System Report for CCL25-Eval Task 9: Leveraging Chain-of-Thought and Multi-task Learning for Optimized Traditional Chinese Medicine Diagnosis and Treatment

Jian Zhang **Wei Zhu** **Zhiwen Tang***

Yunnan Key Laboratory of Intelligent Systems and Computing, Yunnan University, Kunming, China

School of Information Science and Engineering, Yunnan University, Kunming, China

18214956871@stu.ynu.edu, zhuwei@stu.ynu.edu.cn, zhiwen.tang@ynu.edu.cn

Abstract

This paper introduces an intelligent diagnostic system for Traditional Chinese Medicine (TCM) that emulates clinical reasoning through a phased multi-turn dialogue process. The system architecture is divided into three sequential stages: syndrome differentiation, disease diagnosis, and prescription generation. Each stage leverages Chain-of-Thought (CoT) techniques to ensure coherent reasoning, maintaining contextual continuity and consistency throughout the diagnostic process. To optimize model performance, we employ a multi-task fine-tuning approach, combining data from all three stages for training the Qwen2.5-7B-Instruct model. Experimental results show that the system achieves strong performance across all diagnostic tasks. Error analysis reveals that the accuracy of the first two stages, syndrome differentiation and disease diagnosis, has a significant impact on the quality of the generated prescriptions. This work provides a scalable framework for intelligent TCM diagnosis, advancing both medical knowledge reasoning and the application of domain-specific large language models.

Keywords: Traditional Chinese Medicine , Chain-of-Thought Prompting , Multi-task Fine-tuning , Multi-turn Dialogue , Diagnostic Reasoning

1 Introduction

With the rapid advancement of intelligent systems in Traditional Chinese Medicine (TCM), AI-powered platforms for TCM-assisted diagnosis have emerged as a significant research focus. These platforms aim to simulate the progressive reasoning process inherent in TCM, which involves analyzing symptoms, identifying syndrome types, associating them with diseases, and generating personalized prescriptions. However, the challenges in modeling TCM language are substantial. TCM terminology is highly specialized, with complex syndrome names such as "肝郁脾虚证" and intricate herbal compatibilities that are difficult to capture. Moreover, clinical texts often contain substantial structural variability, with unstructured or redundant descriptions of patient complaints, which complicates the extraction of key diagnostic information. The diagnostic process also requires multi-turn reasoning, where the model must maintain coherent context across multiple interactions to arrive at accurate conclusions—a task that traditional single-turn models struggle to achieve, often resulting in fragmented context and logical inconsistencies.

To address these challenges, this paper proposes an intelligent TCM diagnostic approach that integrates a three-stage prompt engineering framework. The system breaks down the diagnostic process into syndrome differentiation, disease diagnosis, and prescription generation, each step building on the previous one and maintaining coherence through multi-turn dialogue. By utilizing Chain-of-Thought (CoT) (Wei et al., 2022) techniques, the system ensures that each stage's reasoning is informed by prior steps, enabling a smooth transition between tasks. The approach also employs a multi-task fine-tuning strategy, allowing the model to be trained on data from all three stages simultaneously, thus enhancing its

Proceedings of the 24th China National Conference on Computational Linguistics, pages 369-375, Jinan, China, August 11-14, 2025.

adaptability to the TCM domain. Through the use of extended context windows, rigorous data cleaning, and standardized formatting, the model is further optimized for clinical use.

Experimental results demonstrate that the system effectively replicates the clinical reasoning process in TCM, with the multi-turn reasoning mechanism showing a strong ability to simulate clinical thinking. The system performs well across all diagnostic tasks, with error analysis highlighting the crucial role of the initial stages, i.e. syndrome differentiation and disease diagnosis, in ensuring the quality of the final prescription generation. These findings underline the importance of maintaining consistency and coherence across the entire diagnostic process, offering a promising approach for deploying large language models in real-world TCM clinical scenarios.

