Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI-powered TCM Diagnosis: A Multi-Task Learning Approach for Syndrome Element Diagnosis and Syndrome Differentiation

Yan Lyu
Peking University
Xiangnan Feng
Fudan University
Ping Jiang
Beijing University of Chinese Medicine
Haoxuan Li
Peking University
Xinxing Lai
new-star@163.com

Beijing University of Chinese Medicine
Xiao-Hua Zhou
Peking University

Research Article

Keywords: Syndrome Differentiation, Syndrome Element, Multi-Task Learning, Artificial Intelligence, TCM Diagnosis

Posted Date: March 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9200456/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Xinxing Lai serves as a member of the Youth Editorial Board of Chinese Medicine. The author was not involved in the review or decision-making process for this manuscript. The authors declare no other conflicts of interest.

AI-powered TCM Diagnosis: A Multi-Task Learning Approach for Syndrome Element Diagnosis and Syndrome Differentiation

Yan Lyu[1†], Xiangnan Feng[4†], Ping Jiang[6,7], Haoxuan Li[1], Xinxing Lai[5,6*], Xiao-Hua Zhou[2,3*]

[1]Center for Data Science, Peking University, Beijing, China.
[2]Department of Biostatistics and Beijing International Center for Mathematical Research, Peking University, Beijing, China.
[3]Chongqing Research Institute of Big Data, Peking University, Chongqing, China.
[4]Department of Statistics and Data Science, Fudan University, Shanghai, China.
[5]Institute for Brain Disorders, Beijing University of Chinese Medicine, Beijing, China.
[6]Department of Neurology, Dongzhimen Hospital, Beijing University of Chinese Medicine, Beijing, China.
[7]Beijing International Center for Mathematical Research, Peking University, Beijing, China.

*Corresponding author(s). E-mail(s): new-star@163.com; azhou@math.pku.edu.cn;
†These authors contributed equally to this work.

Abstract

Background. Syndrome differentiation is the cornerstone of Traditional Chinese Medicine (TCM). The development of robust artificial intelligence aided diagnosis (AIAD) syndrome differentiation method is therefore a pivotal direction for TCM modernization. Contemporary TCM theory identifies syndrome

elements as fundamental components essential for accurate syndrome differentiation. However, existing computational methods have not fully exploited their potential through joint learning approaches. This study proposes a novel AIAD framework that simultaneously learns syndrome element diagnosis and syndrome differentiation. The framework aims to enhance diagnostic accuracy and provide empirical validation of the theoretical role of syndrome elements within the TCM diagnostic process.

Methods. We developed a novel AIAD framework trained and tested on 6,226 electronic health records (EHRs) and further evaluated on an independent dataset of 1,057 EHRs derived from clinical guidelines. A multi-task learning approach was utilized to simultaneously model syndrome element diagnosis and syndrome differentiation. We proposed and compared two models: AIAD-A1, which considers syndrome elements as intermediate variables, and AIAD-A2, which treats them as mediators. Model performance was assessed using top-K accuracy and NDCG@K metrics.

Results. The AIAD-A2 model, which treats syndrome elements as mediators, significantly outperformed both the AIAD-A1 model and baseline models across multiple evaluation metrics. AIAD-A2 improved top-K accuracy by up to 6.31% and demonstrated remarkable enhancements in handling long-tail data, improving top-15 accuracy by 21.34% within the tail subgroup. Additionally, it exhibited superior generalizability, with the smallest performance decline (7.77% in top-15 accuracy) on an independent test dataset facing a different data distribution, compared to drops of 22.7% for the baseline and 12.44% for AIAD-A1. Moreover, model visualization confirmed AIAD-A2's capability to focus on the most clinically relevant symptoms.

Conclusions. Our findings reveal the role of syndrome elements as mediators in TCM syndrome differentiation. Accurate diagnosis necessitates the integrated consideration of both syndrome elements and the original symptom descriptions. The proposed AIAD-A2 framework offers an effective and generalizable approach for AI-powered TCM diagnosis, successfully addressing challenges such as data imbalance and distribution shifts. This work contributes to the modernization of TCM by delivering a robust AI method and deepening the theoretical understanding of the diagnostic process.

Keywords: Syndrome Differentiation, Syndrome Element, Multi-Task Learning, Artificial Intelligence, TCM Diagnosis

1 Introduction

Traditional Chinese Medicine (TCM), as well as its variants, is a widely adopted treatment regime in Asian countries [1] and has shown its efficacy in a wide range of diseases [2–5]. However, the understanding, education, and application of TCM are still limited [6], due to the following two reasons. First, TCM bases its philosophy on the holistic

and dialectical view and the theory of Yin-Yang and five elements [7, 8], which is summarized from thousands of years of clinical experience with the goal of maintaining the balance of the human body as a whole [9]. This conventional and abstract TCM theory is hardly understandable by modern science. Second, the diagnostic and treatment practice of TCM is largely subjective and experience-based [10, 11], which makes it difficult to theorize, generalize, and inherit the practice. Therefore, it is essential to systematically understand the logic of pathogenesis, diagnosis, and the inherent connections within TCM, which represents the modernization of TCM theory.

There are mainly two routes for modernizing TCM theory. The first route focuses on uncovering the biological foundations of TCM diagnosis and treatment, and has achieved significant success in identifying TCM's pharmacological, cellular, molecular, and genetic bases [12–14]. The other route lies in the modernization of the diagnosis and therapeutic methods of TCM. Existing studies have developed computer aided diagnosis (CAD) methods that provides automated, standardized, and objective diagnosis and therapeutic recommendations [15, 16]. These CAD methods utilize machine learning methods to quantitatively analyze the relationships between symptoms and TCM diagnosis and treatment. The learned relationships not only objectify the TCM diagnosis and treatment process, but also reveal the underlying patterns and mechanisms. Consequently, TCM theory can be refined from an interdisciplinary perspective combining conventional TCM theory and modern data science. An emerging and promising direction in this route is to develop more advanced artificial intelligence (AI) aided diagnosis (AIAD) methods. AI methods are appealing for analyzing massive unstructured data, such as text, and for handling complex model structures, such as the joint learning of multiple tasks. These complexities are highly relevant in modeling the TCM diagnosis process. In particular, transforming the subjective and empirical practice of syndrome differentiation into an objective and intelligent process is a critical step in this modernization route.

Syndrome differentiation is one of the fundamental concepts in the practice of TCM that consists of a series of diagnostic procedures [10]. The TCM syndrome can be viewed as the summary of the symptoms and signs of a patient at a specific stage during the course of an illness [17]. Syndrome differentiation is thus to determine the patient's syndrome based on the symptoms gathered from the four diagnostic methods: inspection, auscultation, inquiry, and palpation [18]. Syndrome differentiation can be used for further stratification of the patients' conditions with specific disease, which thus could guide the following personalized treatments. Therefore, an accurate syndrome differentiation is essential for TCM therapy [19]. An illustration of the syndrome differentiation process is depicted in Figure 1a.

Recently, several studies have attempted to develop AIAD methods for syndrome differentiation [20–22]. However, these methods have important limitations. First, the existing AIAD methods heavily rely on the structured and manually selected numeric indicators derived from symptoms as input. However, the original data containing full information of symptoms are text collected from the four diagnostic methods. Second, these methods only focus on a few syndromes (≈ 10) related to a specific western medical disease. TCM is a general medicine that involves multitudinous syndromes (in the hundreds) and thus highly complex syndrome differentiation, which has not been systematically examined before. Third, previous studies simply link symptoms to syndromes and ignored the detailed analyzing procedure in the diagnosis process, which miss the valuable opportunity to closely mimic the TCM diagnostic practice and reveal the underlying patterns and mechanisms.

To address these limitations and effectively model the complex mapping between massive symptoms and multitudinous syndromes, it is essential to revisit the underlying diagnostic logic of TCM. This brings us to the theory of syndrome elements. Originating from the work of Zhu [23] , this theory aims at demystifying syndrome differentiation by decomposing a complex syndrome into basic elements that describe

