



Pathogenesis Reasoning Chain-of-thought Supervision for Large Language Models: Syndrome Manifestation Recognition and Multidimensional Evaluation in Spleen-stomach Disorders*

YANG Shu-Han^{1)**}, HU Yu-Xin^{2)**}, YU Xin-Yu¹⁾, TU Yu-Ying³⁾, ZANG Yi-Chang¹⁾, LI Pan-Fei^{1)***}

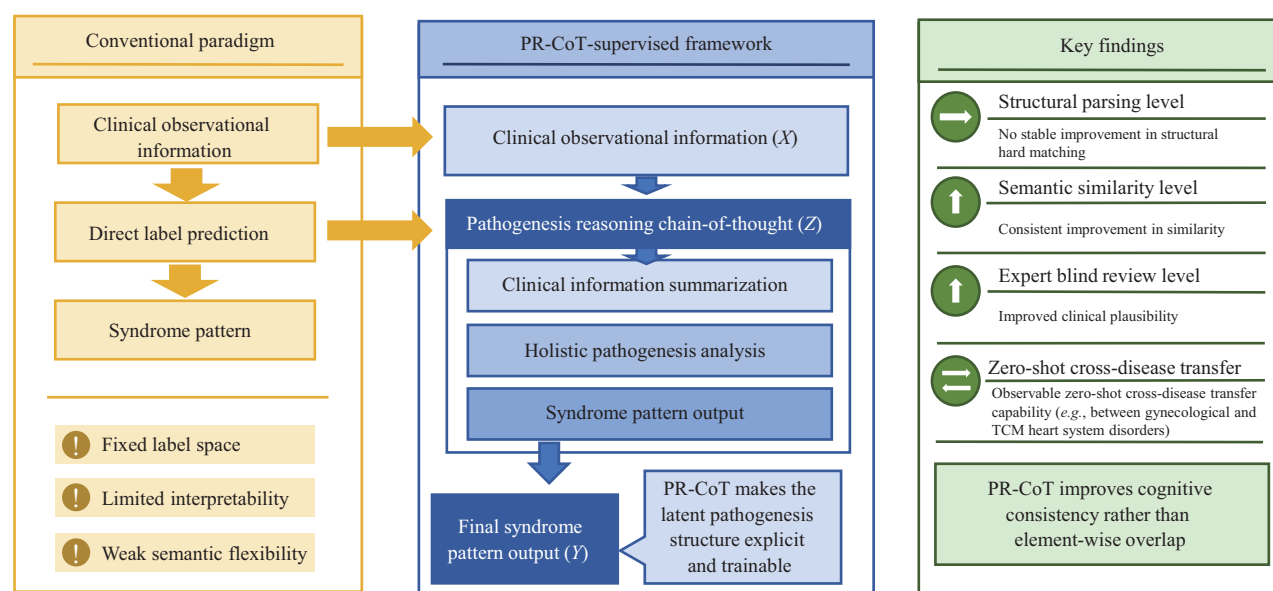
¹⁾School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, Hangzhou 310053, China;

²⁾College of Information Engineering, China Jiliang University, Hangzhou 310018, China;

³⁾Department of Hepatology, Xiyuan Hospital of China Academy of Chinese Medical Sciences, Beijing 100091, China)

Graphical abstract

From fixed-label classification to pathogenesis-oriented TCM syndrome differentiation



TCM syndrome differentiation is better understood as the generation and expression of latent pathogenesis structures

* This work was supported by grants from The National Natural Science Foundation of China (82505804), Major Science and Technology Co-construction Project of the National TCM Comprehensive Reform Demonstration Zone (GZY-KJS-ZJ-2025-018), the Natural Science Foundation of Zhejiang Province (QN26H270007), and University-level Scientific Research Project of Zhejiang Chinese Medical University (2024RCZXZK42).

** These authors contributed equally to this work.

*** Corresponding author.

Tel: 86-571-86613573, E-mail: lipanfei@zcmu.edu.cn

Received: March 6, 2026 Accepted: May 5, 2026

Abstract Objective The essence of syndrome manifestation recognition in traditional Chinese medicine (TCM) is to infer the body's latent pathogenesis state from clinical observational information, rather than to perform simple label matching. However, previous studies have largely modeled this task as syndrome pattern classification within a fixed label space, which does not adequately reflect the cognition process of TCM syndrome differentiation centered on pathogenesis reasoning, and is also insufficient to capture the openness, semantic variability, and cross-disease reusability of syndrome manifestation expression. This study aimed to investigate whether introducing pathogenesis reasoning chain-of-thought (PR-CoT) supervision into large language models (LLMs) could improve the quality and cognitive consistency of syndrome manifestation recognition and support cross-disease transfer. **Methods** Syndrome manifestation recognition was formulated as a conditional generation task under the framework of clinical observational information (X)→pathogenesis structure (Z)→syndrome pattern output (Y), where Z serves as an explicit intermediate structural variable linking the clinical evidence and syndrome judgment. Within this framework, a PR-CoT-supervised dataset for syndrome manifestation recognition was constructed based on medical case records of spleen-stomach disorders. After preprocessing, information extraction, manual proofreading, and data cleaning, the dataset comprised 4 800 training cases, 400 development cases, and 400 test cases. Each sample was annotated with a structured PR-CoT consisting of three progressive levels: clinical information summarization, comprehensive pathogenesis analysis, and syndrome pattern output. Supervised fine-tuning was conducted on open-source LLMs, with an end-to-end model serving as the baseline. Qwen3-32B was used as the primary experimental model, and Qwen3-14B as the scale comparison model. A progressive multidimensional evaluation framework was further established, comprising a structural parsing level, a semantic similarity level, and an expert blind review level. At the structural parsing level, syndrome pattern expressions were decomposed into structural elements and evaluated using Precision, Recall, F1 score, and Jaccard similarity. At the semantic similarity level, independent LLMs scored the theoretical proximity between predicted and reference syndrome patterns. At the expert blind review level, three TCM experts independently evaluated model outputs on two dimensions: syndrome differentiation consistency and terminology standardization of syndrome patterns. In addition, zero-shot cross-disease transfer evaluation was conducted on gynecological and heart-system disorder test sets. **Results** At the structural parsing level, PR-CoT supervision did not lead to a stable improvement in the element-wise overlap of syndrome pattern structural components. Compared with the corresponding baselines, neither Qwen3-32B nor Qwen3-14B showed consistent advantages in structural matching metrics after the introduction of PR-CoT supervision. In contrast, at the semantic similarity level, PR-CoT supervision produced stable positive gains across different model scales and evaluation systems. The average semantic score of Qwen3-32B increased from 6.425 8 in the baseline model to 6.585 0 after PR-CoT supervision, and that of Qwen3-14B increased from 5.870 0 to 5.964 2. At the expert blind review level, the overall score of Qwen3-32B (PR-CoT) was 7.026 0±0.107 7, higher than 6.416 3±0.288 9 for its baseline. In zero-shot cross-disease testing, the PR-CoT model still showed advantages in semantic evaluation and expert evaluation on both gynecological and heart-system disorder test sets, indicating a certain degree of transferability. **Conclusion** The benefits of PR-CoT supervision are mainly reflected in TCM semantic consistency and clinical plausibility, rather than in improved hard matching of structural elements. These findings support understanding syndrome manifestation recognition as a process of generating and expressing latent pathogenesis structures, rather than as a classification task within a traditional fixed label space. By introducing pathogenesis reasoning as an explicit intermediate structure into the modeling process and combining it with a progressive multidimensional evaluation framework, this study provides a methodological pathway for intelligent TCM syndrome differentiation that integrates theoretical alignment, interpretability, and multi-level evaluation.

Key words traditional Chinese medicine syndrome manifestation recognition, large language models, chain-of-thought, syndrome pattern evaluation, spleen-stomach disorders

DOI: 10.3724/j.pibb.2026.0141

CSTR: 32369.14.pibb.20260141

The core of syndrome differentiation and treatment in traditional Chinese medicine (TCM) lies in making an overall inference of the body's internal pathological state on the basis of the four diagnostic methods^[1-2]. This internal pathological state is termed syndrome (“证”), and its standardized naming form is referred to as a syndrome pattern (“证型”)^[3-4]. The reasoning process through which clinical information is integrated and analyzed until a syndrome pattern is established constitutes syndrome judgment (“证候判断”). Syndrome judgment is not only a key step in the clinical practice of syndrome differentiation and treatment, but also an important cognitive bridge linking clinical perceptual information with theoretical inference in TCM. In essence, a syndrome

is not a simple summary of symptoms and signs, but a comprehensive expression of the body's dysfunctional state, and its formation depends on the analysis and inference of pathogenesis relationships.

In clinical practice, syndrome judgment involves a certain degree of subjectivity, and different physicians may differ in both their diagnostic reasoning and conclusions; moreover, traditional manual syndrome differentiation is relatively limited in efficiency^[5-7]. With the development of medical informatization and intelligent healthcare, the use of artificial intelligence (AI) to assist or automate syndrome judgment has gradually become an important direction in intelligent TCM diagnosis research. However, previous studies have largely formulated

this problem as a supervised learning task within a fixed label space^[8-10], directly predicting syndrome pattern labels through machine learning^[11-13] or deep learning methods^[14-16], and most such studies have been confined to single-disease settings^[17-18]. Although these approaches can achieve a certain level of predictive performance on specific datasets, their modeling paradigm still has several structural limitations. First, single-disease label systems to some extent separate syndrome manifestations from their theoretical basis for cross-disease reuse, making them less suitable for the complex comorbidity and integrative syndrome differentiation scenarios encountered in real clinical practice. Second, directly predicting syndrome patterns without a structured representation of the pathogenesis reasoning process makes the model's decision path difficult to interpret^[19-20], which deviates to some extent from the theoretical characteristics of TCM, which emphasize causal logic and mechanistic analysis. Third, syndrome pattern expression itself is dynamic and combinatorial^[21]; therefore, a static classification framework cannot adequately capture its open-ended expression and semantic variation, thereby affecting the model's generalization ability and practical value.

From an informatics perspective, syndrome judgment is not merely a process of label matching, but rather one of comprehensive analysis and inference of latent pathological structures based on incomplete clinical information. Therefore, in the context of computational modeling, defining this task as syndrome manifestation recognition^[22], rather than syndrome pattern classification^[23-24], better reflects its cognitive connotation in TCM. The former emphasizes the inference and generation of latent pathogenesis structures, whereas the latter mainly concerns category assignment at the result level. How to represent the pathogenesis reasoning structure embedded in syndrome differentiation within a computational model has therefore become a key issue in TCM AI research.

In recent years, the rapid development of large language models (LLMs) has provided new technical conditions for overcoming the structural limitations described above. Operating in a unified semantic representation space, LLMs possess the ability to integrate knowledge across tasks and perform complex reasoning^[25], thereby providing a basis for modeling syndrome manifestations as a cognitive structure that can be reused across diseases. Meanwhile, the chain-of-thought mechanism can explicitly present intermediate reasoning steps, transforming the inferential process originally hidden

within model parameters into an observable logical trajectory, and thereby enhancing the structural interpretability of model decision-making^[26]. In light of the characteristics of TCM syndrome differentiation, syndrome judgment essentially depends on pathogenesis analysis and logical inference. Accordingly, introducing pathogenesis reasoning structures into the LLM framework may enable the model not only to output syndrome pattern conclusions but also to learn how those conclusions are derived through pathogenesis analysis. However, empirical studies are still lacking on whether pathogenesis reasoning chain-of-thought (PR-CoT) can systematically improve the quality of syndrome manifestation recognition.

In addition, in real clinical contexts, due to differences in academic schools and theoretical expression, the same pathological state often corresponds to multiple syndrome pattern descriptions that are semantically close but lexically different. Syndrome manifestation expression therefore exhibits clear openness and non-strict discreteness, distinguishing it from traditional closed-label classification problems and posing new requirements for model evaluation systems. Traditional evaluation approaches relying on metrics such as accuracy^[27], precision, recall^[28], and F1 score^[29-30] cannot fully capture the model's performance in terms of semantic similarity and clinical plausibility. It is therefore necessary to construct a multidimensional evaluation framework to assess model performance in syndrome manifestation recognition more systematically.