2 Related Work

2.1 Prompt Engineering for Complex Medical Reasoning

Prompt engineering, the technique of crafting input prompts to guide large language models (LLMs), has emerged as a powerful and cost-effective alternative to traditional fine-tuning. With the advent of models like GPT-3 (Floridi and Chiriatti, 2020) and GPT-4 (Achiam et al., 2023), prompt engineering has been widely adopted in various domains, particularly in healthcare. In Traditional Chinese Medicine (TCM), the inherent complexity of syndrome differentiation, disease diagnosis, and prescription recommendation poses a major challenge for LLMs. By designing structured prompts and multi-stage templates, researchers have successfully transformed these domain-specific processes into machine-understandable tasks, thereby enhancing interpretability and controllability.

A notable advancement in this domain is the use of Chain-of-Thought (CoT) prompting(Wei et al., 2022), exemplified by the 'Let's Think Step by Step' strategy, which shows that LLMs improve performance on complex reasoning tasks by breaking them down into intermediate steps. Furthermore, multi-turn dialogue structures have been shown to enable models to engage in contextual reasoning over multiple stages. For instance, a follow-up questioning mechanism was introduced for medical QA systems(Sorathiya et al., 2021), which mitigated the information loss commonly observed in single-turn interactions. In the TCM setting, this multi-turn interaction naturally aligns with the stepwise clinical workflow—syndrome identification $\rightarrow$ disease inference $\rightarrow$ formula generation—allowing the model to simulate a more human-like diagnostic reasoning process.

2.2 Fine-tuning Techniques for Domain-Specific Adaptation

As LLMs grow in size, full-parameter fine-tuning becomes computationally prohibitive. To address this, parameter-efficient fine-tuning (PEFT) methods have been developed. Among them, LoRA (Low-Rank Adaptation) (Hu et al., 2022) introduces low-rank matrices into the transformer layers to reduce trainable parameters while maintaining performance. QLoRA (Quantized LoRA) (Dettmers et al., 2023) further integrates 4-bit quantization into the LoRA framework, enabling efficient fine-tuning on commodity GPUs.

These techniques have been instrumental in adapting LLMs to domain-specific applications such as biomedicine, law, and TCM. Building on these foundations, recent research has moved toward multi-stage fine-tuning, which introduces progressively aligned training stages to enhance complex reasoning capabilities. The TULIP framework (Wang et al., 2023) , for example, defines a three-tier strategy involving general pretraining, task-aligned training, and scenario-specific refinement. Similarly, UL2 (Unified Language Learning) (Tay et al., 2022) merges diverse training paradigms—including prefix tuning, prompt tuning, and denoising—into a unified framework for cross-task generalization. Inspired by these approaches, our project employs a staged fine-tuning pipeline tailored to the hierarchical nature of TCM tasks: (1) syndrome differentiation, (2) disease diagnosis, and (3) prescription generation, each enhanced with targeted prompt engineering to infuse knowledge incrementally and ensure task-specific precision.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 369-375, Jinan, China, August 11-14, 2025.

2.3 Applications of LLMs in Healthcare and TCM

Traditional Chinese Medicine (TCM) presents unique challenges to NLP systems due to its reliance on ambiguous concepts, diverse linguistic styles, and highly contextual reasoning. TCM texts cover classical literature, modern clinical notes, and structured case records, with each type imposing distinct requirements for semantic comprehension. Existing studies have addressed entity recognition (Zhang et al., 2019), question answering (Li et al., 2024), and formula recommendation, with varying degrees of success. Despite these efforts, most approaches fall short in complex reasoning tasks such as syndrome-disease-prescription mapping.

The integration of LLMs introduces a promising shift in capabilities. With their strong generalization and reasoning capacities, LLMs can be leveraged for semantic understanding, clinical decision support, and knowledge inference in TCM contexts. By structuring the diagnostic process as a multi-turn dialogue and employing CoT prompting, models can now mirror expert-level clinical reasoning more closely.

Beyond TCM, LLMs have demonstrated substantial utility across broader medical applications. They have been employed for disease question answering, medical imaging report generation, and pharmacological recommendation. For example, Kiyak and Emekli studied the impact of structured prompting on medical multiple-choice QA tasks(Kıyak and Emekli, 2024), revealing that carefully designed prompts significantly enhance model performance. These developments collectively highlight the transformative potential of LLMs in both traditional and modern healthcare systems.