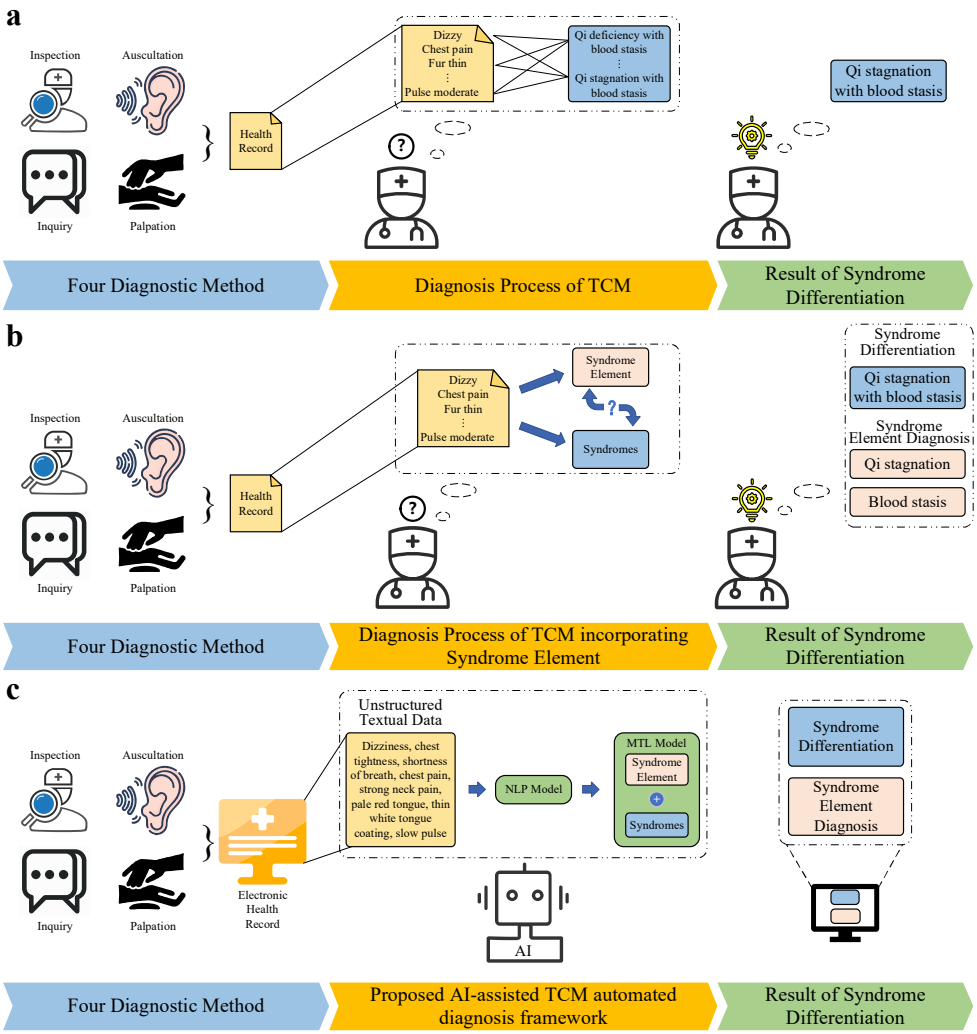


Fig. 1: **Graphical illustration of the TCM diagnosis process. a)** Syndrome differentiation diagnostic process matches symptoms and syndromes directly. **b)** Syndrome differentiation diagnostic process incorporating syndrome elements. **c)** Our proposed TCM AI-aided diagnosis framework.

the location, essence, or pathogenesis of the syndrome. Since the resulting syndrome elements are much fewer in number than the syndromes , they provide a more micro and fundamental viewpoint that significantly facilitates the diagnostic process. In clinical practice, physicians typically infer syndrome elements from symptoms first and

then summarize them into an appropriate syndrome (Figure 1b). Inspired by this theoretical framework, we develop a novel AIAD framework that explicitly incorporates syndrome elements into the syndrome differentiation process. A recent perspective article also highlights the potential of introducing syndrome elements into AIAD TCM diagnosis methods [24]. They conceptually proposed (without implementing and testing) a TCM diagnostic framework, which involves first identifying syndrome elements from standardized symptoms and then performing syndrome differentiation based on these syndrome elements. This conceptual framework can be viewed as a motivation for our current research.

To date, no previous study has considered the joint learning of syndrome element diagnosis and syndrome differentiation. This may be due to the ambiguity and challenge in how syndrome elements work in syndrome differentiation and the difficulty of jointly handling these two tasks in one model. Although the conventional theory of syndrome elements assumes that a syndrome can be directly determined by the diagnosed syndrome elements, the clinical practice of syndrome differentiation is not as simple. In TCM clinical practice, a physician may first diagnose the corresponding syndrome elements from the symptoms and then conclude a syndrome solely based on the diagnosed elements (i.e., the syndrome elements are regarded as the intermediate variable between symptoms and the syndrome). Another possibility is that after diagnosing the corresponding syndrome elements, a physician may retain the related information in symptoms that cannot be summarized by syndrome elements. This information, along with the syndrome elements, are considered together to conclude an appropriate syndrome (i.e., the syndrome elements are regarded as the mediator variable between symptoms and the syndrome, which determines the syndrome together with the symptoms). These two possibilities of diagnosis procedure demonstrate the complexity of the decision-making process behind TCM syndrome element diagnosis and syndrome differentiation. In this study, to model these two possible

analyzing procedures, we propose two alternatives concerning the role of syndrome elements in syndrome differentiation, namely the intermediate variable alternative and the mediator alternative.

To implement these two alternatives, we need to jointly analyze two correlated tasks, syndrome element diagnosis and syndrome differentiation. Therefore, we utilize the cutting-edge MTL framework in deep learning. The MTL framework jointly analyzes several correlated tasks simultaneously with the aim of cross-leveraging useful information contained in these tasks to help build an improved learner for each task [25, 26]. The joint modeling structure of MTL can be flexibly set by users to align with the real situations. In our scenario, we implement the two alternatives by jointly modeling syndrome element diagnosis and syndrome differentiation under two MTL structures. The utilization of MTL framework potentially promotes the diagnostic accuracy of both tasks and helps explore the mechanisms of the theory of syndrome elements. Another significant advantage associated with the multi-task learning of syndrome element diagnosis and syndrome differentiation is that the true labels of syndrome elements (usually unknown in real applications) are only needed in the training stage and are not required in the application of the trained model. Our trained AIAD method takes original text of symptoms as input and determine the syndrome elements and the syndrome automatically. No extra information, including the manually extracted numeric indicators from symptoms and the true labels of syndrome elements, is required, which provides a broad prospect for our proposed AIAD method. The illustration of our AIAD method is depicted in Figure 1c.

This paper makes both methodological and theoretical contributions to the modernization of TCM diagnosis theory. Methodologically, we propose a novel AIAD TCM diagnosis method. To our knowledge, this is the first study that jointly learns syndrome element diagnosis and syndrome differentiation. Extensive experiments confirm

the superiority of our AIAD method: it outperforms baselines in both complex multi-syndrome differentiation and targeted subset differentiation, mitigates popularity bias from imbalanced samples, and exhibits strong generalizability under dataset distribution shifts. Additionally, our visualization method enhances model interpretability by revealing the underlying TCM diagnostic logic. Theoretically, we make three key contributions: (1) Demonstrating AIAD's reliability in handling complex multitudinous syndrome differentiation; (2) Proposing and providing the first empirical verification using field data that syndrome elements act as mediators (not mere intermediate variables) in syndrome differentiation, which supplements the theories of syndrome elements and syndrome differentiation; (3) Highlighting valuable information in raw symptom descriptions that aids the selection and combination of syndrome elements, providing a theoretical interpretation for the mediating role of syndrome elements in syndrome differentiation.

2 Methods

2.1 Data collection and preparation

The objective of this study is to develop an AIAD method for syndrome differentiation that incorporates syndrome elements. We trained our proposed method on 6,226 electronic health records (EHRs) from the clinical medical record database supported by the *National Science and Technology Support Program of China* (NSTSPC). Additionally, we used another 1,057 EHRs extracted from TCM clinical guidelines issued by the China Association of Chinese Medicine (CACM) as an independent test dataset (ITD). The collection and preparation of these two datasets are described as follows.

2.1.1 NSTSPC dataset

Data Source. We utilized the dataset constructed under NSTSPC (Project No. 2007BA I10B01), a national-level program collecting EHRs from 210 senior TCM

physicians across 75 hospitals. Widely acknowledged as a gold standard in the literature [27, 28], this dataset aims to systematize expert clinical experience of senior TCM physicians by computationally analyzing EHRs to decode the underlying dialectical reasoning and therapeutic rationales. This facilitates the standardization of TCM diagnosis, enhancing its reproducibility, learnability, and generalizability. After rigorous quality control (see Appendix A.1 in the supplementary material for detailed standardization and inclusion/exclusion criteria), the final cohort for this study included 6,226 EHRs (2,589 males, 3,637 females; mean age: 44±19 years) covering 161 distinct syndromes (see Appendix B for the full list of syndrome names included in this study).

Dual-Level Supervision. A critical distinction of the NSTSPC dataset, compared to conventional medical datasets that only map symptoms to final diagnoses, is the inclusion of textual *Syndrome Differentiation Analysis* reports (see Appendix Figure A1b in the supplementary material). These reports record the physicians' intermediate reasoning process, from which we extracted *Syndrome Elements*. Consequently, the dataset provides dual-level supervision: (1) Syndrome Elements (intermediate reasoning labels) and (2) Syndrome Names (final diagnostic labels). This rich annotation structure enables our model to explicitly learn the mapping from symptoms to syndrome elements and finally to syndromes, rather than treating diagnosis as a black-box classification task. A detailed description of the syndrome element extraction procedure is provided in Appendix A.1, and the complete list of syndrome elements included in this study is tabulated in Appendix B.