Against this theoretical and methodological background, the present study proposes an LLM-based framework for syndrome manifestation recognition guided by PR-CoT supervision, and conducts experiments using the clinically prevalent scenario of spleen-stomach disorders as a demonstration setting. Under a unified base-model condition, we systematically compare two strategies—end-to-end prediction and the introduction of chain-of-thought supervision—through rigorous controlled experiments across dimensions including structural consistency, semantic similarity, and expert overall evaluation, and further assess model generalization through zero-shot cross-disease testing. The core question addressed in this study is whether explicitly introducing pathogenesis reasoning structures can improve the quality and cognitive consistency of syndrome manifestation recognition within a unified semantic space.

By transforming syndrome manifestation recognition from a static classification problem into a

theory-aligned reasoning modeling task, this study seeks to provide a more interpretable and better-generalizing technical pathway for intelligent TCM syndrome differentiation, and to offer empirical support for introducing TCM-theoretical constraints into the LLM framework.

1 Materials and methods

1.1 Theoretical basis and task modeling

In this study, syndrome manifestation recognition is not understood as a classification task within a fixed label space, but rather as a generative process of

inferring and expressing latent pathogenesis structures. To support subsequent model construction and experimental design, this section develops a unified theoretical model from four aspects: cognitive generative mechanism, structural implementation, cross-disease sharing hypothesis, and task formalization. Clinical observational information is denoted as X , latent pathogenesis structure as Z , and syndrome pattern output as Y , so as to characterize the conditional generative relationship from observed information to diagnostic results. On this basis, a unified modeling framework based on PR-CoT supervision is constructed, as shown in Figure 1.

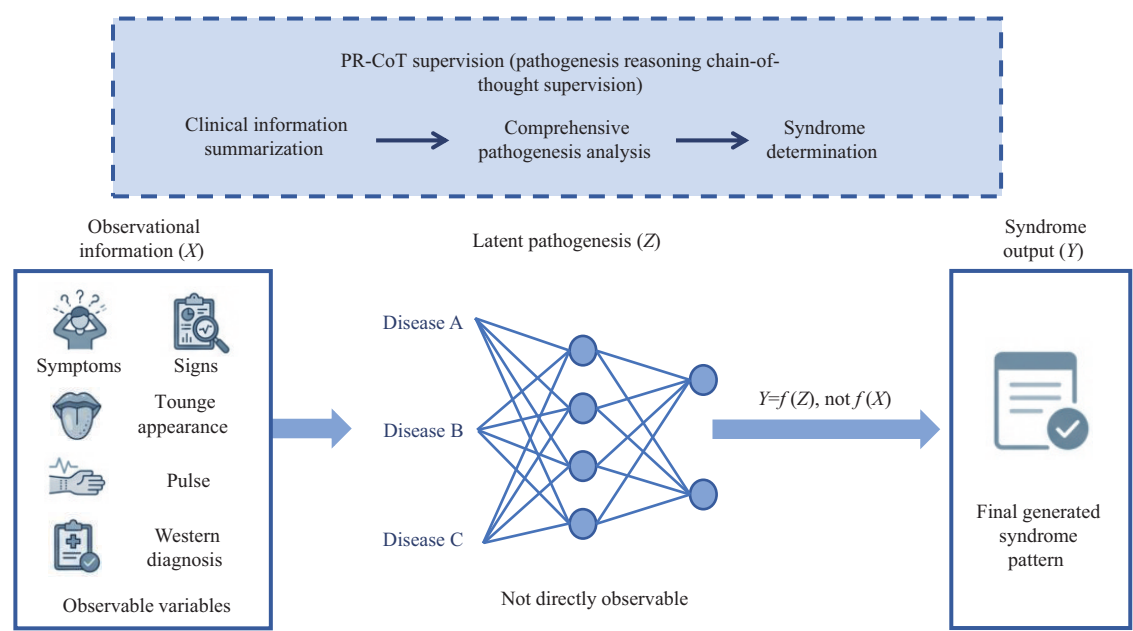


Fig. 1 Generative modeling framework for syndrome recognition based on the pathogenesis reasoning chain-of-thought (PR-CoT)

This figure illustrates the generative theoretical model of syndrome recognition based on PR-CoT. Clinical observational information (X) is treated as the conditional input, pathogenesis structure (Z) as the latent structural variable, and syndrome pattern output (Y) as the final generated result. In this framework, syndrome recognition is not modeled as direct label classification, but as a process in which the latent pathogenesis structure is inferred and then expressed. By jointly constraining the intermediate structure and the final output during training, the model enables explicit modeling of the PR-CoT process and provides the theoretical basis for subsequent methodological design and experiments.

1.1.1 Cognitive generative model of TCM syndrome differentiation

From the perspective of cognitive process, TCM syndrome differentiation can be formalized as a latent-variable inference process. Clinical symptoms and signs, tongue and pulse information, as well as related examinations, constitute observable variables, whereas pathogenesis structure belongs to a latent state that cannot be directly observed. Syndrome manifestation is neither a mechanical accumulation of symptoms nor a statistically defined feature cluster; rather, it is a structured expression of the body's

dysfunctional state and a compressed representation of pathogenesis within a specific diagnostic context.

Within this variable framework, the syndrome differentiation process can be understood as inferring Z from X , and then generating Y from Z . Rather than directly learning a simple mapping from clinical observations to syndrome pattern output, the model generates diagnostic results through inference of the intermediate pathogenesis structure. This perspective transforms TCM diagnosis from a problem of empirical label matching into one of structured reasoning, thereby providing the theoretical

foundation for subsequent generative modeling.

1.1.2 Pathogenesis reasoning chain-of-thought as the implementation mechanism of structural constraint

If syndrome manifestation recognition is formalized as a pathogenesis-driven latent generative process, the key issue is how to make the latent structure Z learnable, constrainable, and observable. Traditional end-to-end paradigms usually learn the mapping $X \rightarrow Y$ directly, while the intermediate pathogenesis reasoning path remains implicit in the parameter space, making it difficult to model and regulate explicitly and thus difficult to ensure logical consistency at the structural level.

In this study, PR-CoT supervision unfolds the originally implicit pathogenesis reasoning process into a trainable intermediate generative sequence, allowing the latent structural variable Z to enter the training constraint stage in textual form. Accordingly, the modeling path is transformed from $X \rightarrow Y$ into $X \rightarrow Z \rightarrow Y$. Under this framework, the model is required to first generate a pathogenesis analysis path and then produce the syndrome pattern output on the basis of that structure, while the intermediate structure is supervised through template constraints and expert annotations. Therefore, PR-CoT in this study is not merely a performance enhancement technique, but a key mechanism for implementing the hypothesis of structured modeling.

1.1.3 Semantic space of syndrome manifestation and the cross-disease sharing hypothesis

Traditional syndrome pattern research has mostly constructed closed label systems within a single disease, treating syndrome manifestations as disease-specific categories. However, according to the principle of “different diseases, same syndrome”, syndrome manifestation is theoretically reusable across diseases and is closer to a functional state pattern than to a label attached to a specific disease. From a structural perspective, the syndrome manifestation system is reusable, combinatorial, and hierarchical, forming a continuous and structured semantic space rather than a set of isolated discrete labels.

Based on this, the present study proposes a unified semantic space hypothesis: syndrome manifestations across different diseases share the same underlying pathogenesis-generating mechanism, whereas disease differences are mainly reflected at the level of the distribution of clinical observational information. This hypothesis provides the theoretical premise for the structured modeling of syndrome manifestation recognition and its zero-shot cross-disease transfer experiment.

1.1.4 Task formalization and modeling

On the basis of the above theoretical model and structural hypothesis, this study formulates syndrome manifestation recognition as a trainable conditional generation task. Specifically, during training, the intermediate structural variable Z and the final output Y are jointly supervised, so that the model both satisfies the requirement of structural consistency and achieves alignment between the syndrome pattern output and the reference result. In the implementation process, given clinical observational information X , the model first generates the pathogenesis reasoning chain Z , and then outputs the syndrome pattern Y on the basis of that structure, thereby transforming the abstract generative hypothesis into an executable computational process.

Compared with traditional end-to-end approaches, this task formulation emphasizes the explicit involvement of intermediate structure, with the aim of improving the logical consistency and stability of model outputs through structural constraints. At the evaluation stage, the final syndrome manifestation recognition result remains the core indicator, while the pathogenesis reasoning chain serves only as a structural constraint carrier during training and is not evaluated as an independent performance target.

Based on the above formulation, this study proposes the following testable hypothesis: under the unified semantic space hypothesis, introducing PR-CoT supervision can improve the cognitive consistency of syndrome manifestation recognition models and enhance their cross-disease generalization ability. Around this hypothesis, the study further constructs an integrated research system covering data construction, experimental strategies, and multidimensional evaluation (Figure 2).

1.2 Construction of a PR-CoT-supervised dataset for syndrome manifestation recognition in spleen–stomach disorders

1.2.1 Data sources and preprocessing

To validate the structural hypothesis described above, this study constructed a clinical case record dataset for syndrome manifestation recognition and pathogenesis reasoning chain learning in spleen–stomach disorders. The case records were mainly collected from the China National Knowledge Infrastructure (CNKI), published collections of medical cases, and works by contemporary renowned TCM physicians. In CNKI, systematic retrieval was conducted using disease names related to spleen–stomach disorders. To collect published case resources and contemporary clinical experience as

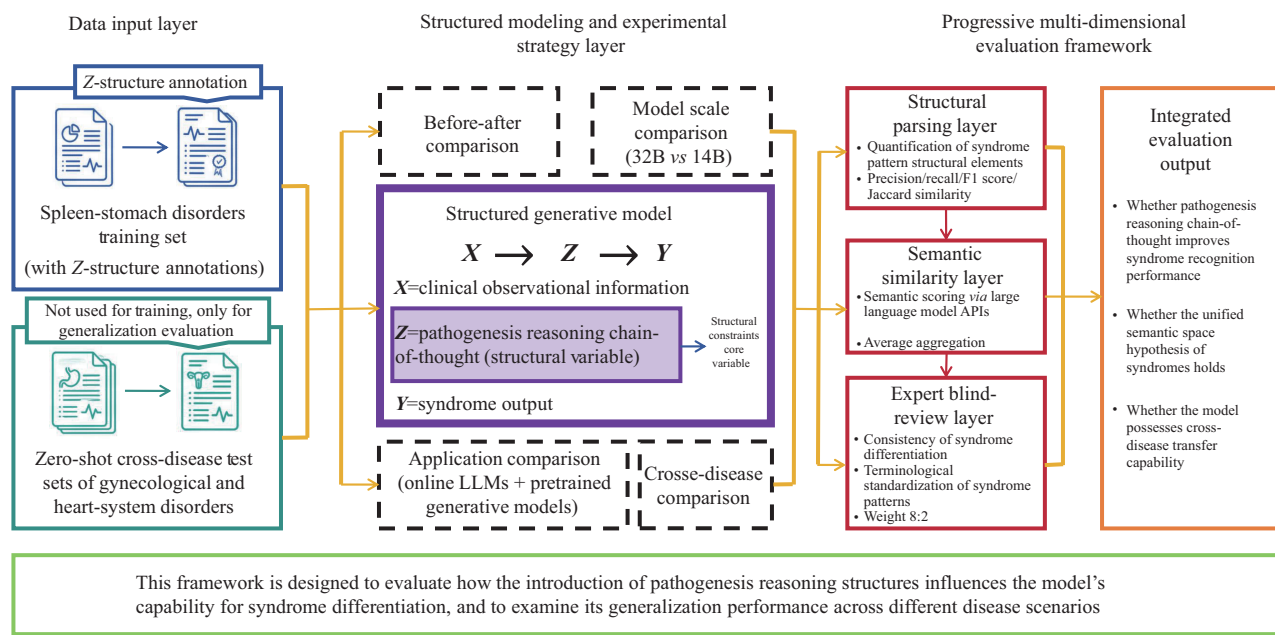


Fig. 2 Generative syndrome recognition and multidimensional evaluation framework based on PR-CoT supervision
This figure illustrates the overall technical pipeline and logical architecture of the study. The input layer consists of a spleen-stomach disorder training dataset annotated with structured pathogenesis information and zero-shot cross-disease test sets; the modeling layer is built on the $X \rightarrow Z \rightarrow Y$ generative framework, in which the pathogenesis reasoning chain (Z) is introduced as a key structural variable for supervised fine-tuning of the large language model; the experimental layer includes pre-/post-training comparison, model scale comparison, application comparison, and zero-shot cross-disease generalization comparison; and the evaluation layer comprises structural parsing, semantic similarity, and expert blind review, enabling a comprehensive assessment of model performance in terms of structural consistency, semantic equivalence, and clinical plausibility.

comprehensively as possible, we also referred to *A General Catalogue of TCM Books in the Sixty Years of New China*, the Duxiu database, and several highly recognized case series in the academic community, such as *Typical Medical Cases of Contemporary Senior TCM Physician* and *Series of One Hundred Modern Chinese Medical Clinicians*, for systematic literature retrieval and screening.