3 Methodology

3.1 Framework Overview

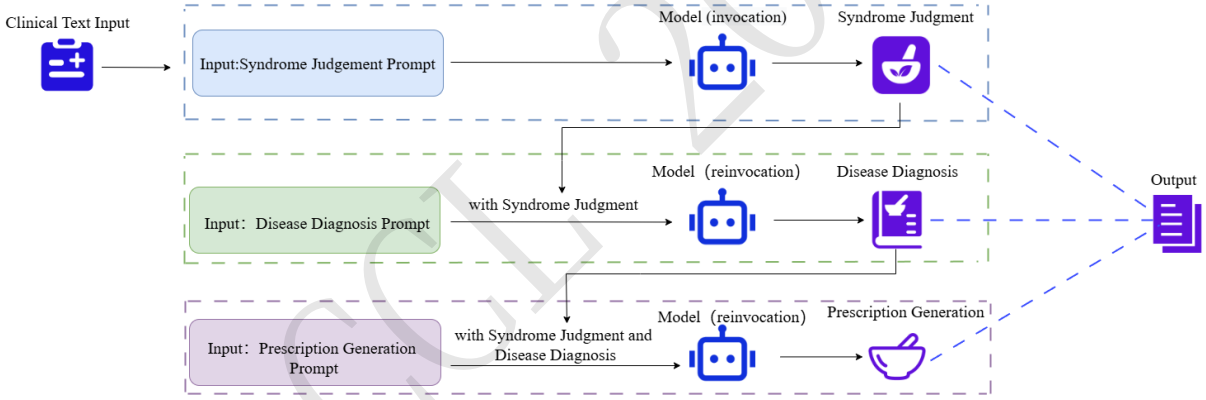

Figure 1: System Architecture Diagram

The system adopts a staged reasoning architecture to automate the Traditional Chinese Medicine (TCM) diagnostic process, leveraging both Chain-of-Thought (CoT) and multi-task learning principles. The overall workflow is depicted in Figure 1. Clinical texts first undergo the Syndrome Differentiation Step, where a diagnostic model performs initial reasoning to evaluate the patient's condition. The results from this step are then integrated into the Disease Diagnosis Step, which builds upon the initial findings and refines the diagnosis through secondary reasoning. Finally, the aggregated outputs from the previous steps are used in the Prescription Generation Step, where the system generates personalized treatment recommendations based on the comprehensive diagnostic context. Throughout the process, each step is closely interlinked, with outputs from one step serving as the context for the next, forming a coherent and continuous reasoning chain. This structure not only simulates the syndrome differentiation and treatment reasoning process in clinical TCM practice but also utilizes the benefits of multi-task learning, enabling the system to address various diagnostic tasks in a structured, interconnected, and integrated manner.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 369-375, Jinan, China, August 11-14, 2025.

3.2 Progressive Diagnostic Dialogue

The multi-turn dialogue prompt design adopts a progressive constraint strategy, incorporating intricate design principles aimed at ensuring output coherence and context consistency across the reasoning stages. As illustrated in Table 3.2, the design evolves through a structured series of constraints tailored to each stage of the diagnostic process. During the syndrome differentiation phase, output stability is achieved by narrowing the selection range (SYNDROMES) and enforcing a precise format specification (pipe-separated), which ensures controlled output variations and consistency. In the subsequent disease diagnosis phase, the model is provided with the prior response (response1) to enhance contextual consistency and to enforce a rigorous single-choice mechanism, ensuring that the system stays focused on the most probable diagnosis. Finally, during the prescription generation stage, a triple constraint framework is employed: a predefined herb whitelist (HERBS), a strict quantity control (10–15 herbs), and explicit hallucination prevention instructions. Throughout each stage, the prompts incorporate explicit role definitions (e.g., ”作为中医专家”in the first stage) and well-defined output format requirements. These structured instructions guide the model toward producing standardized results, promoting high levels of consistency, relevance, and clinical validity across the entire diagnostic workflow.