Data Imbalance and Stratification. The dataset exhibits a significant long-tail distribution [29], where a few common syndromes dominate while rare conditions (tail classes) are prevalent. To systematically evaluate model performance across different frequencies, we stratified the syndromes into three subgroups, namely the *Head*, *Middle*, and *Tail* subgroup, based on their sample prevalence (see Figure 2a for the sample distribution). Specifically, the dataset was partitioned such that each subgroup

contains approximately one-third of the total EHRs, corresponding to 13, 31, and 117 distinct syndromes, respectively. This stratification serves as a crucial benchmark and will be discussed in more detail in Section 3.2.

Clinical Coverage. To validate the clinical representativeness of the dataset, we mapped the diagnostic records within the EHRs to the ICD-11 classification system[1] using a semantic mapping strategy (detailed in Appendix A.1 of the supplementary material). The dataset encompasses a wide spectrum of physiological systems, including the digestive, circulatory, and nervous systems. Subsequently, we constructed a Sankey diagram to visualize the associations between ICD-11 categories, TCM disease entities, and syndrome differentiation results (Appendix Figure A2 in the supplementary material). The plot reveals complex interconnections across these three levels, validating the TCM principle of "one syndrome, multiple diseases". These characteristics ensure the extensive diagnostic coverage of the NSTSPC dataset in terms of both physiological and TCM pathologies.

2.1.2 ITD dataset

To validate the performance of our proposed AIAD method, we collected ITD, an external validation dataset. Acknowledging that it is crucial to assess whether study results are generalizable to new patients [30], we specifically utilized this ITD to verify our model's robustness on unseen data. This property is vital for safe clinical adoption, as models failing to generalize may introduce pernicious bias or pose safety risks when deployed in heterogeneous real-world settings [31, 32]. Unlike the NSTSPC dataset derived from raw clinical records, the ITD consists of 1,057 canonical cases extracted from 11 clinical guidelines issued by CACM. These guidelines cover diverse medical specialties (e.g., gynecology, pediatrics, see Appendix A.2 in the supplementary material for details) and provide expert-reviewed, standardized symptom descriptions for the 161 target syndromes.

[1] https://icd.who.int/browse/2025-01/mms/en

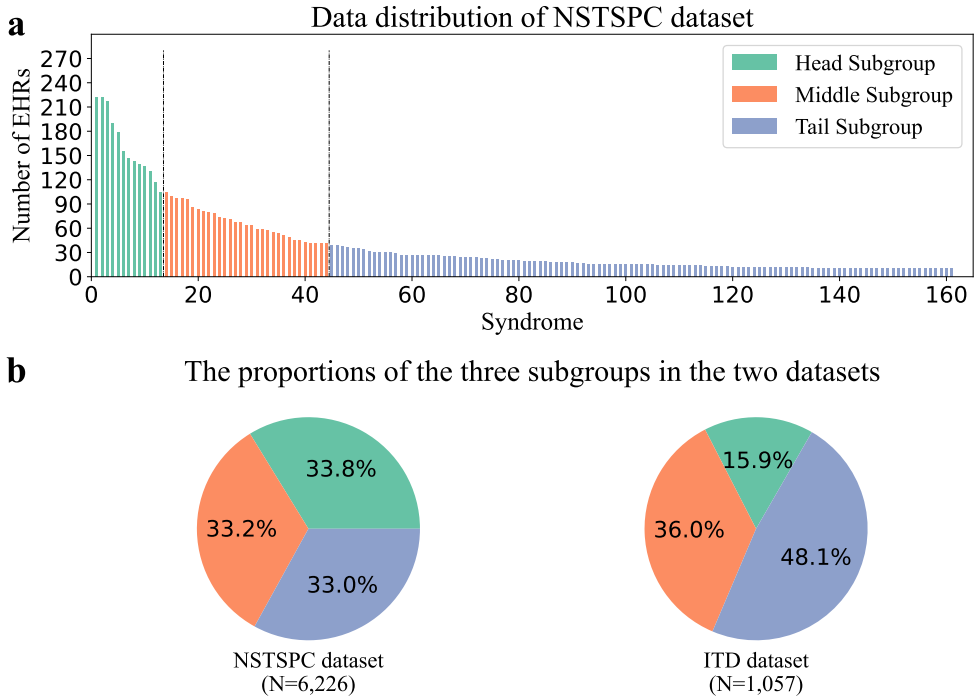


Fig. 2: **Data statistics. a)** Data distribution of NSTSPC dataset, with the head, middle and tail subgroups displayed in different colors. The head (middle, and tail) subgroup containing the first (second, and third) 1/3 of all EHRs corresponds to 13 (31, and 117) different syndromes. **b)** The proportions of the three subgroups in the NSTSPC and the ITD datasets.

Crucially, the ITD imposes a significant challenge due to a severe distribution shift relative to the training data. We also partitioned the ITD into the head, middle, and tail subgroups, according to the same syndrome classifications for NSTSPC. While the NSTSPC dataset was stratified into equal subgroups (33% each for Head/Middle/Tail subgroups), the ITD exhibits a skew towards the Tail subgroup where the Head, Middle, and Tail subgroups account for 15.9%, 36.0%, and 48.1% of the test data, respectively (see Figure 2b). Given this significant distribution shift between the two datasets, the ITD functions as a rigorous external benchmark for validating the model's ability to handle long-tail distributions and generalize to unseen heterogeneous settings.

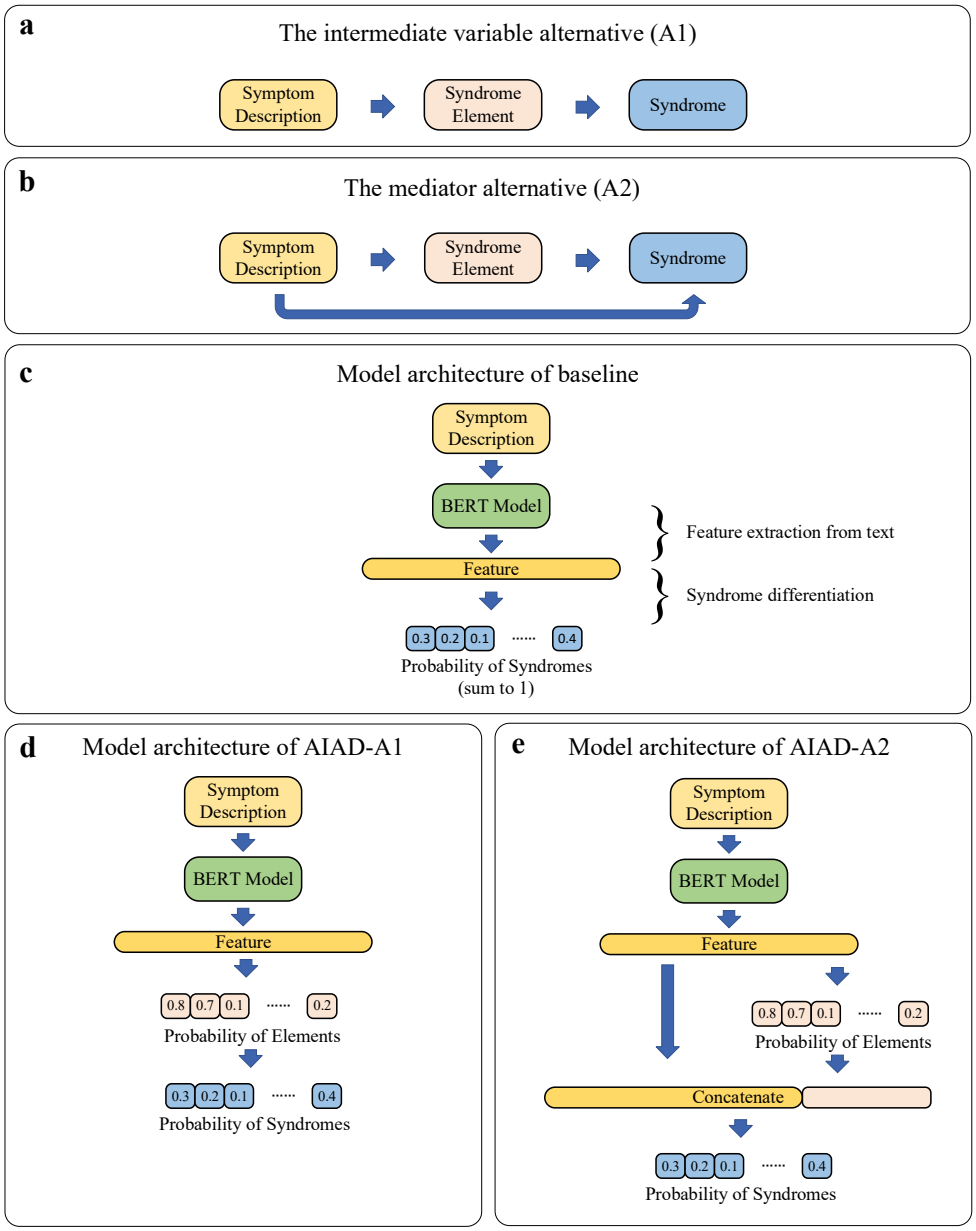


Fig. 3: **Graphical illustration of the model architectures. a)** The intermediate variable alternative (A1). **b)** The mediator alternative (A2). The model architecture for the **c)** baseline, **d)** AIAD-A1, and **e)** AIAD-A2.