The inclusion criteria were as follows: the primary diagnosed disease belonged to spleen-stomach disorders, the case contained relatively complete clinical information, and the syndrome differentiation conclusion was explicitly provided. The exclusion criteria included incomplete texts lacking key symptoms or syndrome differentiation information, non-clinical popular science materials, and case records published before 1949.

During preprocessing, paper-based or scanned materials were first converted into text through optical character recognition (OCR) and then manually proofread. To meet the needs of syndrome manifestation recognition and pathogenesis reasoning modeling, a prompt template for large language models^[31] was designed to automatically extract key entities from the case records, including patient demographics (name, sex, and age), past medical

history, present illness history, symptoms and signs, tongue manifestations, pulse manifestations, Western medical diagnoses and examination findings, and syndrome patterns. To improve the accuracy and consistency of entity extraction, ERNIE-4.5-Turbo^[32] and GLM-4.5^[33] were first used to perform synchronous entity extraction on the text. When the extraction results from the two models were consistent, they were regarded as reliable and retained. When inconsistencies occurred, manual intervention was introduced for supplementary extraction and correction, followed by final quality control through dedicated expert review. The proofread information was then stored in a structured text format for subsequent dataset construction.

During data cleaning, duplicate entries were merged and samples with missing information or clearly abnormal records were removed. For synonymous expressions in symptom and syndrome pattern descriptions, the original semantics were retained in principle rather than being forcibly standardized, and only obvious spelling errors and incomplete expressions were corrected, so as to avoid the loss of syndrome differentiation semantics caused by over-standardization. After the above procedures, a high-quality clinical dataset of spleen-stomach

disorders was obtained, with information from the four diagnostic methods as input and syndrome patterns as output.

To support experiments on end-to-end syndrome manifestation recognition and supervised learning with pathogenesis reasoning chains, the spleen-stomach disorder dataset was split in a fixed manner while maintaining, as far as possible, a similar distribution of syndrome patterns across subsets. For the end-to-end syndrome manifestation recognition task, the dataset comprised 4 800 cases for training, 400 for development, and 400 for testing. Correspondingly, the pathogenesis reasoning chain dataset was aligned one-to-one with the case samples, on the same scale of 4 800 training samples, 400 development samples, and 400 test samples, thereby ensuring comparability between the two types of supervision signals in both sample size and data partitioning.

In addition, to evaluate the model's cross-disease transfer ability under the unified semantic space hypothesis, this study further constructed gynecological and heart-system disorder test sets. Using relevant disease names as retrieval constraints, 150 clinical case records were collected for each category from CNKI and processed according to the same preprocessing, extraction, and cleaning procedures described above. Neither dataset was used in model training or fine-tuning and was reserved exclusively for testing, in order to evaluate the model's zero-shot cross-disease generalization performance in different disease contexts after the introduction of PR-CoT supervision.

Through the experimental design of "training on related disease categories and testing across diseases", this study aims to determine whether the model has learned a syndrome manifestation recognition mechanism with cross-disease shared characteristics, rather than merely capturing statistical associations within a local label space.

1.2.2 Construction of the PR-CoT

Within the generative task framework described above, the PR-CoT serves not only as an important supervision signal for model training, but also as a core structural unit in dataset construction. This section describes its construction from three aspects: structural definition, annotation procedure, and quality control.

(1) Structural definition

In this study, the PR-CoT is defined as a structured representation of the TCM syndrome differentiation process, used to characterize the intermediate reasoning path from clinical

observational information to the syndrome pattern conclusion. Unlike traditional dataset construction that annotates only syndrome pattern labels, the present study explicitly encodes the diagnostic process at the data level, so that each sample contains not only the "syndrome differentiation result" but also the "basis for syndrome differentiation" that leads to that result.

Within the structural modeling framework, clinical observational information is denoted as X , the PR-CoT as the intermediate structural variable Z , and syndrome pattern output as Y . Here, Z functions as the key structure linking observational information and syndrome judgment, enabling the syndrome differentiation process, which is otherwise implicit in experience-based reasoning, to enter model training in a learnable and supervisable form.

A complete PR-CoT consists of three progressive levels (Figure 3), corresponding to the cognitive path of "clinical information summarization-comprehensive pathogenesis analysis-syndrome pattern output". First, at the clinical information summarization level, patient demographics, symptoms and signs, tongue manifestations, pulse manifestations, and relevant auxiliary examination findings are structurally organized, and key diagnostic information is extracted from the original descriptions to provide the basis for subsequent pathogenesis analysis. Second, at the comprehensive pathogenesis analysis level, based on the summarized information, systematic inferences are made regarding etiology, pathological location, pathological nature, pathological factors, and pathological states, which are then progressively integrated into an overall pathogenesis judgment; this is the core component of the reasoning chain. Finally, at the syndrome pattern output level, the final syndrome pattern conclusion is generated on the basis of the preceding pathogenesis analysis, thereby establishing a clear and traceable mapping between syndrome pattern expression and pathogenesis analysis.

Through the above design, the PR-CoT transforms the syndrome differentiation process of "symptom summarization-pathogenesis analysis-syndrome pattern formation" into an expressible structural sequence. This structure preserves the logical order of clinical syndrome differentiation while making the intermediate reasoning process a learnable, supervisable, and verifiable modeling unit. To further illustrate its data organization, Table 1 presents the structural format of a single training sample. Each sample consists of three parts: structured clinical observational information (X), a

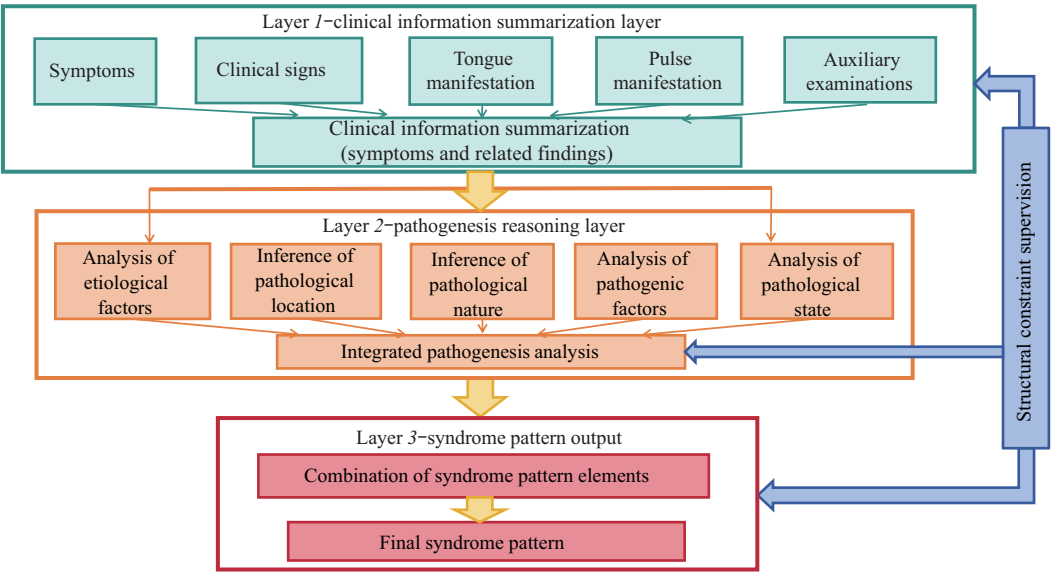


Fig. 3 Hierarchical structure and logical dependency of the PR-CoT

This figure illustrates the hierarchical organization and logical dependencies of the PR-CoT as the intermediate structural variable (Z). The first layer is the clinical information summarization layer, in which symptoms, clinical signs, tongue manifestations, pulse manifestations, and auxiliary examination findings are structurally integrated. The second layer is the pathogenesis reasoning layer, where inference of etiology, pathological location, pathological nature, pathological factors, and pathological states is conducted on the basis of the summarized information, and further integrated into an overall pathogenesis interpretation. The third layer is the syndrome pattern output layer, in which syndrome pattern elements are combined on the basis of the preceding pathogenesis analysis to generate the final syndrome pattern conclusion. Arrows in the figure indicate the semantic dependencies and structural mappings among different layers. PR-CoT supervision operates on intermediate structural nodes, transforming the pathogenesis reasoning process from merely explanatory text into a learnable and structurally constrained generative mechanism, and providing the data foundation for subsequent supervised training.

Table 1 Example structure of a single sample

Module	Example content
Input X	Sex: male; age: 28 years; symptoms and signs: vomiting immediately after eating, fullness and oppression in the chest and abdomen, fatigue and lack of strength, emaciation; tongue manifestation: pale tongue; pulse manifestation: weak pulse
Intermediate structure Z	Clinical information summarization: the patient’s chief manifestations were vomiting immediately after eating and fullness and oppression in the chest and abdomen, accompanied by fatigue, emaciation, a pale tongue, and a weak pulse, suggesting severe impairment of the spleen and stomach in transportation and transformation, insufficiency in the generation of qi and blood, and inadequate nourishment of the zang-fu organs Comprehensive pathogenesis analysis: vomiting immediately after eating was caused by deficiency-cold in the middle jiao, failure of the stomach to harmonize and descend, and adverse upward movement of qi; fullness and oppression in the chest and abdomen reflected insufficiency of spleen yang, impaired transportation and transformation, stagnation of water-grain essence in the middle jiao, and obstruction of qi movement; fatigue and emaciation indicated decline of yang qi, insufficiency in the generation of qi and blood, and inadequate nourishment of the body; a pale tongue and weak pulse signified insufficiency of yang qi and deficiency of qi and blood, failing to nourish the tongue upward and to fill the vessels. Taken together, the disease location was in the spleen and stomach, the disease nature was deficiency-cold, and the key pathogenesis was insufficiency of middle yang, with comprehensive decline in warming, motivating, and transforming functions
Output Y	Syndrome pattern of insufficiency of middle yang

hierarchical PR-CoT (Z), and syndrome pattern output (Y), all encoded according to a unified template for model training.

(2) Annotation procedure

The PR-CoT was constructed through a staged procedure of “expert-led, model-assisted, and expert-calibrated” annotation.

In the first stage, high-quality seed samples were constructed by experts with long-term experience in the clinical practice and teaching of spleen-stomach disorders. Under a unified template, the experts manually wrote standard reasoning chains to define structural levels, semantic granularity, and expression norms. The focus of this stage was to establish a

stable annotation paradigm rather than to pursue data scale.

In the second stage, based on the expert seed samples and the constraints of a unified template, raw medical case records equal in number to the seed samples and already annotated with syndrome patterns were selected as input, and an online large language model was used to structurally expand the reasoning chains and generate candidate pathogenesis reasoning chains in batch. During model selection, the research team organized a blind comparative evaluation by clinical experts and comprehensively assessed the performance of different models in terms of the consistency of TCM pathogenesis reasoning, the accuracy of syndrome manifestation summarization, and the standardization of TCM language expression. The results showed that ERNIE-4.5-Turbo performed better in TCM reasoning and language organization, and it was therefore selected as the candidate model for generating reasoning chains in the second stage. At this stage, the model served only an auxiliary role in structural expansion and language organization, strictly outputting hierarchical reasoning steps according to a fixed template, while retaining source markers and version information for the generated content.

In the third stage, experts reviewed and revised the candidate reasoning chains on a case-by-case basis. The review focused on whether the pathogenesis logic was self-consistent, whether the reasoning steps were complete, whether the syndrome pattern conclusion was consistent with the preceding analysis, and whether the language expression conformed to clinical usage. Samples with logical defects or semantic deviations were rewritten by experts or removed.

Through this procedure, the model played only an auxiliary role in content generation, whereas final annotation authority remained entirely in the hands of experts, thereby ensuring the cognitive reliability of the dataset.

(3) Review and quality control

To ensure data stability and reproducibility, this study established a multi-level quality control system. Each reasoning chain was cross-checked by at least two independent experts, with particular attention to the correspondence between pathogenesis nodes and symptom evidence, the rationality of reasoning order, and the logical support for the syndrome pattern conclusion. When disagreements arose, a third senior expert made the final decision, and the types of disagreement were recorded to continuously optimize the annotation criteria.

