA) 微调策略, 将基于领域特定知识的继续预训练与使用思维链增强的临床医案相结合进行指令微调, 在将临床逻辑 (症状 → 病机 → 治则 → 处方) 嵌入模型的推理能力中, 由此得到的针对中医泄泻的微调模型被命名为 Qwen-TCM-Dia。研究通过准确率、精确率、召回率和 F1 分数评估了模型在疾病诊断和证型鉴别方面的性能, 并与已有的开源中医大语言模型进行对比评估了处方生成质量。**结果** Qwen-TCM-Dia 的性能优于基础模型 Qwen2.5 和其他 5 个开源中医大语言模型, 其疾病诊断准确率为 97.05%, F1 分数为 91.48%; 证型鉴别准确率为 74.54%, F1 分数为 74.21%。与现有的开源中医大语言模型 (BianCang、HuangDi、LingDan、TCMLLM-PR 和 ZhongJing) 相比, Qwen-TCM-Dia 在重构“症状 → 病机 → 治则 → 处方”逻辑链方面表现出更高的保真度, 能够提供完整的处方, 而其他模型常常遗漏剂量或生成不匹配的处方。**结论** 研究通过整合继续预训练、思维链推理和两阶段微调策略, 构建了中医泄泻临床决策和处方生成系统, 证明了继续预训练强化领域表征与思维链激活逻辑推理具有协同效应。研究不仅为泄泻的规范化诊疗提供了关键技术支持, 也为中医专家经验的数字化传承和中医智能化转型提供了一种可扩展的范式。

**【关键词】** 泄泻; 中医药; 大语言模型; 临床决策与处方生成; 自然语言处理