2.2 Baseline syndrome differentiation model

Our baseline model implements the direct diagnostic procedure depicted in Figure 1a, serving as a performance benchmark for the comparison with our proposed methods in later sections. In this model, syndrome differentiation is modeled as a multi-class classification task where the input is the unstructured symptom description and the output is the probability distribution over all possible syndromes.

The model architecture comprises a sequential pipeline of feature extraction and classification. First, we utilize the Bidirectional Encoder Representations from Transformers (BERT) [33] to extract deep semantic features from the unstructured text. For the i-th EHR, the BERT model transforms the raw symptom description into a high-dimensional real-valued vector $\mathbf{x}_i^{PO} \in \mathbb{R}^{768}$ as:

$$\mathbf{x}_i^{PO} = BERT(\text{symptom description}_i). \tag{1}$$

Subsequently, a classification head, consisting of a linear layer and a Softmax activation function, maps this feature vector $\mathbf{x}_i^{PO}$ to the probability distribution over the N_s syndromes (where $N_s = 161$). The estimated probability $\hat{y}_{ij}$ that the i-th EHR belongs to the j-th syndrome is calculated as:

$$\hat{y}_{ij} = Softmax(LN(\mathbf{x}_i^{PO}))_j, \tag{2}$$

where $LN(\cdot)$ represents the linear layer. The model is optimized by minimizing the cross-entropy loss function $\mathcal{L}_s$ against the ground truth labels y_{ij} as:

$$\mathcal{L}_s = -\sum_{i=1}^{N}\sum_{j=1}^{N_s} y_{ij} \log \hat{y}_{ij}. \tag{3}$$

The structure of the baseline model is depicted in Figure 3c and technical details regarding the BERT architecture is provided in Appendix C in the supplementary material.

2.3 Incorporation of syndrome element diagnosis

To mimic the reasoning process of expert TCM physicians, we propose the AIAD framework, which explicitly incorporates syndrome element diagnosis into the syndrome differentiation process. We treat syndrome element diagnosis as a multi-label classification task, as a patient may manifest multiple syndrome elements simultaneously.

Regardless of the specific role syndrome elements play, whether as intermediate variables or mediators, we first estimated the probabilities of syndrome elements using the same feature vector $\mathbf{x}_i^{PO}$ extracted by BERT. The probability $\hat{y}_{ik}^e$ that the i-th EHR contains the k-th syndrome element is computed as:

$$\hat{y}_{ik}^e = \sigma(\mathbf{x}_i^e)_k, \tag{4}$$

where $\mathbf{x}_i^e = LN(\mathbf{x}_i^{PO})$ is the logit of the syndrome elements, $\sigma(\cdot)$ is the sigmoid function and $LN(\cdot)$ is the linear layer that projects the features to the syndrome element space ($N_e = 51$). The corresponding loss $\mathcal{L}_e$ is defined as:

$$\mathcal{L}_e = -\sum_{i=1}^{N}\sum_{k=1}^{N_e}\left[y_{ik}^e \log \hat{y}_{ik}^e + (1 - y_{ik}^e)\log(1 - \hat{y}_{ik}^e)\right]. \tag{5}$$

However, a critical question remains regarding how these diagnosed syndrome elements contribute to the final syndrome differentiation. To address this, we proposed and implemented two alternative hypotheses regarding the role of syndrome elements.

2.3.1 The intermediate variable alternative (AIAD-A1)

The intermediate variable alternative (A1) aligns with the conventional hierarchical view of the theory of syndrome elements. It posits that syndrome elements function as complete intermediaries that summarize all necessary information from symptoms. Under this assumption, the diagnostic pathway is *Symptom* $\rightarrow$ *Syndrome Element* $\rightarrow$ *Syndrome* (Figure 3a). This implies that once the syndrome elements are identified, the original symptom descriptions provide no additional information for the final syndrome differentiation.

Based on this concept, we designed the AIAD-A1 model structure as shown in Figure 3d. In this architecture, the final syndrome differentiation depends solely on the predicted syndrome elements. Specifically, the estimated probability distribution over syndromes using AIAD-A1 is given by:

$$\hat{y}_{ij} = Softmax(LN(\mathbf{x}_i^e))_j, \tag{6}$$

with $\mathbf{x}_i^e$ defined in Equation (4). The model was trained using an MTL objective that minimizes the joint sum of both losses:

$$\mathcal{L}_{MTL} = \mathcal{L}_s + \mathcal{L}_e, \tag{7}$$

where $\mathcal{L}_s$ and $\mathcal{L}_e$ are defined in Equation (3) and Equation (5), respectively.

2.3.2 The mediator alternative (AIAD-A2)

The mediator alternative (A2) proposes a more complex and realistic decision-making process. While syndrome elements capture the pathological essence, we argue that they are structured summaries of the patient's condition. The original symptom descriptions contain fine-grained details that may not be fully encapsulated by the syndrome

elements and are crucial for differentiating syndromes. Therefore, we posit that syndrome elements act as mediators, which means the final diagnosis is jointly determined by both the syndrome elements and the original symptoms (Figure 3b).

To model this dual-pathway logic (*Symptom* $\rightarrow$ *Syndrome* and *Symptom* $\rightarrow$ *Syndrome Element* $\rightarrow$ *Syndrome*), we designed the AIAD-A2 model as shown in Figure 3e. Unlike AIAD-A1, AIAD-A2 utilizes a feature concatenation strategy. We concatenated the predicted syndrome element logit vector $\mathbf{x}_i^e$ with the original BERT extracted feature vector $\mathbf{x}_i^{PO}$. This combined vector was then used to predict the probability distribution over syndromes as:

$$\hat{y}_{ij} = Softmax(LN([\mathbf{x}_i^e, \mathbf{x}_i^{PO}]))_j. \tag{8}$$

This architecture allows the model to leverage the high-level semantic guidance provided by syndrome elements while retaining access to the rich, raw information embedded in the symptom text. The model was optimized using the same MTL loss function defined in Equation (7).

2.4 Evaluation metrics and experimental setup

We evaluated the performance of the Baseline, AIAD-A1, and AIAD-A2 methods, using two commonly adopted metrics, the Top-K accuracy [34, 35] and the NDCG@K (normalized discounted cumulative gain at K) [36, 37]. Higher values for both metrics indicate superior model performance. We conducted a 5-fold cross-validation to assess our proposed models. Additionally, in each fold, the trained models were evaluated on the ITD. Detailed descriptions of the evaluation metrics and experimental setup are provided in Appendix D in the supplementary material.

3 Results

In this section, we validate the effectiveness of our proposed method through five experiments. In Section 3.1, we compare the performance of the baseline, AIAD-A1, and AIAD-A2 on the complete NSTSPC dataset. In Section 3.2, we examine the effectiveness of the baseline, AIAD-A1, and AIAD-A2 in the head, middle, and tail subgroups of the NSTSPC dataset, respectively, to test the model's performance across different subgroups. In Section 3.3, we explore the intrinsic diagnostic logic of the models through the developed visualization tool. Specifically, we generate saliency maps to identify the most important words in the symptom descriptions that contribute to the final syndrome differentiation results, thereby helping us understand the diagnostic logic learned by the developed AIAD method. In Section 3.4, we compare the performance of the baseline, AIAD-A1, and AIAD-A2 on the ITD dataset to verify the generalizability of our proposed models. Finally, in Section 3.5, we conduct extensive supplementary experiments to validate the framework's adaptability and robustness. Given our large label space of 161 syndromes, we first evaluate performance on truncated subsets to simulate specialized clinical settings. Then we verify technical robustness through sensitivity analyses across diverse NLP backbones, Large Language Models (LLMs), and loss function hyperparameters.

3.1 Syndrome differentiation performance

As summarized in Figure 4, the mediator alternative (AIAD-A2) demonstrated substantial improvements over the baseline. Specifically, AIAD-A2 increased Top-K accuracy by 3.60% to 6.31% across different K values, peaking at K=10 (rising from 65.19% to 71.50%). Similarly, NDCG@K saw consistent gains, with a maximum increase of 0.0505 at K=10. AIAD-A2 significantly outperformed the baseline across

all metrics ($p < 0.005$, see Appendix Table E3 in the supplementary material), validating the effectiveness of incorporating syndrome element diagnosis in syndrome differentiation.

In contrast, the AIAD-A1 method yielded only marginal and inconsistent improvements. Gains were minimal (e.g., merely 0.99% for Top-10 accuracy) and lacked broad statistical significance ($p > 0.05$ for most metrics), with Top-1 accuracy even exhibiting a slight decline. These results indicate that the inclusion of syndrome elements does not automatically guarantee performance gains, rather, their effectiveness is strictly contingent on the appropriateness of the model architecture. Direct paired comparisons reinforce this conclusion that AIAD-A2 significantly outperformed AIAD-A1 across all metrics ($p < 0.005$, Appendix Table E3 in the supplementary material), confirming the superiority of the mediator alternative over the intermediate variable alternative.