During dataset construction, template standardization, logical constraint checking, abnormal sample screening, and retrospective revision were continuously implemented. Through the dual safeguards of structural constraints and expert review, a PR-CoT dataset with adequate scale, logical completeness, and cognitive consistency was ultimately established, providing a reliable foundation for subsequent structurally supervised training of large language models.

1.3 Experimental design and comparison strategy

To systematically evaluate the role of PR-CoT supervision in syndrome manifestation recognition, this study constructed a hierarchical comparative experimental system. The main purpose was to examine whether structured training could lead to stable improvement in syndrome differentiation performance, and to assess the consistency and generalizability of such improvement across different model scales, application settings, and disease contexts.

With respect to model selection, Qwen3-32B^[34] from the Qwen series was adopted as the main experimental platform. Preliminary exploratory evaluation showed that this model performed relatively stably in TCM question answering and syndrome differentiation tasks, with strong semantic understanding and expressive capacity. Its 32B parameter scale also achieved a reasonable balance between performance and computational cost, making it suitable for controlled fine-tuning experiments on complex reasoning tasks. On this basis, Qwen3-14B was selected as the scale comparison model to analyze the effect of PR-CoT supervision under different parameter scales. Both models were general-purpose large language models in their initial state and had not received task-specific training in TCM pathogenesis. In this study, “pathogenesis reasoning chain-of-thought training” refers to supervised fine-tuning of the base model on the structured dataset, so that stepwise reasoning structures are internalized at the parameter level, rather than being induced only through superficial prompting. Except for the introduction of this structured training mechanism, all other aspects of model architecture and pretraining background were kept unchanged as far as possible, so that the observed experimental differences mainly arose from changes in the reasoning path.

To address the above objectives, four types of experiments were designed: pre- and post-training comparison, scale comparison, application comparison, and zero-shot cross-disease comparison.

First, a pre- and post-training comparison was

conducted. The model's performance before and after structured fine-tuning was compared to examine whether PR-CoT supervision could improve syndrome manifestation recognition performance and to verify the basic effectiveness of this mechanism.

Second, a scale comparison was performed. After both Qwen3-32B and Qwen3-14B had undergone structured training, their results were compared to determine whether the benefits of this mechanism could be observed across different model scales, thereby assessing whether its effect was stable across scales rather than merely dependent on parameter enlargement.

Third, an application comparison was conducted. Qwen3-32B after PR-CoT training was taken as the representative model and compared with current online general-purpose large language models, including OpenAI's GPT-5.2^[35], DeepSeek's DeepSeek-V3.2^[36], Baidu's ERNIE-4.5-Turbo, and Google's Gemini 3 Flash^[37]. All online models were tested within the same time window using a unified input format and consistent prompting strategy, without additional prompt optimization, so as to maintain comparability as far as possible. This comparison was intended to evaluate the practical competitiveness of the structured reasoning mechanism in real-world clinical assistance scenarios. In addition, Mengzi-T5-FT, a pretrained generative baseline from the pre-LLM era, was introduced as a supplementary comparison model. Considering that the original T5 was mainly pretrained on English corpora, this study adopted Mengzi-T5^[38], which had undergone continued pretraining on Chinese corpora, and fine-tuned it using the same training data and output format, in order to observe the performance differences between current large language models and earlier pretrained generative models.

Finally, a cross-disease zero-shot comparison was conducted. The experiments were extended to the test sets of gynecological and TCM heart-system disorders, using a setting in which the model was trained only on spleen-stomach disorder data and evaluated in a zero-shot manner on gynecological and heart-system disorder data. The performance changes of the structurally trained Qwen3-32B in new disease contexts were then observed to determine whether the model had learned a relatively stable way of generating and organizing pathogenesis, rather than merely memorizing disease-specific data characteristics, thereby testing its zero-shot cross-disease generalization ability.

Through the above hierarchical comparisons and cross-context validation, this study systematically

evaluated the performance differences associated with PR-CoT supervision across different model scales, disease contexts, and application environments. The specific evaluation metrics and multidimensional evaluation methods are described in the next section.

At the technical implementation level, model training and fine-tuning were performed in a high-performance computing environment using NVIDIA RTX 6000 (96 GB) GPUs, together with the LoRA parameter-efficient fine-tuning strategy implemented in the LLaMA-Factory framework, in order to improve the training efficiency and procedural stability of complex syndrome pattern reasoning tasks.

1.4 Design of a progressive multidimensional evaluation framework for syndrome manifestation recognition results

Because syndrome pattern expression in TCM is highly flexible and combinatorial, directly applying traditional classification metrics such as accuracy, recall, and F1 score through whole-string matching cannot adequately reflect model performance at the structural and semantic levels. To address this issue, this study constructed a progressive multidimensional evaluation framework consisting of a structural parsing level, a semantic similarity level, and an expert blind review level to conduct a systematic and interpretable quality assessment of syndrome manifestation recognition results (Figure 4). These three levels correspond respectively to structural consistency, semantic equivalence, and clinical plausibility, forming a progressive evaluation framework from objective computation to expert adjudication.

(1) Multidimensional parsing and quantitative evaluation based on structural elements of syndrome patterns

To transform flexible and diverse syndrome pattern expressions into computable objects, this study formulated a set of multidimensional parsing rules for syndrome pattern structural elements on the basis of the theoretical principles concerning syndrome pattern composition in *Diagnostics of Traditional Chinese Medicine*^[39] and *Syndrome Element Differentiation*^[40], together with expert discussion. Syndrome pattern expressions were decomposed into five relatively independent structural dimensions: disease location (DL), disease nature (DN), pathological factor (PF), pathological state (PS), and other elements (OE). Two researchers independently performed the parsing according to the unified rules, and discrepancies were resolved through discussion to ensure the stability of the parsing standard. Through this structured decomposition,

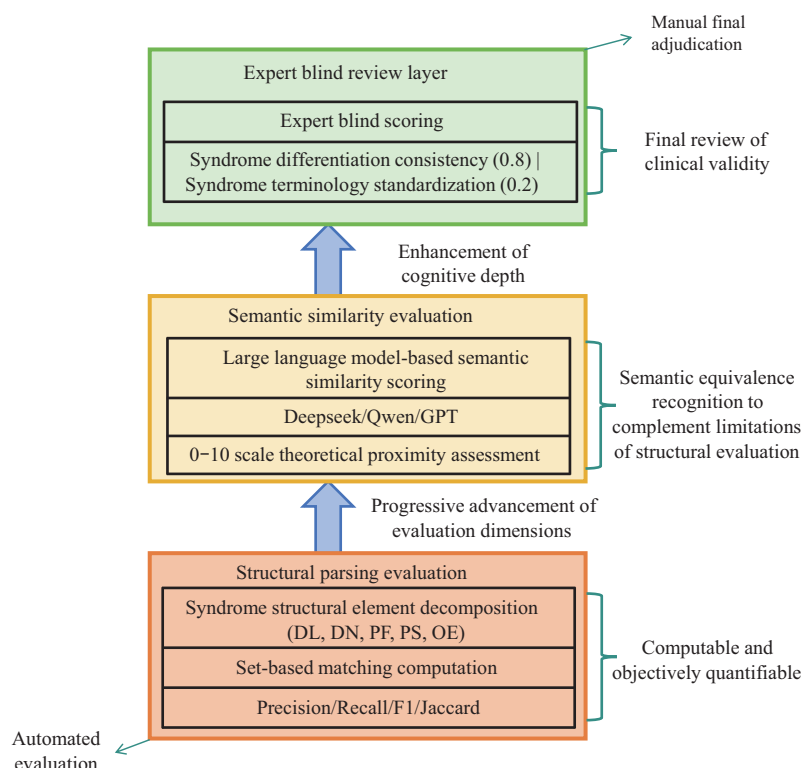


Fig. 4 Architecture of the progressive multidimensional evaluation framework for syndrome manifestation recognition

This figure illustrates a progressive multidimensional evaluation framework for syndrome manifestation recognition results, comprising three levels: the structural parsing level, the semantic similarity level, and the expert blind review level. The structural parsing level measures structural consistency through decomposition of syndrome pattern structural elements and set-based matching; the semantic similarity level evaluates the theoretical proximity between model outputs and the gold standard using independent large language models; and the expert blind review level provides final adjudication from two dimensions, namely syndrome differentiation consistency and terminology standardization of syndrome patterns. Together, the three levels form a progressive pathway from automated computation to expert adjudication, enabling an integrated assessment of structural consistency, semantic equivalence, and clinical plausibility.

syndrome pattern expressions in natural language could be mapped into multidimensional structural element sets. For example, “liver depression and qi stagnation, with cold in the middle jiao” (“肝郁气滞, 中焦有寒”) could be parsed as: $S = \{(DL = \{\text{liver (肝), qi (气), middle jiao (中焦)}\}), (DN = \{\text{cold (寒)}\}), (PF = \emptyset), (PS = \{\text{depression (郁), stagnation (滞)}\}), (OE = \emptyset)\}$.

On this basis, the structural element set generated by the model was matched with the gold-standard set dimension by dimension, and metrics including Precision, Recall, F1 score, and Jaccard similarity were calculated. To ensure objectivity and computability, the five dimensions were treated with equal weights at the structural level, without introducing subjective weight assignment. Specifically, Precision reflects the accuracy of structural expression, Recall measures the coverage of gold-standard elements, F1 score balances accuracy and completeness, and Jaccard similarity quantifies overall overlap through the intersection-over-union ratio of sets, making it particularly suitable for analyzing structural consistency in multi-element

combination scenarios. Through structural parsing and set-based matching, syndrome pattern expressions that were originally difficult to compare directly were transformed into computable multidimensional objects, thereby enabling quantitative evaluation at the structural level.

(2) TCM semantic similarity evaluation of syndrome patterns based on large language models

Although evaluation at the structural parsing level has advantages in computability and interpretability, syndrome pattern expression in real TCM contexts often involves the phenomenon of semantic equivalence with lexical variation. For example, “deficiency-cold of the spleen and stomach with insufficiency of qi and blood” and “yang deficiency of the middle jiao with insufficiency of transformation and generation” are highly similar in theoretical meaning, yet may not show a high degree of overlap after decomposition into structural elements. Therefore, relying solely on structural consistency metrics may underestimate the validity of semantically equivalent expressions.

To compensate for the limitations of structural parsing in identifying semantic equivalence, this study introduced application programming interfaces of mainstream high-performance large language models at the semantic similarity level for scoring. Based on preliminary consistency testing and comparative analysis, Qwen3.5-Plus^[41], DeepSeek-V3.2, and OpenAI GPT-5.1 were selected as the semantic evaluation tools. None of these models participated in the construction of the training data used in this study, and all were independent of the parameter systems of the experimental models, thereby reducing the risk of homology bias.

To ensure evaluation consistency, this study adopted a unified prompt template and a unified evaluation procedure, and completed batch evaluation within the same time window. Specifically, for each medical case, the syndrome pattern expressions generated by different models were collected into a single evaluation unit, and integer scores on a 0–10 scale were assigned in parallel within a single evaluation round, so as to minimize systematic bias caused by cross-round and contextual differences. No additional prompt optimization or iterative tuning was performed during evaluation. The scoring criterion was the degree of semantic equivalence between the predicted syndrome pattern and the gold-standard syndrome pattern at the level of essential TCM pathogenesis meaning. To reduce the potential bias introduced by any single evaluation model, the results

from the three models were aggregated for analysis, thereby measuring the semantic proximity of syndrome pattern expressions in terms of theoretical connotation and complementing the limitations of the structural parsing level in equivalence recognition.

(3) Comprehensive quality evaluation of syndrome patterns based on expert blind review

Although structural parsing and semantic similarity evaluation have advantages such as high efficiency and batch applicability, both may still be limited in their assessment of overall depth of pathogenesis understanding, implicit logical consistency, and terminological standardization. To enhance the clinical relevance and credibility of the evaluation results, this study further introduced expert blind review as the final criterion for quality assessment.

An expert rating rubric for syndrome pattern quality was designed in this study, evaluating the outputs from two dimensions: syndrome differentiation consistency and terminology standardization of syndrome patterns (Table 2). The former was used to assess the overall consistency between model outputs and the gold standard in terms of syndrome differentiation results and structural logic, whereas the latter was used to assess whether the syndrome pattern expression conformed to the terminological norms and structural conventions of TCM. According to expert consultation, the two dimensions were weighted at 8 : 2.