Phase	**Input Variables**	**Core Constraints**	**Example Prompt**	**Example Prompt (English)**
Syndrome Judgment	clinical_text, SYNDROMES	1-2 syndromes, pipe-separated, no empty values	作为中医专家，请根据以下临床资料识别证型：{clinical_text} 要求： 1. 从以下选项中选择1-2个证型：{SYNDROMES} 2. 用竖线分隔多个证型 3. 必须做选择 请直接回答证型判断。	As a TCM expert, identify syndromes based on: {clinical_text} Requirements: 1. Select 1-2 syndromes from: {SYNDROMES} 2. Separate multiple syndromes with vertical bar 3. Must make a selection Respond directly with syndromes.
Disease Diagnosis	response1, DISEASES	Single-select, strict list constraints	根据证型判断：{response1}，请诊断对应的疾病。 要求： 1. 必须从以下列表选择1项：{DISEASES} 2. 必须选择一个 请直接回答疾病名称。	Based on syndrome: {response1}, diagnose the disease. Requirements: 1. Must select one from: {DISEASES} 2. Must choose exactly one Respond directly with disease name.
Prescription Generation	response1, response2, HERBS	Herb whitelist, quantity limits	根据证候：{response1} 和疾病诊断：{response2}，开具处方。 要求： 1. 仅从以下药材选择：{HERBS} 2. 按标准格式输出 3. 列表外的药材不要多写 4. 药材数量10-15个 请直接回答药方建议。	Based on syndrome: {response1} and diagnosis: {response2}, prescribe formula. Requirements: 1. Select herbs only from: {HERBS} 2. Output in standard format 3. Do not include extra herbs 4. Use 10-15 herbs Respond directly with prescription.

Table 1: Multi-turn Dialogue Prompt Design Paradigm

3.3 Multi-task Fine-tuning

In this work, we adopt a multi-task fine-tuning approach, where the data from the three diagnostic steps, syndrome differentiation, disease diagnosis, and prescription generation, are combined and used to fine-tune a single model. This strategy leverages the strengths of multi-task learning by enabling the model to simultaneously learn from the intertwined relationships between these distinct yet complementary tasks. By training on a unified dataset that integrates all stages of the diagnostic process, the model

Proceedings of the 24th China National Conference on Computational Linguistics, pages 369-375, Jinan, China, August 11-14, 2025.

is able to develop a deeper understanding of the holistic nature of Traditional Chinese Medicine (TCM) diagnosis.

This multi-task learning approach offers several key advantages. First, it allows the model to benefit from the contextual interdependence between tasks, ensuring that the reasoning from one stage informs and enhances the performance of subsequent stages. Second, it improves the model's generalization ability by exposing it to a diverse range of tasks, thereby making it more robust to variations in input data and better equipped to handle complex, multi-faceted diagnostic scenarios. Furthermore, this approach reduces the need for task-specific fine-tuning, as the model is optimized to perform all relevant tasks simultaneously. This not only streamlines the training process but also leads to a more efficient and scalable solution. The integration of these tasks within a single model also enables a more cohesive and consistent output across the entire diagnostic pipeline, ensuring that the different stages are aligned and coherent in their reasoning.

4 Experiments and Results

4.1 Experimental Setup

We finetune the `Qwen2.5-7B-instruct` backbone model on an `NVIDIA GeForce RTX 3090` GPU platform with `PyTorch 2.0`, leveraging mixed precision training with `BF16` acceleration. The learning rate is set to $5e-5$ and the training epoch is set to 3 epochs.