It is worth noting that despite the challenge of differentiating 161 syndromes, our model achieved a Top-15 accuracy of approximately 80%. This performance is competitive with established symptom checkers and sufficient to effectively narrow the diagnostic search space for clinical decision support (see Appendix D.3 in the supplementary material for detailed benchmarking).

To further verify the mediating role of syndrome elements, we conducted a causal mediation analysis [38, 39], revealing that the pathway *Symptom* $\rightarrow$ *Syndrome* serves as a robust foundational baseline while the pathway *Symptom*$\rightarrow$*Syndrome Element* $\rightarrow$ *Syndrome* acts as a critical semantic booster. These findings empirically substantiate that syndrome elements function as significant causal mediators in the process of syndrome differentiation. Detailed analysis and experimental setups are provided in Appendix F in the supplementary material.

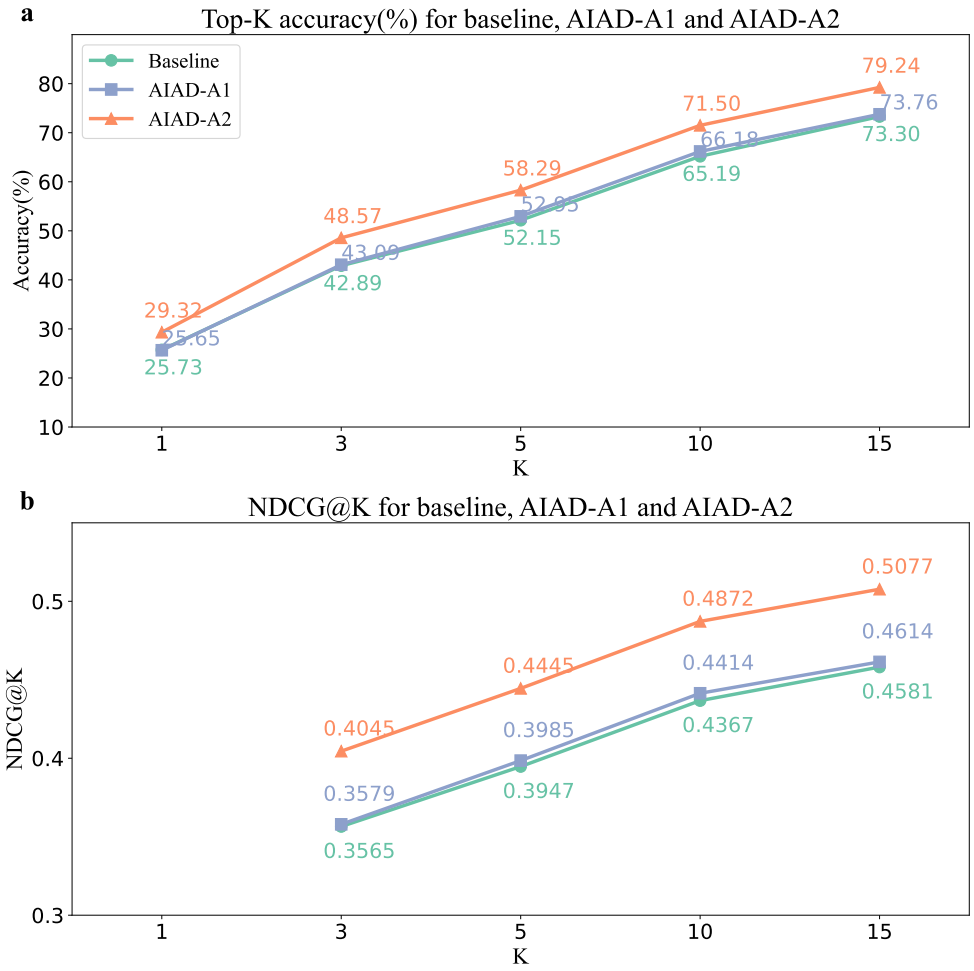


Fig. 4: **Model performance of the baseline, AIAD-A1 and AIAD-A2. a)** Model performance in terms of Top-K accuracy for $K = 1, 3, 5, 10, 15$ and **b)** Model performance in terms of NDCG@K for $K = 3, 5, 10, 15$.

3.2 Performance evaluation under long-tail distribution

To investigate the rationale behind AIAD-A2's superiority, we evaluated model performance across the head, middle, and tail subgroups of the NSTSPC dataset with results summarized in Table 1.

As discussed in Section 2.1.1, our NSTSPC dataset exhibits a long-tail distribution, which is a common scenario in practice. This phenomenon often leads to popularity

bias [40], where models excessively focus on classes with large sample sizes (head) while neglecting rare ones (tail), causing considerable performance degradation. Indeed, we observed dramatic performance declines in the analysis of tail subgroups for both the baseline and AIAD-A1 models. For instance, the Top-15 accuracy of the baseline plummeted from 91.39% in the head subgroup to 41.05% in the tail subgroup, confirming their vulnerability to the bias.

In contrast, AIAD-A2 demonstrates remarkable robustness. While showing slightly inferior performance in the head subgroup (a marginal decrease of $\approx 2\%$ across metrics), AIAD-A2 achieved consistent performance gains in the middle subgroup and substantial improvements in the tail subgroup. Notably, AIAD-A2 improved the Top-15 accuracy in the tail subgroup by 21.34% (from 41.05% to 62.39%) compared to the baseline. These results verify that AIAD-A2 effectively mitigates popularity bias, via its inherent mediator architecture, by wisely allocating attention to all classes rather than fitting only the majority.

This behavior aligns well with clinical reality. A TCM physician does not diagnose a syndrome simply based on its prevalence. Instead, as captured by the AIAD-A2 model, the diagnosis is reached by comprehensively considering unique symptoms and syndrome elements. This capability suggests that the mediator based framework better learns the intrinsic decision-making process of TCM diagnosis.

3.3 Unveiling diagnostic logic through visualization

To explore the intrinsic diagnostic logic of AIAD-A2, we utilized gradient based saliency maps to visualize the model's attention distribution assigned to each word (see Appendix G in the supplementary material for the technical formulation). Figure 5 displays a representative case presenting with symptoms such as "white spots" and "tongue body red with stasis spots".

Table 1: Top-K accuracy(%) of baseline, AIAD-A1 and AIAD-A2 on the head, middle and tail subgroups.

Metric	Model	Head subgroup		Middle subgroup		Tail subgroup	
		Acc.	**Imp.**	**Acc.**	**Imp.**	**Acc.**	**Imp.**
Top-1	Baseline	51.07		23.82		1.68	
	AIAD-A1	50.58	-0.49	23.66	-0.16	2.15	0.47
	AIAD-A2	49.06	-2.01	31.68	7.86	6.76	5.08
Top-3	Baseline	68.79		51.73		7.45	
	AIAD-A1	69.17	0.38	52.97	1.24	6.50	-0.95
	AIAD-A2	67.08	-1.71	59.13	7.4	18.97	11.52
Top-5	Baseline	76.69		64.31		14.76	
	AIAD-A1	77.40	0.71	67.31	3	13.47	-1.29
	AIAD-A2	75.17	-1.52	69.16	4.85	30.05	15.29
Top-10	Baseline	86.97		79.69		28.24	
	AIAD-A1	87.74	0.77	81.71	2.02	28.44	0.2
	AIAD-A2	84.31	-2.66	80.61	0.92	49.25	21.01
Top-15	Baseline	91.39		86.94		41.05	
	AIAD-A1	91.45	0.06	86.97	0.03	42.28	1.23
	AIAD-A2	89.35	-2.04	85.70	-1.24	62.39	21.34

Note: "Acc." denotes accuracy (%), "Imp." denotes improvement over baseline (percentage points).

Regarding symptom attention, both the baseline and AIAD-A1 were susceptible to interference from non-specific symptoms like "white spots" or "sweating", resulting in a misclassification of *liver-kidney insufficiency*. In contrast, AIAD-A2 successfully filtered out this noise and concentrated heavily on "stasis spots", the pivotal indicator for *blood stasis*. This demonstrates the superior capability of AIAD-A2 in capturing clinically determinative features. Furthermore, this visualization serves as a vital human-in-the-loop mechanism for clinical application. By exposing the model's focus, physicians can verify whether predictions rely on clinically relevant symptoms (e.g., "stasis spots") or irrelevant noise, thereby preventing blind acceptance of black-box errors.

Regarding syndrome element reasoning, we further examined the probability shifts depicted in Figure 5d. While the patient exhibited signs of both *liver-kidney insufficiency* and *qi stagnation and blood stasis* (see Appendix Figure A1 in the supplementary material for details), the latter constitutes the principal pathology requiring priority treatment. AIAD-A1 failed to distinguish this diagnostic hierarchy, assigning

high confidence to *liver* (1.00) and *kidney* (0.64) but lower confidence to *qi stagnation* (0.52). Crucially, AIAD-A2 rectified this bias. By leveraging the mediator structure, it significantly boosted the probabilities of the dominant pathogenic elements, specifically *blood stasis* ($0.63 \rightarrow 0.95$) and *qi stagnation* ($0.52 \rightarrow 0.69$), while suppressing secondary elements like *kidney* ($0.64 \rightarrow 0.27$). This explicit probability distribution serves as a vital mechanism for logical verification. It allows physicians to validate that the predicted syndrome is supported by consistent pathogenic elements rather than contradictory noise. This alignment with the dialectical perspective of "distinguishing the primary from the secondary" confirms that AIAD-A2 is different from conventional black-box models. Instead, it provides a clinically interpretable rationale that mirrors a physician's own dialectical reasoning, thereby facilitating trust and enabling safer clinical adoption.