Table 2 Expert rating rubric for syndrome pattern quality

Evaluation dimension	Evaluation criteria	Weight score
Syndrome differentiation consistency	Used to evaluate the degree of equivalence between the model output and the gold standard at the level of essential TCM pathogenesis. A score of 10 indicates complete overall pathogenesis consistency, with differences only in wording; 8–9 indicates consistency in the core pathogenesis with only slight differences or missing elements; 6–7 indicates partial pathogenesis consistency with some deviation; 4–5 indicates obvious local conflicts or insufficient directional consistency; 1–3 indicates an incorrect core pathogenesis judgment or an opposite directional judgment; and 0 indicates a severe deviation from TCM theory, obvious hallucination, or irrelevance to the original information. Scoring in this dimension is primarily based on overall pathogenesis consistency	0.8
Terminology standardization of syndrome patterns	Used to evaluate whether the syndrome pattern expression conforms to TCM terminological norms and common expression structures. A score of 10 indicates the use of standard syndrome pattern terminology with rigorous and well-standardized expression; 8–9 indicates generally standardized expression with only slight lack of conciseness; 6–7 indicates basically correct terminology but an insufficiently standardized structure; 4–5 indicates confusing expression or non-standard combinations; 1–3 indicates obvious non-standard wording or inclusion of non-TCM terminology; and 0 indicates that the expression is difficult to recognize as a TCM syndrome pattern. This dimension does not judge the correctness of pathogenesis, but only the degree of terminological standardization	0.2

In practice, three TCM experts with long-term experience in the diagnosis and treatment of spleen-stomach disorders were invited to independently conduct a blind evaluation of the outputs generated by each model, without knowing their model sources. The final score was calculated as the average of the

three experts' ratings and then aggregated according to the predefined weights.

Through the progressive evaluation across the structural parsing level, semantic similarity level, and expert blind review level, this study ensured the efficiency of automated assessment while introducing expert final adjudication to confirm output quality, thereby constructing an evaluation system for syndrome manifestation recognition that integrates computability, semantic sensitivity, and clinical plausibility.

2 Results and analysis

2.1 Quantitative evaluation based on structural elements of syndrome patterns

To examine the effect of PR-CoT supervision at the structural parsing level, this study conducted a quantitative evaluation based on the structural elements of syndrome patterns. The analysis focused on element-level structural overlap and did not involve judgments of semantic equivalence or clinical plausibility.

Table 3 shows that the overall differences among models at the structural parsing level were not large. Among the online models, Gemini 3 Flash performed relatively well on Precision and Jaccard, DeepSeek-V3.2-Think achieved relatively high Recall, and ERNIE-4.5-Turbo remained in a relatively high range across the metrics overall. As a pretrained generative baseline from the pre-LLM era, Mengzi-T5-FT

exhibited the characteristic of “relatively high hit rate but insufficient coverage”, suggesting that its overall coverage of structural elements was relatively limited.

Table 3 Syndrome pattern structural element matching results across models in the structural parsing layer

Model	Precision	Recall	F1	Jaccard
Qwen3-32B (PR-CoT)	0.460 2	0.460 9	0.446 2	0.325 3
Qwen3-32B (Baseline)	0.480 6	0.475 3	0.461 9	0.343 5
Qwen3-14B (PR-CoT)	0.436 2	0.427 8	0.414 8	0.295 4
Qwen3-14B (baseline)	0.437 4	0.453 6	0.428 9	0.309 2
GPT-5.2	0.431 0	0.456 1	0.429 6	0.307 9
DeepSeek-V3.2-Think	0.477 8	0.494 2	0.470 2	0.345 0
ERNIE-4.5-Turbo	0.483 6	0.486 6	0.471 6	0.346 8
Gemini 3 Flash	0.540 9	0.451 4	0.477 6	0.358 7
Mengzi-T5-FT	0.508 7	0.296 5	0.357 2	0.242 0

The pre- and post-training comparison within the same model scale showed that after introducing PR-CoT supervision, neither Qwen3-32B nor Qwen3-14B exhibited consistent improvement at the structural parsing level; compared with their corresponding baselines, the PR-CoT models were generally lower on multiple structural metrics. The incremental plot (Figure 5) further shows that the structural metric increments of PR-CoT-baseline were overall negative for both model scales, suggesting that PR-CoT supervision did not demonstrate a stable advantage at the structural parsing level.

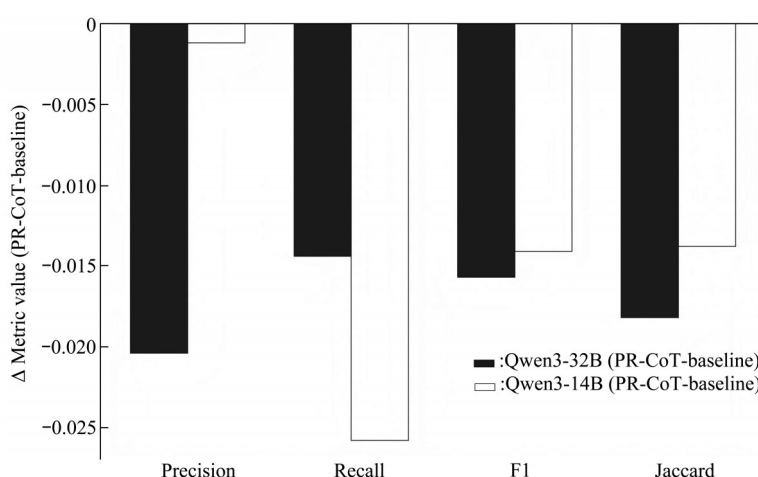


Fig. 5 Metric increments in the structural parsing layer (Δ =PR-CoT-baseline)

This figure shows the increments in Precision, Recall, F1 score, and Jaccard similarity for Qwen3-32B and Qwen3-14B.

It should be noted that evaluation at the structural parsing level is based on element-by-element matching of discrete element sets, emphasizing

explicit lexical overlap and element coverage. It mainly reflects the formal consistency of syndrome pattern expression, rather than directly measuring the

semantic equivalence and expressive flexibility of different syndrome pattern formulations at the level of theoretical meaning. In TCM, syndrome patterns are often compressed expressions of the overall pathogenesis structure. Different formulations may be highly similar in theoretical essence, yet may not produce high overlap after structural decomposition. Therefore, the absence of improvement at the structural level does not necessarily mean that PR-CoT supervision is ineffective.

Taken together, these results suggest that PR-CoT training may not strengthen the “hard matching” of structural elements, but may instead encourage the model to focus more on the organization of pathogenesis logic and the expression of overall semantics. On this basis, the next section further examines whether it demonstrates advantages in cognitive consistency at a higher level through semantic similarity evaluation.

2.2 TCM semantic similarity evaluation of syndrome patterns based on large language models

To further assess the extent to which model outputs approximated TCM syndrome differentiation conclusions at the level of theoretical meaning, this study introduced independent large language models (DeepSeek-V3.2, Qwen3.5-Plus, and GPT-5.1) to score the theoretical consistency of syndrome pattern expressions at the semantic similarity level, in addition to the structural parsing evaluation. None of these evaluation models participated in the training process of this study, thereby enhancing the relative independence of the evaluation.

Table 4 shows that, except for Mengzi-T5-FT, the semantic scores of the other models were mainly distributed in the range of 5.8–6.9. Among them, GPT-5.2 achieved the highest average score (6.896 7), while ERNIE-4.5-Turbo and Qwen3-32B (PR-CoT) were also ranked relatively high. In contrast, as a pretrained generative baseline from the pre-LLM era, Mengzi-T5-FT scored markedly lower across all three evaluation systems (average 2.552 5), indicating a considerable gap between it and current large language models in terms of the theoretical closeness of syndrome pattern expression.

The pre- and post-training comparison within the same model scale showed that, unlike the results at the structural parsing level, the Qwen series models all exhibited positive increments across the three semantic scoring systems after the introduction of PR-CoT supervision. Compared with its baseline, Qwen3-32B (PR-CoT) improved under all three scoring systems, with the average score increasing from 6.425 8 to 6.585 0; Qwen3-14B also showed consistent

Table 4 Semantic similarity layer scoring results

Model	Qwen3.5-	Deepseek-	GPT-5.1	Average
	Plus	V3.2		
Qwen3-32B (PR-CoT)	6.922 5	6.312 5	6.520 0	6.585 0
Qwen3-32B (Baseline)	6.740 0	6.022 5	6.515 0	6.425 8
Qwen3-14B (PR-CoT)	6.312 5	5.475 0	6.105 0	5.964 2
Qwen3-14B (Baseline)	6.095 0	5.422 5	6.092 5	5.870 0
GPT-5.2	7.127 5	6.630 0	6.932 5	6.896 7
DeepSeek-V3.2-Think	6.787 5	6.197 5	6.610 0	6.531 7
ERNIE-4.5-Turbo	6.837 5	6.310 0	6.655 0	6.600 8
Gemini 3 Flash	6.757 5	5.782 5	6.292 5	6.277 5
Mengzi-T5-FT	2.962 5	2.207 5	2.487 5	2.552 5

improvement, with the average score increasing from 5.870 0 to 5.964 2. To present the differences more intuitively, this study further calculated the semantic increment of PR-CoT relative to Baseline (Δ =PR-CoT-baseline) and plotted the incremental chart (Figure 6). The results show that the Δ values were positive across all three evaluation systems for both model scales, among which the DeepSeek-V3.2 scoring system showed the most pronounced increment for Qwen3-32B (+0.290 0), suggesting that this benefit was to some extent consistent across different evaluation systems and model scales.

These results stand in contrast to the element-matching performance observed at the structural parsing level: although PR-CoT supervision did not significantly improve the element-wise overlap of syndrome pattern structural elements, it showed a sustained advantage in terms of consistency of theoretical meaning. This suggests that after pathogenesis reasoning training, the model’s organization of syndrome differentiation logic and overall grasp of pathogenesis may become closer to the TCM theoretical framework. In other words, the benefits of PR-CoT supervision are more likely to be reflected in the semantic coherence of syndrome pattern expression and the plausibility of pathogenesis inference, rather than in the mechanical alignment of structural terms.

Based on the above findings, the next section further combines expert blind review to examine whether the advantage observed at the semantic level can be translated into overall clinical acceptance at the final adjudication level, from the two dimensions of syndrome differentiation consistency and terminology standardization of syndrome patterns.

2.3 Comprehensive quality evaluation of syndrome patterns based on expert review

To further evaluate the quality of model outputs

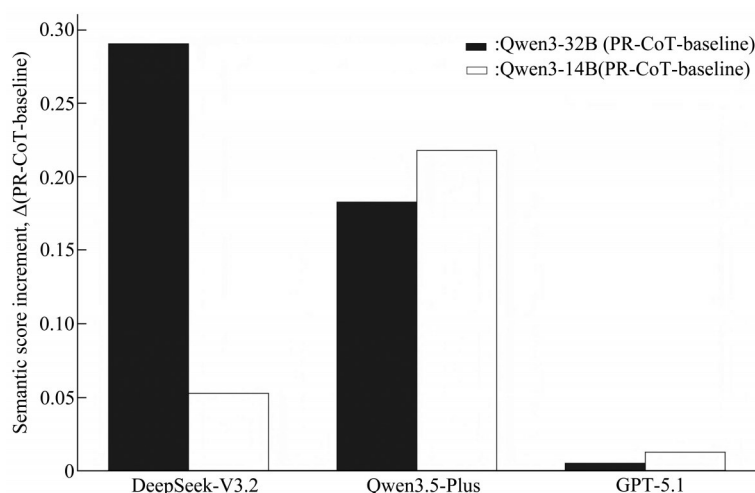


Fig. 6 PR-CoT-baseline increments in syndrome pattern semantic similarity (Δ =PR-CoT-baseline)

Semantic score increments for Qwen3-32B and Qwen3-14B evaluated by three independent models (DeepSeek-V3.2, Qwen3.5-Plus, and GPT-5.1).

from the perspective of clinical syndrome differentiation, this study invited three TCM experts with long-term experience in the diagnosis and treatment of spleen-stomach disorders to conduct independent blind reviews without knowledge of the model sources. The scoring involved two dimensions, syndrome differentiation consistency and terminology standardization of syndrome patterns, and the overall score was calculated according to the predefined weights.

Because expert blind review is a high-cost manual evaluation procedure, and because the preceding structural parsing and semantic similarity results had already shown generally consistent trends across different model scales, this study focused at the expert review stage on the main experimental model Qwen3-32B and its end-to-end baseline model, while selecting representative online models and a pretrained generative baseline as controls, in order to examine the practical effect of PR-CoT supervision at the level of final clinical adjudication.