4.2 Main Results

Team	Score1	Score2	Score (Average)
Team A	0.5830	0.4261	0.5045
Team B	0.5690	0.4306	0.4998
Team C	0.5550	0.4177	0.4863
Our Team	**0.5390**	**0.4189**	**0.4789**
Team D	0.5420	0.4006	0.4726
Team E	0.5550	0.3859	0.4705
Team F	0.5230	0.3845	0.4537

Table 2: Comparison of Results with Selected Teams

As can be seen from Table 4.2, our team achieved a strong position in this competition with an average score of 0.4789. However, our score of 0.5390 on Task 1 indicates a certain gap compared to some leading teams. Notably, in Task 2, our team scored 0.4189. Compared with other top - ranked teams, such as Team A (0.4261) and Team B (0.4306), the gap is relatively small. This suggests that our team's methods and strategies in the research and practice of this task are approaching the forefront. In the future, we will conduct in - depth analysis and optimization for the deficiencies in Task 1 to further enhance the overall competitiveness. Meanwhile, efforts will also be made to improve our performance in Task 2.

4.3 Error Analysis and Visualization

In this experiment, we employed the official evaluation methodology to separately calculate key performance metrics across different stages. Specifically, we measured the accuracy of syndrome judgement for the first stage, the accuracy of disease diagnosis under both correct and incorrect syndrome judgments, and overall accuracy of prescription generation across four scenarios combining correct/incorrect results from the first two stages. The detailed accuracy at each stage of the diagnostic process is shown in Figure 2.

The first-stage syndrome pattern identification exhibited relatively low accuracy (the accuracy of syndrome judgement = 0.29), which can be attributed to the inherent complexity of syndrome differentiation in the task. Unlike subsequent stages, this initial step requires nuanced discrimination among highly

Proceedings of the 24th China National Conference on Computational Linguistics, pages 369-375, Jinan, China, August 11-14, 2025.

similar traditional Chinese medicine syndromes, making it inherently challenging and contributing to the lower score. Notably, the first-stage syndrome accuracy had minimal impact on the second-stage disease diagnosis. Despite the low accuracy of syndrome judgement, the accuracy of disease diagnosis remained high in both scenarios: 0.93 when syndrome judgment was correct and 0.94 when incorrect. This robustness stems from the second stage's design as a single-choice task among only four disease categories, which limits the cascading effect of upstream errors and allows the model to maintain stable performance even with flawed syndrome inputs. In contrast, the third-stage prescription generation showed a clear dependency on the cumulative accuracy of the first two stages. The overall accuracy of prescription generation reached its peak (0.439) when both syndrome identification and disease diagnosis were correct, demonstrating the critical role of sequential correctness in downstream performance. Conversely, when both prior stages produced errors, overall accuracy of prescription generation dropped to its lowest level(0.350), highlighting the compounding effect of cascading inaccuracies. Even in intermediate cases—such as correct disease diagnosis following incorrect syndrome judgment—prescription accuracy (0.419) still suffered significantly compared to the fully correct scenario, underscoring the stage's sensitivity to upstream reliability.

In summary, while the second stage mitigates some first-stage errors due to its constrained task design, the third stage's performance is profoundly shaped by the cumulative accuracy of earlier phases. The low accuracy of syndrome judgement thus exerts a pronounced chain reaction on overall system performance, necessitating targeted improvements in the first stage's discriminative capabilities. Future work will prioritize optimizing syndrome pattern identification algorithms to enhance initial accuracy, while also refining disease diagnosis logic and prescription generation models to further strengthen the system's robustness and practical utility across all stages.