3.4 Generalizability evaluation on ITD

Generalizability is a critical prerequisite for the safe deployment of medical AI systems. It reflects a model's ability to learn the intrinsic, invariant relationships between clinical features and diagnoses, rather than merely memorizing the specific distribution or noise patterns of the training data. When facing a distribution shift, a model lacking generalizability often suffers from performance collapse as it relies heavily on the bias of the source domain as prior knowledge. In contrast, a highly generalizable model can overcome these superficial statistical correlations, maintaining robustness by leveraging the underlying diagnostic logic to guide classification in unseen environments.

In this context, our test on the ITD dataset serves as an external validation. It is noteworthy that the ITD represents a canonical domain derived from guidelines, which differs intrinsically from the raw, noisy EHRs in the training set. While guidelines offer standardized symptom descriptions, this domain shift poses a distinct

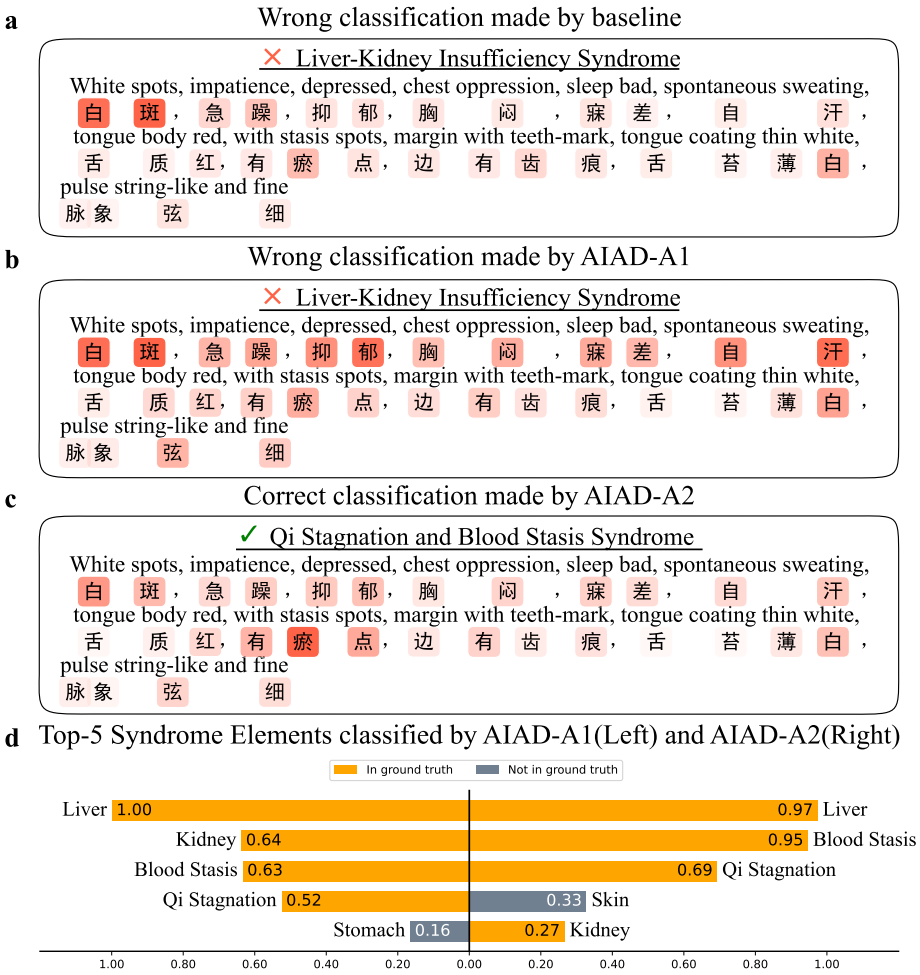


Fig. 5: **Saliency maps of the classification results of a representative sample.** **a)** Wrong classification made by baseline, **b)** Wrong classification made by AIAD-A1, and **c)** Correct classification made by AIAD-A2. **d)** Top-5 syndrome elements with the highest probabilities made by AIAD-A1 and AIAD-A2.

challenge of bridging the gap between noisy clinical reality and standardized theoretical knowledge. A model that merely memorizes specific phrasings or noise in the training EHRs would struggle to adapt to the generalized descriptions in the ITD.

Table 2: Performance comparison between the baseline, AIAD-A1 and AIAD-A2 methods in terms of top-K accuracy (%).

Model	Top 1	Top 3	Top 5	Top 10	Top 15
Baseline	25.73	42.89	52.15	65.19	73.30
Baseline (ITD)	9.14	20.81	29.54	41.36	50.60
AIAD-A1	25.65	43.09	52.95	66.18	73.76
AIAD-A1 (ITD)	13.02	28.53	37.65	52.09	61.32
AIAD-A2	29.32	48.57	58.29	71.50	79.24
AIAD-A2 (ITD)	16.35	32.70	44.56	62.06	71.47

Note: '(ITD)' indicates that the model was tested on the ITD dataset.

Therefore, strong generalization to the ITD would indicate that a model has successfully captured the fundamental diagnostic logic shared between clinical practice and theoretical guidelines, rather than overfitting to specific noise.

The performance stability across this domain shift is summarized in Table 2. The baseline model exhibited a significant collapse, with Top-15 accuracy dropping by 22.7% (from 73.30% to 50.60%). AIAD-A1 showed moderate stability with a 12.44% drop. In contrast, AIAD-A2 demonstrated remarkable resilience, suffering only a minor 7.77% decrease (from 79.24% to 71.47%). This superior stability suggested that by incorporating syndrome elements as mediators, AIAD-A2 effectively bridged the gap between clinical reality and theoretical standards.

Further subgroup analysis (Table 3) reinforced this conclusion. While AIAD-A1 showed balanced improvements, AIAD-A2 exhibited a distinct advantage in the tail subgroup, achieving a 33.38% improvement over the baseline. This confirms that AIAD-A2's robust capacity to mitigate popularity bias and handle rare syndromes effectively enables it to maintain superior generalizability, even under significant distribution shifts.

3.5 Robustness evaluation and sensitivity analysis

To further validate the clinical applicability and technical robustness of the AIAD-A2 framework, we conducted extensive supplementary experiments. First, to simulate specialized clinical settings (e.g., specific departments) where the diagnostic scope is

Table 3: Top-K accuracy (%) of the baseline, AIAD-A1 and AIAD-A2 methods on the head, middle and tail subgroups of the ITD dataset.

Metric	Model	Head subgroup		Middle subgroup		Tail subgroup	
		Acc.	Imp.	Acc.	Imp.	Acc.	Imp.
Top-1	Baseline	36.67		9.08		0.08	
	AIAD-A1	42.02	5.35	17.06	7.98	0.39	0.31
	AIAD-A2	38.93	2.26	26.04	16.96	1.61	1.53
Top-3	Baseline	57.26		31.02		1.10	
	AIAD-A1	69.17	11.91	43.99	12.97	3.50	2.40
	AIAD-A2	64.29	7.03	52.07	21.05	7.72	6.62
Top-5	Baseline	66.55		47.77		3.62	
	AIAD-A1	79.05	12.50	60.73	12.96	6.65	3.03
	AIAD-A2	74.88	8.33	70.18	22.41	15.31	11.69
Top-10	Baseline	78.21		66.77		10.12	
	AIAD-A1	87.38	9.17	82.36	15.59	17.72	7.60
	AIAD-A2	84.29	6.08	86.25	19.48	36.57	26.45
Top-15	Baseline	82.98		80.21		17.68	
	AIAD-A1	90.36	7.38	90.55	10.34	29.80	12.12
	AIAD-A2	87.74	4.76	91.50	11.29	51.06	33.38

Note: "Acc." denotes accuracy (%), "Imp." denotes improvement over baseline (percentage points).

naturally limited, we evaluated the models on truncated subsets comprising the Top-20 and Top-40 most prevalent syndromes in the training set. AIAD-A2 demonstrated superior adaptability in these focused tasks, achieving Top-5 accuracies of 84.38% and 87.66% on the NSTSPC and ITD datasets, respectively (see Appendix H.1 in the supplementary material). Second, we performed comprehensive sensitivity analyses to assess model robustness. These included benchmarking against diverse state-of-the-art TCM diagnostic methods and NLP backbones, comparing against LLMs (GPT-4o, Qwen2.5-72B) in a zero-shot setting, and verifying stability with respect to the loss function hyperparameter. The results consistently confirmed that AIAD-A2 achieved state-of-the-art performance and that its performance gains were robust across architectural variations and hyperparameter settings (see Appendix H.2 in the supplementary material).