Table 5 summarizes the scores assigned by the

three experts on the two dimensions and the overall statistical results, while Figure 7 further presents the overall scores of each model in the form of mean $\pm$ standard deviation (SD). Overall, Qwen3-32B (PR-CoT) achieved an average overall score of 7.026 0 $\pm$ 0.107 7, which was higher than its baseline (6.416 3 $\pm$ 0.288 9), suggesting that PR-CoT supervision brought a relatively stable gain at the level of expert final review. Among the online models, GPT-5.2 and DeepSeek-V3.2-Think obtained overall scores close to that of Qwen3-32B (PR-CoT), namely 7.023 5 and 6.938 0, respectively, but their standard deviations were relatively larger, especially for DeepSeek-V3.2-Think (0.248 5), indicating slightly weaker agreement among experts in their evaluations. In contrast, as a pretrained generative baseline from the pre-LLM era, Mengzi-T5-FT had a clearly lower overall score (4.006 8 $\pm$ 0.350 0), indicating that it still lagged considerably behind current large language models in terms of clinical syndrome differentiation consistency and syndrome pattern expression quality.

Table 5 Overall syndrome pattern quality scores in the expert blind-review layer

	Qwen3-32B (PR-CoT)	Qwen3-32B (baseline)	GPT-5.2	DeepSeek- V3.2-Think	Mengzi-T5-FT
Expert 1 Consistency of syndrome differentiation	6.775 0	6.135 0	7.010 0	6.815 0	3.827 5
Terminological standardization of syndrome patterns	8.067 5	8.375 0	7.032 5	8.677 5	6.632 5
Expert 2 Consistency of syndrome differentiation	6.567 5	5.977 5	6.880 0	6.317 5	3.087 5
Terminological standardization of syndrome patterns	8.182 5	6.140 0	6.865 0	7.725 0	5.365 0
Expert 3 Consistency of syndrome differentiation	6.905 0	6.252 5	7.230 0	6.680 0	3.507 5
Terminological standardization of syndrome patterns	8.150 0	8.270 0	6.975 0	8.417 5	6.415 0
Mean $\pm$ SD	7.026 0 $\pm$ 0.107 7	6.416 3 $\pm$ 0.288 9	7.023 5 $\pm$ 0.123 5	6.938 0 $\pm$ 0.248 5	4.006 8 $\pm$ 0.350 0

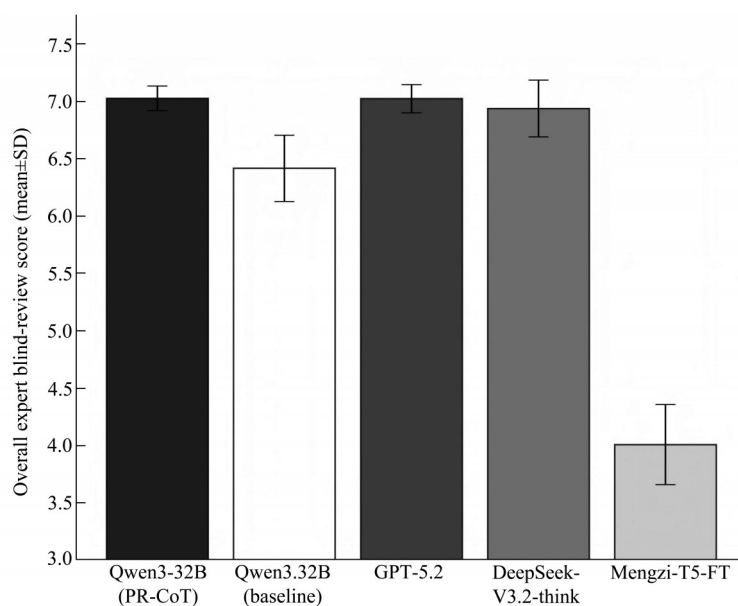


Fig. 7 Overall expert blind-review score (mean±SD)

From the dimension-specific results, the PR-CoT model was overall superior to the baseline in the dimension of syndrome differentiation consistency, whereas in the dimension of terminology standardization of syndrome patterns, the baseline was slightly superior in the ratings of some individual experts. This suggests that PR-CoT supervision is more likely to enhance the model's grasp of overall pathogenesis logic and syndrome differentiation direction, but does not necessarily improve the standardization of syndrome pattern terminology at the same time.

Taken together with the results of the previous two sections, a relatively clear hierarchical progression emerges: no stable improvement was observed at the structural parsing level, consistent positive gains appeared at the semantic similarity level, and the expert blind review level further supported the effectiveness of PR-CoT supervision from the perspective of final clinical adjudication. Overall, the improvement brought by this mechanism is more likely to be reflected in the overall rationality of syndrome differentiation reasoning and the consistency of pathogenesis expression, rather than in the mechanical matching of structural elements.

From the perspective of TCM cognition, the core of syndrome differentiation lies in the integrated judgment of pathogenesis relationships, rather than in the simple concatenation of terminological elements. The expert blind review results support this point to some extent: when the pathogenesis reasoning path is

explicitly introduced into model training, the outputs are more likely to gain clinical acceptance at the level of overall syndrome differentiation logic. Accordingly, the expert blind review level not only provides clinical support for the advantages observed at the semantic similarity level, but also further strengthens the theoretical explanatory power of understanding syndrome manifestation recognition as a process of generating latent pathogenesis structures.

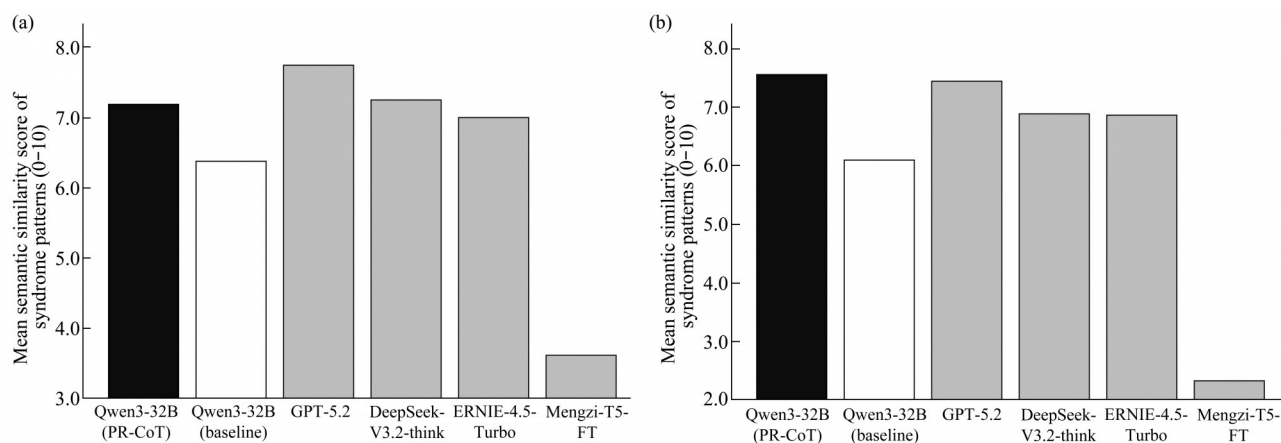
2.4 Zero-shot cross-disease transfer results of syndrome patterns in gynecological and heart-system disorders

To determine whether the model had learned a pathogenesis reasoning pattern that could be reused across diseases, this study further conducted a zero-shot cross-disease transfer evaluation on independently constructed test sets of gynecological and heart-system disorder medical cases. Neither of these two test datasets participated in the model training; they were used only for cross-disease generalization validation. Because the focus of this section is transferability, the complete layered evaluation procedure is not repeated here. Instead, the results of the structural parsing level, semantic similarity level, and expert blind review level are summarized comprehensively (Table 6), and a bar chart is used to present the average semantic similarity scores of each model on the two cross-disease test sets (Figure 8).

At the structural parsing level, the overall structural matching performance in both cross-disease

Table 6 Summary of model transfer performance on cross-disease gynecological and TCM heart-system test sets

Model	Gynecological disorder			TCM heart-system disorder		
	Jaccard	Mean semantic score	Mean expert score	Jaccard	Mean semantic score	Mean expert score
Qwen3-32B (PR-CoT)	0.270 0	7.188 9	6.636 0	0.252 6	7.560 0	6.158 2
Qwen3-32B (baseline)	0.327 3	6.377 8	6.216 4	0.294 8	6.097 8	5.923 1
GPT-5.2	0.261 5	7.744 4	7.064 4	0.256 0	7.446 7	7.281 8
DeepSeek-V3.2-think	0.292 5	7.251 1	7.118 2	0.330 2	6.886 7	6.630 7
ERNIE-4.5-Turbo	0.311 4	7.000 0	6.827 6	0.339 8	6.864 4	6.776 0
Mengzi-T5-FT	0.213 8	3.613 3	4.452 4	0.172 2	2.324 4	3.382 7

**Fig. 8** Semantic similarity of syndrome patterns in zero-shot cross-disease tests on gynecological and TCM heart-system datasets

This figure shows the mean semantic similarity scores of syndrome patterns produced by different models in zero-shot cross-disease tests: (a) the gynecological case test set and (b) the TCM heart-system disorder case test set. Semantic scores were obtained by averaging the ratings of three independent large language models (Qwen3.5-Plus, DeepSeek-V3.2, and GPT-5.1), each of which evaluated the theoretical proximity between the model-predicted syndrome pattern and the reference syndrome differentiation conclusion on a 0–10 scale under a unified prompt. Neither dataset was used for model training or fine-tuning; both were reserved exclusively for zero-shot cross-disease transfer evaluation.

tests declined relative to the spleen-stomach disorder task, suggesting that alignment of structural elements became more difficult in cross-disease settings. The Jaccard values of Qwen3-32B (PR-CoT) were lower than those of the corresponding baseline in both the gynecological and heart-system test sets. This result is consistent with the overall observation made earlier at the structural parsing level, indicating that PR-CoT supervision did not improve the direct matching ability of structural elements under cross-disease conditions.

At the semantic similarity level, the PR-CoT model showed advantages in both cross-disease tests. In the gynecological test set, the average semantic score of Qwen3-32B (PR-CoT) was higher than that of the baseline; in the heart-system disorder test set, this advantage was even more pronounced. At the same time, the online general-purpose large language models still maintained relatively high overall scores, whereas Mengzi-T5-FT scored markedly lower in both cross-disease tests, indicating that the pretrained

generative baseline from the pre-LLM era had clear limitations in terms of cross-disease semantic proximity.

At the expert blind review level, the results of the two cross-disease tests were generally consistent with those of the semantic similarity level: Qwen3-32B (PR-CoT) scored slightly higher than the baseline in both the gynecological and heart-system disorder tests. Overall, under zero-shot cross-disease conditions, PR-CoT supervision did not improve the direct matching ability of structural elements, but still maintained certain advantages at the levels of semantic understanding and expert final adjudication. This suggests that the syndrome differentiation reasoning pattern formed through explicit pathogenesis reasoning path training has a certain degree of cross-disease transferability, and it also provides further support for the hypothesis that syndrome manifestation can be regarded as a cross-disease shared cognitive structure.

3 Discussion

The key to syndrome manifestation recognition in TCM lies in inferring the body's internal pathogenesis state from clinical observational information. Its essence is not simple label matching, but a process of integrated judgment centered on pathogenesis relationships. Using medical case records of spleen-stomach disorders as the main experimental setting, this study constructed a large-language-model-based framework for syndrome manifestation recognition under PR-CoT supervision, and conducted systematic evaluation at three levels: structural parsing, semantic similarity, and expert blind review. The results showed that, after the introduction of PR-CoT supervision, the model did not exhibit a stable advantage at the structural parsing level, but showed relatively consistent improvement at the semantic similarity and expert blind review levels. Zero-shot cross-disease transfer experiments further indicated that the pathogenesis reasoning mode induced by this training mechanism possessed a certain degree of generalizability across different disease contexts. These findings suggest that syndrome manifestation recognition is more appropriately understood as a process of generating and expressing latent pathogenesis structures, rather than as a classification task within a fixed label space. On this basis, the following discussion is organized around four aspects: modeling paradigm, cognitive alignment mechanism, cross-disease semantic sharing, and evaluation methodology.