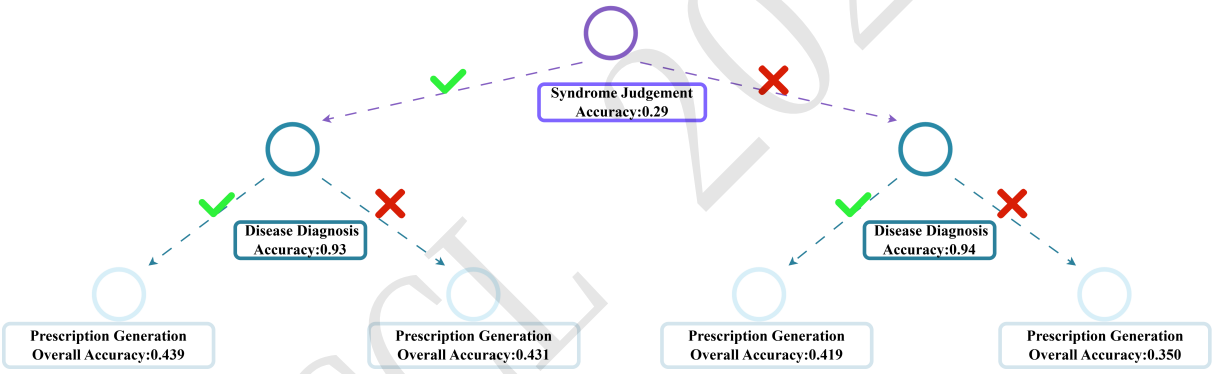

Figure 2: System Performance Flowchart showing the accuracy at each stage of the diagnostic process.

5 Conlusion and Future Work

This paper presents a TCM diagnostic framework that leverages multi-turn prompt optimization, progressive diagnostic dialogue, and multi-task fine-tuning. The proposed approach achieves exceptional performance in syndrome identification and prescription generation, effectively simulating TCM clinical thinking through context-aware multi-turn reasoning. The use of parameter-efficient fine-tuning enables deep domain adaptation without compromising the base model's capabilities. The framework demonstrates a reusable methodology and technical support for deploying large-scale TCM models in clinical practice.

While the proposed framework shows significant advantages, two key limitations remain: the 4K token window limits long-context processing, and the prescription generation module relies on a predefined candidate list, restricting its applicability in open scenarios. Future work will focus on extending the model's context window and enhancing the flexibility of the prescription generation module to move beyond predefined candidate lists, improving its adaptability in diverse clinical scenarios.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 369-375, Jinan, China, August 11-14, 2025.

Acknowledgements

This work was supported by the Open Project Program of Yunnan Key Laboratory of Intelligent Systems and Computing (Grant No. ISC24Y03) and the Yunnan Fundamental Research Project (Grant No. 202501AT070231).

References

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and others. 2022. *Chain-of-thought prompting elicits reasoning in large language models*. *Advances in neural information processing systems*, volume 35, pages 24824–24837.

Luciano Floridi and Massimo Chiriatti. 2020. *GPT-3: Its nature, scope, limits, and consequences*. *Minds and Machines*, volume 30, pages 681–694. Springer.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and others. 2023. *Gpt-4 technical report*. *arXiv preprint arXiv:2303.08774*.

Nazib Sorathiya, Chuan-An Lin, Daniel Xiong, Scott Zin, Yi Zhang, He Sarina Yang, and Sharon Xiaolei Huang. 2021. *Multi-turn dialog system on single-turn data in medical domain*. *arXiv preprint arXiv:2105.12887*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and others. 2022. *LoRA: Low-rank adaptation of large language models*. *ICLR*, volume 1(2), pages 3.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. *QLoRA: Efficient finetuning of quantized LLMs*. *Advances in neural information processing systems*, volume 36, pages 10088–10115.

Wenjian Wang, Lijuan Duan, Yuxi Wang, Junsong Fan, and Zhaoxiang Zhang. 2023. *MMT: cross domain few-shot learning via meta-memory transfer*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 45(12), pages 15018–15035. IEEE.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, and others. 2022. *UL2: Unifying language learning paradigms*. *arXiv preprint arXiv:2205.05131*.

Hong Zhang, Wandong Ni, Jing Li, Youlin Jiang, Kunjing Liu, and Zhaohui Ma. 2019. *On standardization of basic datasets of electronic medical records in traditional Chinese medicine*. *Computer Methods and Programs in Biomedicine*, volume 174, pages 65–70. Elsevier.

Ran Li, Gao Ren, Junfeng Yan, Beiji Zou, and Qingping Liu. 2024. *Intelligent question answering system for traditional Chinese medicine based on BSG deep learning model: taking prescription and Chinese materia medica as examples*. *Digital Chinese Medicine*, volume 7(1), pages 47–55. Elsevier.

Yavuz Selim Kıyak and Emre Emekli. 2024. *ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review*. *Postgraduate medical journal*, volume 100(1189), pages 858–865. Oxford University Press.