4 Discussion

In this study, we developed the AIAD framework to address the challenge of intelligent syndrome differentiation in TCM. Unlike conventional "black-box" models, our

approach explicitly integrates syndrome elements into the diagnostic reasoning process. Through extensive evaluations, we empirically demonstrated that the mediator alternative (AIAD-A2) consistently outperforms both the baseline and the intermediate variable alternative (AIAD-A1). A pivotal theoretical finding is the validation of the mediator alternative: while syndrome elements function as structured summaries capturing the core pathology, they fail to capture the subtle nuances present in the raw symptom descriptions. Consequently, the superior performance of AIAD-A2 confirms that accurate diagnosis necessitates the joint modeling of structured syndrome elements and the fine-grained semantic details inherent in raw symptom descriptions.

Our work advances existing TCM diagnostic research in terms of scale, methodology, and capability to address critical clinical challenges. Prior studies [20–22] typically relied on manually extracted features and focused on a limited scope of syndromes (often fewer than 10). Such approaches usually overlook the popularity bias inherent in real-world data and lack transparency. In contrast, our framework leverages deep learning to process unstructured clinical text directly and successfully scales to 161 syndromes. Crucially, AIAD-A2 effectively tackles the long-tail distribution problem. While baseline models struggled on rare conditions, our approach achieved a substantial 21.34% accuracy improvement in the tail subgroup, demonstrating remarkable robustness against class imbalance. Furthermore, we bridge the gap between "black-box" AI and clinical trust through visualization. By generating saliency maps that highlight key symptoms, our method reveals the underlying diagnostic logic, offering interpretability that is absent in many deep learning models. Moreover, we validated the model on truncated subsets (Top-20/40 prevalent syndromes), proving superior adaptability even in the focused settings adopted by traditional methods. Additionally, by benchmarking against the state-of-the-art LLMs (GPT-4o, Qwen2.5), we demonstrated that general-purpose LLMs still fall behind our domain-specialized framework, validating the necessity of our dedicated training approach.

From a clinical perspective, the framework exhibits robust generalizability in the presence of domain shifts. Notably, our validation on the ITD provides critical evidence of this capability. The ITD represents a canonical domain derived from guidelines, differing significantly from the noisy, empirical EHRs used for training. While the baseline model struggled with this shift (suffering a 22.7% drop in accuracy), AIAD-A2 demonstrated remarkable stability (declining by only 7.77%). This resilience suggests that incorporating syndrome elements helps the model capture the intrinsic diagnostic logic shared between clinical practice and theoretical standards, rather than merely overfitting to the writing styles or noise of specific hospitals or physicians.

Based on these findings, we propose specific recommendations for the clinical application of AI-TCM systems. First, regarding clinical workflow, given the model's high Top-K accuracy, we recommend deploying AIAD-A2 to rapidly narrow down the search space from hundreds of syndromes to a focused candidate list, assisting junior physicians in avoiding omissions. Second, regarding medical education, the model's interpretability features (e.g. saliency maps) can serve as pedagogical aids, visualizing key symptoms to teach diagnostic reasoning. Finally, for future AI development, we strongly advocate for the mediator based architecture, as our results prove that preserving raw symptom semantics alongside structured knowledge is essential for robustness.

Our study has several limitations that highlight future research directions. First, although the ITD dataset confirms alignment with guidelines, our current evaluation may not fully equate to validation on unseen real-world EHRs. Future work should involve prospective validation on larger, multi-center independent cohorts to verify the model's robustness against the noise and variability of real-world clinical settings. Second, our framework currently relies solely on textual descriptions. Since TCM diagnosis inherently integrates visual (tongue diagnosis) and tactile (palpation) signals, relying on text alone presents as a bottleneck. Future applications should

aim to integrate multimodal deep learning (e.g., incorporating tongue images and pulse waveforms) to achieve a more holistic modeling of patient conditions. Finally, from a theoretical perspective, our model is grounded in syndrome elements derived from expert consensus. While this provides a standardized foundation, the theory of TCM is dynamic and evolving. Future research could leverage data-driven approaches to discover latent, previously unidentified syndrome elements from massive EHRs, thereby not only automating diagnosis but also actively expanding and enriching the theoretical system of TCM.

5 Conclusion

In this study, we propose a novel AI-based TCM diagnostic method that integrates syndrome elements into the TCM syndrome differentiation process. Our findings demonstrate both the feasibility of achieving objective TCM diagnoses with AI assistance and the crucial role of syndrome elements in this diagnostic framework. Our study enriches TCM diagnostic theory by treating syndrome elements as mediators. We show that, with a carefully selected taxonomy, such as the syndrome elements used in our study, the previously enigmatic TCM diagnosis can be partially demystified. This paves the way for future studies aimed at modernizing TCM diagnostic theory.

Acknowledgements. This study was supported by the grants from the Science and Technology Innovation Key R&D Program of Chongqing (No. CSTB2024TIAD-STX0040); the State Key Research Program (No. 2021YFF0901400); the National Natural Science Foundation of China (No. 82173623, No. 72271060).

List of abbreviations

AI: Artificial Intelligence, AIAD: Artificial Intelligence Aided Diagnosis, BERT: Bidirectional Encoder Representations from Transformers, CACM: China Association of Chinese Medicine, CAD: Computer Aided Diagnosis , EHR: Electronic Health Record,

ITD: Independent Test Dataset, LLM: Large Language Model, MTL: Multi-task Learning, NLP: Natural Language Process, NSTSPC: National Science and Technology Support Program of China, TCM: Traditional Chinese Medicine, WM: Western Medicine

Declarations

- Funding: This study was supported by the grants from the Science and Technology Innovation Key R&D Program of Chongqing (No. CSTB2024TIAD-STX0040); the State Key Research Program (No. 2021YFF0901400); the National Natural Science Foundation of China (No. 82173623, No. 72271060).

- Conflict of interest: Xinxing Lai serves as a member of the Youth Editorial Board of Chinese Medicine. The author was not involved in the review or decision-making process for this manuscript. The authors declare no other conflicts of interest.
- Ethics approval and consent to participate: Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants or patients/participants legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.
- Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.
- Author contribution: **Yan Lyu**: Data curation, Formal analysis, Methodology, Visualization, Writing - original draft. **Xiangnan Feng**: Formal analysis, Methodology, Validation, Writing - original draft. **Ping Jiang**: Validation, Writing - original draft. **Haoxuan Li**: Methodology, Writing – review and editing. **Xinxing Lai**: Conceptualization, Supervision, Writing - review and editing. **Xiao-Hua Zhou**: Conceptualization, Funding acquisition, Supervision, Writing - review and editing.

References

[1] Cheung F. TCM: made in China. Nature. 2011;480(7378):S82–S83.

[2] Lai X, Dong Z, Wu S, Zhou X, Zhang G, Xiong S, et al. Efficacy and safety of Chinese herbal medicine compared with losartan for mild essential hypertension: A randomized, multicenter, double-blind, noninferiority trial. Circulation: Cardiovascular Quality and Outcomes. 2022;15(3):e007923.

[3] Wu L, Song H, Zhang C, Wang A, Zhang B, Xiong C, et al. Efficacy and safety of Panax notoginseng saponins in the treatment of adults with ischemic stroke in China: A randomized clinical trial. JAMA Network Open. 2023;6(6):e2317574–e2317574.

[4] Bach RG. Traditional Chinese medicine meets evidence-based medicine in the acutely infarcted heart. JAMA. 2023;330(16):1529–1530.

[5] Tian Z, Zhang S. Integrating traditional and modern medicines to treat heart failure. Nature Medicine. 2024;30(8):2129–2130.

[6] Yan E, Song J, Liu C, Hong W. A research on syndrome element differentiation based on phenomenology and mathematical method. Chinese Medicine. 2017;12:1–18.

[7] Sun Dz, Li Sd, Liu Y, Zhang Y, Mei R, Yang Mh. Differences in the origin of philosophy between Chinese medicine and western medicine: exploration of the holistic advantages of Chinese medicine. Chinese journal of integrative medicine. 2013;19:706–711.

[8] Qiu J. When the East meets the West: the future of traditional Chinese medicine in the 21st century. National Science Review. 2015;2(3):377–380.

[9] Xu Q, Bauer R, Hendry BM, Fan TP, Zhao Z, Duez P, et al. The quest for modernisation of traditional Chinese medicine. BMC Complementary and Alternative Medicine. 2013;13:1–11.

[10] Jiang M, Lu C, Zhang C, Yang J, Tan Y, Lu A, et al. Syndrome differentiation in modern research of traditional Chinese medicine. Journal of ethnopharmacology. 2012;140(3):634–642.

[11] Song G, Wang Y, Zhang R, Liu B, Zhou X, Zhou X, et al. Experience inheritance from famous specialists based on real-world clinical research paradigm of traditional Chinese medicine. Frontiers of medicine. 2014;8:300–309.

[12] Liu Y, Jiang Jj, Du Sy, Mu Ls, Fan Jj, Hu Jc, et al. Artemisinins ameliorate polycystic ovarian syndrome by mediating LONP1-CYP11A1 interaction. Science. 2024;384(6701):eadk5382.