3.1 Large language models provide a new modeling paradigm for TCM syndrome manifestation recognition

For a long time, research on TCM syndrome manifestation recognition has mostly modeled syndrome differentiation as a classification problem within a fixed label space, establishing mappings between clinical observational information and syndrome pattern labels through machine learning or deep learning methods. Although such approaches can achieve a certain level of predictive performance on specific datasets, they simplify the cognitive process of syndrome differentiation to some extent. TCM syndrome differentiation is not merely a matter of matching symptom features to labels; rather, it involves inferring the internal pathogenesis state from the four diagnostic methods and, on that basis, forming a syndrome judgment. Therefore, if syndrome manifestation recognition is treated simply as a classification task, the core role of pathogenesis

analysis and logical inference in syndrome differentiation cannot be adequately represented.

Large language models offer a new technical pathway for addressing this problem. Unlike traditional discriminative models, they operate within a unified semantic representation space and possess the ability for generative reasoning and multi-source information integration. This enables them to progressively organize pathogenesis analysis during generation, thereby establishing a closer connection between syndrome pattern expression and the reasoning process. Such a generative modeling characteristic is highly compatible with the cognitive process of TCM syndrome differentiation, which emphasizes establishing the syndrome pattern on the basis of pathogenesis.

The results of this study support this view to some extent. The structural parsing results showed that neither the large language models themselves nor their versions trained under PR-CoT supervision demonstrated a stable advantage in element-wise matching of the syndrome pattern structural components. However, at the semantic similarity and expert blind review levels, the PR-CoT-supervised models showed higher syndrome differentiation consistency and closer theoretical proximity. A question worth discussing is whether the failure of chain-of-thought supervision to produce a stable improvement in the element-wise overlap of syndrome pattern structural components would undermine the model's clinical utility. In this study, we argue that this is not necessarily the case. The element-wise overlap of syndrome pattern structural components mainly reflects structural consistency at the formal level, whereas TCM clinical practice is more concerned with whether the essential pathogenesis and the direction of syndrome differentiation are consistent. Because syndrome pattern expression is inherently flexible, the same pathogenesis may correspond to multiple semantically similar expressions. Therefore, a decline in the "hard matching" of structural elements does not necessarily imply a reduction in clinical value. In contrast, semantic similarity evaluation and expert blind review are more capable of reflecting the actual performance of model outputs in terms of theoretical consistency and clinical plausibility. This suggests that the advantage of large language models in syndrome manifestation recognition may not lie primarily in the mechanical alignment of structural terms, but rather in their capacity for overall understanding, organization, and expression of pathogenesis semantic structures. Furthermore, compared with pretrained generative

baselines from the pre-LLM era, current large language models performed better in semantic consistency and clinical plausibility, which also indirectly indicates that syndrome manifestation recognition relies more strongly on unified semantic representation and generative reasoning ability.

From a methodological perspective, the present findings support modeling syndrome manifestation recognition as a process of generating and expressing latent pathogenesis structures, rather than as a classification problem within a fixed label space. The unified semantic representation and generative reasoning abilities provided by large language models offer a new technical foundation for the integration and sharing of syndrome information across disease contexts, and also provide a feasible pathway for intelligent TCM syndrome differentiation to move from “label prediction” toward “structured generation”.

3.2 PR-CoT supervision enables cognitive alignment in syndrome differentiation

One noteworthy phenomenon observed in this study is that introducing PR-CoT supervision did not improve the element-wise overlap of syndrome pattern structural components at the structural parsing level; however, at the semantic similarity and expert blind review levels, the PR-CoT-supervised models showed relatively stable advantages. This suggests that the improvements brought by PR-CoT supervision may not be primarily reflected in the “hard matching” of structural elements, but rather in cognitive consistency at the levels of syndrome differentiation logic and pathogenesis understanding.

From the perspective of evaluation methodology, the structural parsing level mainly computes explicit element overlap in syndrome pattern expressions, focusing on whether structural elements are present and whether their combinations are consistent. However, in actual TCM contexts, syndrome patterns are often comprehensive expressions of pathogenesis relationships. Different terminological combinations may be highly similar in theoretical meaning, yet may not yield high overlap after structural decomposition. Therefore, structural metrics are more suitable for reflecting formal consistency, but are less capable of fully capturing the degree of theoretical proximity in syndrome differentiation judgment.

Against this background, the significance of introducing PR-CoT supervision lies in enabling the model to explicitly learn the reasoning path of syndrome differentiation during training, rather than relying solely on statistical associations between clinical observations and syndrome pattern labels. By

first generating pathogenesis analysis and then outputting the syndrome pattern conclusion, the model is guided to follow the logical sequence of “symptom summarization-pathogenesis inference-syndrome pattern formation”, thereby showing greater consistency in semantic similarity evaluation and expert blind review.

From the perspective of cognitive modeling, this process can be understood as a mechanism of cognitive alignment: through a structured reasoning path, the model’s internal generative pattern gradually approaches the cognitive structure of TCM syndrome differentiation. After repeatedly undergoing pathogenesis reasoning training, the model’s outputs not only present the final syndrome pattern result, but also retain, to some extent, the semantic organization of pathogenesis analysis, making syndrome pattern expression more consistent with TCM theoretical logic. To more intuitively illustrate the influence of PR-CoT supervision on syndrome differentiation expression, this study selected a representative medical case for example-based analysis (Table 7).

The gold-standard syndrome differentiation for this case was the syndrome pattern of exuberant yin with outward floating of yang. Although the syndrome pattern expressions varied across models, the outputs of most models pointed, to different degrees, to the same core pathogenesis structure, namely spleen-kidney yang deficiency, exuberant internal yin-cold, and outward floating of deficient yang. In contrast, Mengzi-T5-FT produced only the broader label cold syndrome, failing to reflect a more complete hierarchy of pathogenesis. Compared with the end-to-end baseline model, the PR-CoT-supervised model generated syndrome pattern outputs that were closer to this pathogenesis profile. This suggests that PR-CoT supervision may help the model organize and express pathogenesis relationships more accurately, thereby making its outputs semantically closer to clinical syndrome differentiation logic.

3.3 The PR-CoT mechanism supports zero-shot cross-disease transfer of syndrome semantics

The zero-shot cross-disease transfer results showed that, in the test sets of gynecological and heart-system disorders, models trained with PR-CoT supervision still maintained certain advantages at the semantic similarity and expert blind review levels, although no obvious improvement was observed at the structural parsing level. This suggests that what the model learned during training may not have been merely the correspondence between clinical observational information and syndrome pattern expression in the spleen-stomach disorder setting, but

Table 7 Representative case example: comparison of syndrome recognition results across models

Item	Content
Clinical information	Female, 61 years old; abdominal pain and diarrhea; clear, loose stool occurring more than ten times per day; aversion to cold; cold extremities; poor appetite; fatigue with a tendency to curl up in bed; pale tongue with white coating; thready and weak pulse
Gold-standard syndrome pattern	Syndrome pattern of exuberant yin with outward floating of yang
Qwen3-32B (PR-CoT)	Syndrome pattern of spleen-kidney yang deficiency with upward floating of deficient yang
Qwen3-32B (Baseline)	Syndrome pattern of spleen-stomach deficiency-cold
Qwen3-14B (PR-CoT)	Syndrome pattern of severe spleen-yang deficiency
Qwen3-14B (Baseline)	Syndrome pattern of spleen-kidney yang deficiency with exuberant internal yin-cold
GPT-5.2	Diarrhea due to spleen-kidney yang deficiency (with exuberant internal yin-cold and outward floating of deficient yang), tending toward yang collapse
DeepSeek-V3.2	Syndrome pattern of outward floating of yang (with spleen-kidney yang deficiency and upward floating of deficient yang)
ERNIE-4.5-Turbo	Syndrome pattern of spleen-kidney yang deficiency with outward floating of deficient yang
Gemini 3 Flash	Spleen-kidney yang depletion with outward floating of deficient yang
Mengzi-T5-FT	Cold syndrome

rather a pathogenesis reasoning pattern with a certain degree of general applicability.

From the perspective of TCM theory, this result has intrinsic plausibility. The syndrome system is not entirely attached to specific diseases; rather, it is a generalized expression of the body’s dysfunctional state. The same syndrome manifestation may appear in different disease contexts, which corresponds to the principle of “different diseases, same syndrome”. Therefore, clinical syndrome differentiation is usually organized around pathogenesis relationships, rather than being determined solely according to disease categories.

Once the model gradually learns the above pathogenesis structural relationships through PR-CoT supervision, its reasoning pattern may be transferred to different disease settings without exposure to training samples from the target disease. This supports, to some extent, the unified semantic space hypothesis proposed in this study: syndrome manifestation expressions in different disease contexts may be regarded as different projections of the same latent pathogenesis space. Accordingly, cross-disease syndrome manifestation recognition is not only technically feasible, but also theoretically interpretable within the framework of TCM.

3.4 The progressive multidimensional evaluation framework provides methodological guidance for TCM syndrome manifestation recognition research

An important reason why syndrome manifestation recognition differs from traditional classification tasks lies in the flexibility and combinatorial nature of syndrome pattern expression.

The same pathogenesis state may correspond to multiple terminological formulations. Therefore, if evaluation relies only on label-matching metrics such as accuracy and F1 score, it is often difficult to fully reflect the model’s actual performance in syndrome differentiation judgment.

To address this characteristic, the present study constructed a progressive multidimensional evaluation framework consisting of a structural parsing level, a semantic similarity level, and an expert blind review level. The structural parsing level is used to quantify the formal consistency of syndrome pattern expression; the semantic similarity level is used to measure the degree of proximity in theoretical connotation; and the expert blind review level provides final adjudication from the perspective of clinical syndrome differentiation, thereby making the evaluation results more consistent with actual clinical cognition.

The results show that the model characteristics revealed at different evaluation levels are not entirely the same, and that a single metric is insufficient to fully characterize the complexity of syndrome manifestation recognition. In contrast, the progressive multidimensional evaluation framework is able to present more systematically the differences in model performance across multiple aspects, including structural consistency, semantic equivalence, and clinical plausibility.

From a methodological perspective, this framework provides a reusable evaluation approach for TCM artificial intelligence research: while maintaining the efficiency of automated evaluation, it

introduces expert blind review as the final reference, thereby establishing a more robust evaluation framework between objective computation and clinical cognition.

3.5 Limitations and future research directions

Although this study explored syndrome manifestation recognition modeling under PR-CoT supervision, several limitations remain.

First, with respect to training strategy, this study started from open-source base models and used LoRA-based supervised fine-tuning to verify the role of PR-CoT supervision, but did not conduct continued pretraining in the TCM domain. In the future, domain-specific continued pretraining on larger-scale corpora of TCM literature and medical case records could be performed and combined with structured chain-of-thought supervision to further improve the model's understanding of TCM terminological systems and pathogenesis relationships.

Second, in terms of data scope, this study used medical case records of spleen-stomach disorders as the main training setting and conducted zero-shot cross-disease testing on gynecological and heart-system disorder case records. Although transferability was preliminarily verified, this design still cannot cover the multisystem diseases and complex comorbidity situations encountered in real clinical practice. Future work may expand the study to a broader disease spectrum and further evaluate the model's ability to comprehensively identify multiple pathogenesis structures in comorbidity samples.

Third, the current PR-CoT is mainly expressed in textual form, and there remains some variability in both structural granularity and mode of expression. In the future, the reasoning chain structure could be further refined by introducing more stable reasoning anchors, such as key nodes for disease location, disease nature, pathological factors, and the relationships among qi, blood, and body fluids, so as to establish a more standardized “symptoms-pathogenesis-syndrome pattern” mapping and thereby improve the stability and interpretability of the reasoning process.

Overall, the results of this study suggest that PR-CoT supervision has methodological potential under the current experimental conditions rather than that it has already become a mature solution applicable to all disease scenarios. Further work is still needed to improve the model foundation, data scale, evaluation consistency, and scope of clinical validation.

4 Conclusion

This study proposed a large-language-model-

based framework for syndrome manifestation recognition under PR-CoT supervision, and constructed a progressive multidimensional evaluation system consisting of a structural parsing level, a semantic similarity level, and an expert blind review level, through which model performance was systematically assessed. The results showed that although the introduction of PR-CoT supervision did not improve the element-wise overlap of syndrome pattern structural components, it demonstrated relatively stable advantages at the semantic similarity and expert blind review levels, and also showed a certain degree of transferability in zero-shot cross-disease tasks. These findings suggest that syndrome manifestation recognition is more appropriately understood as a process of generating and expressing latent pathogenesis structures, rather than as a classification task within a fixed label space. At the same time, the progressive multidimensional evaluation framework can more systematically characterize the differences in syndrome manifestation recognition across structural consistency, semantic equivalence, and clinical plausibility, thereby providing methodological reference for the evaluation of intelligent TCM diagnostic models in open-expression settings. Overall, this study provides empirical support for introducing TCM-theoretical constraints and structured reasoning mechanisms into the large language model framework, and also offers a research foundation for building more interpretable intelligent TCM syndrome differentiation systems.

Data availability The relevant data that directly support the research conclusion (CSTR: 31253.11. sciencedb. 35044, DOI: 10.57760/sciencedb. 35044) can be accessed and obtained in the Science Data Bank database (<https://www.scidb.cn/s/3IbuMj>). The fine-tuning data used in this study, together with the source-related raw materials from which they were derived, constitute important foundational data resources for the research group's ongoing work. In addition, because the original sources involve CNKI and other books and literature materials, public sharing of these data—even after structural processing—may still give rise to copyright risks and issues concerning the legitimate rights and interests of third parties. For these reasons, this portion of the data is not publicly shared at present; however, for legitimate research purposes, access requests may be submitted to the authors with justified reasons, and the data may be made available in a conditional manner upon the authors' approval.

References

- [1] Jiang M, Lu C, Zhang C, *et al.* Syndrome differentiation in modern research of traditional Chinese medicine. *J Ethnopharmacol*, 2012, **140**(3): 634-642
- [2] Liu Y L, Li R H, Tao L. Reflections on the establishment of an integrated speculation system in traditional Chinese medicine. *J Nanjing Univ Tradit Chin Med*, 2021, **37**(5): 658-661
- [3] Lu A, Jiang M, Zhang C, *et al.* An integrative approach of linking traditional Chinese medicine pattern classification and biomedicine diagnosis. *J Ethnopharmacol*, 2012, **141**(2): 549-556
- [4] Lai X M, Yang X M, Chen M M, *et al.* Practice and reflection on standardization of name and connotation of TCM syndrome based on the principle of differentiation of syndrome elements. *China Journal of Traditional Chinese Medicine and Pharmacy*, 2017, **32**(6): 2470-2472
- [5] Xu Y L, Wu X Y, Li Y L, *et al.* Three methods of latent structure to determine the TCM syndrome based for the unlabeled data. *World Science and Technology-Modernization of Traditional Chinese Medicine*, 2019, **21**(1): 101-108
- [6] Zhang G G, Lee W L, Lao L X, *et al.* The variability of TCM pattern diagnosis and herbal prescription on rheumatoid arthritis patients. *Altern Ther Health Med*, 2004, **10**(1): 58-63
- [7] Liu G P, Wang Y Q, Zhao N Q, *et al.* Study on the diagnosis agreement of clinical doctor of traditional Chinese medicine. *World Science and Technology-Modernization of Traditional Chinese Medicine*, 2010, **12**(3): 356-360
- [8] Liu Z, He H, Yan S, *et al.* End-to-end models to imitate traditional Chinese medicine syndrome differentiation in lung cancer diagnosis: model development and validation. *JMIR Med Inform*, 2020, **8**(6): e17821
- [9] Li X, Chen K, Yang J, *et al.* TLDA: a transfer learning based dual-augmentation strategy for traditional Chinese Medicine syndrome differentiation in rare disease. *Comput Biol Med*, 2024, **169**: 107808
- [10] Hu C, Zhang S, Gu T, *et al.* Multi-task joint learning model for Chinese word segmentation and syndrome differentiation in traditional Chinese medicine. *Int J Environ Res Public Health*, 2022, **19**(9): 5601
- [11] Lim J, Li J, Zhou M, *et al.* Machine learning research trends in traditional Chinese medicine: a bibliometric review. *Int J Gen Med*, 2024, **17**: 5397-5414
- [12] Tang J, He S. DC-TSCM: an interpretable dual-channel traditional Chinese medicine syndrome classification model *via* semantic-structural fusion. *PeerJ Comput Sci*, 2026, **12**: e3555
- [13] Yao L, Jin Z, Mao C, *et al.* Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. *J Am Med Inform Assoc*, 2019, **26**(12): 1632-1636
- [14] Yu X, He L, Wang Q, *et al.* Heat syndrome types prediction of traditional Chinese medicine in acute ischemic stroke through deep learning: a pilot study. *Front Pharmacol*, 2025, **16**: 1601601
- [15] Zhang G, Jiang H, Kuai L, *et al.* KGSD-Net: a knowledge graph syndrome differentiation network for syndrome classification. *Front Med*, 2025, **12**: 1555781
- [16] Ye H, Feng Q S, Yan X Y, *et al.* Study of tongues and pulses syndrome prediction of diabetes mellitus combined with coronary heart disease based on artificial neural network. *China Journal of Traditional Chinese Medicine and Pharmacy*, 2020, **35**(10): 5184-5187
- [17] Yu N, Huang J, Liu J, *et al.* Development of a cold-heat syndrome classification model for children with allergic rhinitis based on multimodal data. *Transl Pediatr*, 2025, **14**(12): 3375-3386
- [18] Ba H, Du H, Cheng C, *et al.* Identification of syndrome types in patients with pancreatic cancer from free text in electronic medical records: model development and validation. *JMIR Form Res*, 2025, **9**: e70602
- [19] Zhu L, He S, Zheng W, *et al.* Predictive model of ulcerative colitis syndrome with ensemble learning and interpretability methods. *Sci Rep*, 2025, **15**(1): 21985
- [20] Zhang T T, Wang Y Q, Wen C B, *et al.* Challenges and ideas in the interpretability of intelligent syndrome differentiation of traditional Chinese medicine. *Journal of Traditional Chinese Medicine*, 2022, **63**(6): 506-510
- [21] Lu Y, Fang Z, Zeng T, *et al.* Chronic hepatitis B: dynamic change in traditional Chinese medicine syndrome by dynamic network biomarkers. *Chin Med*, 2019, **14**: 52
- [22] Song Y, Ma S, Dai Y, *et al.* AI-assisted TCM syndrome differentiation: key issues and technical challenges. *Strateg Study CAE*, 2024, **26**(2): 234
- [23] Deng J, Dai S, Liu S, *et al.* Clinical study of intelligent tongue diagnosis and oral microbiome for classifying TCM syndromes in MASLD. *Chin Med*, 2025, **20**(1): 78
- [24] Chen H, He Y. Machine learning approaches in traditional Chinese medicine: a systematic review. *Am J Chin Med*, 2022, **50**(1): 91-131
- [25] Xu F, Hao Q, Shao C, *et al.* Toward large reasoning models: a survey of reinforced reasoning with large language models. *Patterns*, 2025, **6**(10): 101370
- [26] Li X, Liu X, Shu Q, *et al.* Automatic contrastive chain-of-thought prompting: learning from reasoning errors of large language models. *Expert Syst Appl*, 2026, **306**: 130919
- [27] Huang Z, Miao J, Chen J, *et al.* A traditional Chinese medicine syndrome classification model based on cross-feature generation by convolution neural network: model development and validation. *JMIR Med Inform*, 2022, **10**(4): e29290
- [28] Zhao Y, He L, Xie Q, *et al.* A novel classification method for syndrome differentiation of patients with AIDS. *Evid Based Complement Alternat Med*, 2015, **2015**: 936290
- [29] Chen Z, Zhang D, Liu C, *et al.* Traditional Chinese medicine diagnostic prediction model for holistic syndrome differentiation based on deep learning. *Integr Med Res*, 2024, **13**(1): 101019
- [30] Xiao L, Wang C W, Deng Y, *et al.* HHO optimized support vector machine classifier for traditional Chinese medicine syndrome differentiation of diabetic retinopathy. *Int J Ophthalmol*, 2024,

- 17(6): 991-1000
- [31] Li P F, Yang X K, Bai Y C, *et al.* Study on named entity extraction in TCM medical records based on large language models. Chinese Journal of Library and Information Science for Traditional Chinese Medicine, 2024, **48**(2): 108-113
- [32] Baidu. ERNIE 4.5 Turbo. Beijing: Baidu Intelligent Cloud, 2025 [2026-02-05]. <https://console.bce.baidu.com/qianfan/modelcenter/model/buildIn/detail/am-xpgukxjf6s0r>
- [33] Zhipu AI. GLM-4.5. Beijing: Zhipu AI, 2025 [2026-02-05]. <https://docs.bigmodel.cn/cn/guide/models/text/glm-4.5>
- [34] Yang A, Li A F, Yang B S, *et al.* Qwen3 technical report. arXiv: 2505.09388, 2025 [2026-03-15]. <https://arxiv.org/abs/2505.09388>
- [35] Singh A, Fry A, Perelman A, *et al.* OpenAI GPT-5 System Card. arXiv: 2601.03267, 2025 [2026-03-15]. <https://arxiv.org/abs/2601.03267>
- [36] DeepSeek-AI, Liu A X, Mei A X, *et al.* DeepSeek-V3.2: pushing the frontier of open large language models. arXiv: 2512.02556, 2025 [2026-03-15]. <https://arxiv.org/abs/2512.02556>
- [37] Google DeepMind. Gemini 3 flash model card. Mountain View: Google DeepMind, 2025 [2026-02-20]. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf>
- [38] Zhang Z S, Zhang H Q, Chen K M, *et al.* Mengzi: towards lightweight yet ingenious pre-trained models for Chinese. arXiv: 2110.06696, 2021 [2026-03-15]. <https://arxiv.org/abs/2110.06696>
- [39] Li C D, Fang C Y. Zhongyi Zhenduanxue. 5th ed. Beijing: China Press of Traditional Chinese Medicine, 2021: 138-190
- [40] Zhu W F. Zhengsu Bianzhengxue. Beijing: People's Medical Publishing House, 2008: 36
- [41] Qwen Team. Qwen3.5: accelerating productivity with native multimodal agents. Hangzhou: Alibaba Cloud, 2026 [2026-02-20]. <https://qwen.ai/blog?id=qwen3.5>

基于病机推理思维链监督的大模型脾胃病证候识别及多维评价研究*

杨淑涵^{1)**} 胡宇鑫^{2)**} 俞欣妤¹⁾ 涂钰莹³⁾ 臧一畅¹⁾ 李盼飞^{1)***}

(¹⁾ 浙江中医药大学医学技术与信息工程学院, 杭州 310053;

(²⁾ 中国计量大学信息工程学院, 杭州 310018; (³⁾ 中国中医科学院西苑医院肝病科, 北京 100091)

摘要 目的 证候识别的本质是基于临床观测信息推断机体内在病机状态, 而非简单标签匹配。本研究旨在探讨在大语言模型中引入病机推理思维链监督能否提升证候识别质量及其跨病种迁移能力。**方法** 以脾胃病医案构建证候识别数据集, 并标注结构化病机推理思维链; 在开源大语言模型上开展监督微调, 并以端到端模型为对照。构建由结构解析层、语义相似度层和专家盲评层组成的递进式多维评价体系: 结构解析层基于证型结构要素匹配进行量化评价, 语义相似度层由独立大模型对证型理论接近程度进行评分, 专家盲评层从辨证一致性与证型术语规范性两个维度进行终审。另以妇科和心系疾病医案开展零样本跨病种迁移测试。**结果** 结构解析层显示, 思维链监督未稳定提升证型结构要素逐项重合度。但在语义相似度层和专家盲评层, 思维链训练模型表现出较一致优势, 提示其证型表达在理论内涵和临床辨证层面更接近参考结论。跨病种测试中, 上述语义与专家评价优势仍可观察到, 显示出一定迁移性。**结论** 病机推理思维链监督的收益主要体现在中医语义一致性与临床合理性层面, 而非结构要素硬匹配的提升。研究结果支持将证候识别理解为潜在病机结构的生成与表达过程, 并为中医智能辨证提供了一条兼顾理论对齐、可解释性与多层评价的方法学路径。

关键词 中医证候识别, 大模型, 思维链, 证型评价, 脾胃病

中图分类号 R229, TP182

DOI: 10.3724/j.pibb.2026.0141

CSTR: 32369.14.pibb.20260141

* 国家自然科学基金 (82505804), 国家中医药综合改革示范区科技共建重大项目 (GZY-KJS-ZJ-2025-018), 浙江省自然科学基金 (QN26H270007) 和浙江中医药大学校级科研项目 (2024RCZXZK42) 资助。

** 并列第一作者。

*** 通讯联系人。

Tel: 0571-86613573, E-mail: lipanfei@zcmu.edu.cn

收稿日期: 2026-03-06, 接受日期: 2026-05-05