[13] Wang L, Zhou GB, Liu P, Song JH, Liang Y, Yan XJ, et al. Dissection of mechanisms of Chinese medicinal formula Realgar-Indigo naturalis as an effective treatment for promyelocytic leukemia. Proceedings of the National Academy of Sciences. 2008;105(12):4826–4831.

[14] Gou H, Su H, Liu D, Wong CC, Shang H, Fang Y, et al. Traditional medicine Pien Tze Huang suppresses colorectal tumorigenesis through restoring gut microbiota and metabolites. Gastroenterology. 2023;165(6):1404–1419.

[15] Pan D, Guo Y, Fan Y, Wan H. Development and Application of Traditional Chinese Medicine Using AI Machine Learning and Deep Learning Strategies. The American Journal of Chinese Medicine. 2024;52(03):605–623.

[16] Zhou W, Yang K, Zeng J, Lai X, Wang X, Ji C, et al. FordNet: recommending traditional Chinese medicine formula via deep neural network integrating phenotype

592 and molecule. Pharmacological research. 2021;173:105752.

593 [17] Cheng F, Wang X, Song W, Lu Y, Li X, Zhang H, et al. Biologic basis of
594 TCM syndromes and the standardization of syndrome classification. Journal of
595 Traditional Chinese Medical Sciences. 2014;1(2):92–97.

596 [18] WHO. WHO international standard terminologies on traditional Chinese
597 medicine. World Health Organization; 2022.

598 [19] Lu AP, Jia HW, Xiao C, Lu QP. Theory of traditional Chinese medicine
599 and therapeutic method of diseases. World journal of gastroenterology: WJG.
600 2004;10(13):1854.

601 [20] Huang Z, Miao J, Chen J, Zhong Y, Yang S, Ma Y, et al. A traditional Chi-
602 nese medicine syndrome classification model based on cross-feature generation by
603 convolution neural network: model development and validation. JMIR medical
604 informatics. 2022;10(4):e29290.

605 [21] Ding L, Zhang Xy, Wu Dy, Liu Ml. Application of an extreme learning machine
606 network with particle swarm optimization in syndrome classification of primary
607 liver cancer. Journal of Integrative Medicine. 2021;19(5):395–407.

608 [22] Xie J, Li Y, Wang N, Xin L, Fang Y, Liu J. Feature selection and syndrome
609 classification for rheumatoid arthritis patients with Traditional Chinese Medicine
610 treatment. European Journal of Integrative Medicine. 2020;34:101059.

611 [23] Zhu W. Establishing a new system of syndrome differentiation with syndrome
612 element as core. J Hunan Univ Chin Med. 2004;24(6):38–39.

613 [24] Wang J, Duan L, Li H, Liu J, Chen H. Construction of an artificial intelligence
614 traditional Chinese medicine diagnosis and treatment model based on syndrome

elements and small-sample data. Engineering. 2022;8:29–32.

[25] Zhang Y, Yang Q. An overview of multi-task learning. National Science Review. 2018;5(1):30–43.

[26] Thung KH, Wee CY. A brief review on multi-task learning. Multimedia Tools and Applications. 2018;77(22):29705–29725.

[27] Li Z, Zeng L, Liang Z, Zhang Y, Chen Q, Fu W. The development of Chen's flying needle based on the traditional Chinese Medicine: A needle-inserting technique created by a distinguished veteran doctor of TCM. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE; 2012. p. 444–447.

[28] Zhao Fy, Zhang Zy, Xia Xj, Hong YF, Zhao Yx, Yan Hx, et al. Memory response to manual acupuncture in chronic insomniacs: evidence from event-related potentials. Journal of Acupuncture and Tuina Science. 2018;16(6):382–388.

[29] Anderson C. The long tail: Why the future of business is selling less of more. Hachette UK; 2006.

[30] Kamper SJ. Generalizability: linking evidence to practice. journal of orthopaedic & sports physical therapy. 2020;50(1):45–46.

[31] Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nature medicine. 2019;25(9):1337–1340.

[32] Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. Practical guidance on artificial intelligence for health-care data. The Lancet Digital Health. 2019;1(4):e157–e159.

[33] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018;.

[34] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.

[35] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020;.

[36] Seltzer ML, Droppo J. Multi-task learning in deep neural networks for improved phoneme recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2013. p. 6965–6969.

[37] Järvelin K, Kekäläinen J. IR evaluation methods for retrieving highly relevant documents. In: ACM SIGIR Forum. vol. 51. ACM New York, NY, USA; 2017. p. 243–250.

[38] Pearl J. Direct and indirect effects. In: Probabilistic and causal inference: the works of Judea Pearl; 2022. p. 373–392.

[39] Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychological methods. 2010;15(4):309.

[40] Zhang Y, Feng F, He X, Wei T, Song C, Ling G, et al. Causal intervention for leveraging popularity bias in recommendation. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval; 2021. p. 11–20.

Appendix A Dataset Description

A.1 NSTSPC Dataset

A.1.1 Data collection.

In the present study, we used the TCM EHRs from a clinical medical record database of senior TCM physicians supported by the *NSTSPC, No. 2007BAI10B01*. The NSTSPC database was constructed under a national project that extensively collects clinical and diagnosis information from senior TCM physicians in different hospitals and forms a database of representative EHRs for further research. One objective of constructing such a database is to summarize the academic insights and clinical experience from these TCM physicians by analyzing the EHRs of patients that treated by these TCM physicians using modern information technologies. In this process, we can study their dialectical thinking characteristics, explore the rationales behind their therapeutic effects, and summarize the theoretical basis for TCM diagnosis. This can improve the reproducibility, learnability and generalizability of TCM diagnosis. The NSTSPC database includes 210 senior TCM physicians from 75 hospitals around China, inlcuding Beijing, Tianjin, Hebei, Shanxi, Nei Menggu, Liaoning, Jilin, Heilongjiang, Shanghai, Jiangsu, Zhejiang, Anhui, Fujian, Jiangxi, Shandong, Henan, Hubei, Hunan, Guangdong, Guangxi, Hainan, Sichuang, Chongqing, Guizhou, Yunnan, Xizang, Shaanxi, Gansu, Qinghai and Xinjiang. More than 40,000 EHRs were collected in the database. These EHRs are representative cases handled by these TCM physicians and have been regarded as a gold standard in the literature [41–45]. A graphical illustration for the data collection process is given in Figure A1a.

Each EHR in the NSTSPC database represents the health record of a patient, which contains the symptom description (corresponds to the information collected from the Four Diagnostic Methods in Figure 1 in the main paper), syndrome name

(corresponds to the Syndrome Differentiation in Figure 1 in the main paper), and syndrome differentiation analysis report (corresponds to the Diagnosis Making Process of TCM in Figure 1 in the main paper) of this patient. Figure A1b provides an example of EHR. Note that the symptom description and the syndrome differentiation analysis report are in unstructured text format. The symptom description (the second panel in Figure A1b) records the information of symptoms collected from the four diagnostic methods. The TCM physician performs syndrome differentiation based on the information in this description. The syndrome name (the first panel in Figure A1b) is the diagnosis result given by the TCM physician. The syndrome differentiation analysis report (the third panel in Figure A1b) records the detailed analyzing procedure of how the physician diagnoses a syndrome based on the symptom description. This report is thus valuable because it contains detailed description and reflection of the TCM diagnosis procedure, from which the true labels (ground truth) of syndrome elements can be manually extracted. This is one particular advantage of the NSTSPC database and the foundation for our study. Therefore, EHRs with full records of syndrome name, symptom description, and syndrome differentiation analysis report are included in our study. The detailed inclusion and exclusion criteria are discussed below.

A.1.2 Inclusion and exclusion criteria.

First, we excluded EHRs with missing values in symptom description, syndrome name, or syndrome differentiation analysis report. This is because we take each patient's symptom description as the input and the syndrome name as the outcome, and rely on the syndrome differentiation analysis report to extract true labels of syndrome elements. Second, to ensure the consistency of the syndrome names among EHRs, we further standardized the syndrome names according to the standard of *Clinic Terminology of Traditional Chinese Medical Diagnosis and Treatment Diseases* (GB/T 16751.2-2021) and discard the EHRs with non-standardizable syndrome names. Third, to keep a minimal number of training and testing samples for each syndrome, we have

to discard the syndromes with less than 10 records. As a result, a total of 6,226 samples (EHRs) are included in this study (2589 males, 3637 females; mean age: 44±19 years), involving 161 different syndromes. A graphical illustration for the data collection process is given in Figure A1a. A full list of the syndrome names included in this study is provided in Appendix B.

A.1.3 Syndrome element extraction.

The primary preprocessing step in this study is to extract true labels of syndrome elements from the syndrome differentiation analysis report. We followed a procedure similar to that described in [46]. Specifically, 51 potential syndrome elements were determined by five senior TCM physicians (see Appendix B for the full list of syndrome elements). For each syndrome element, a set of keywords was determined. Subsequently, for each EHR, an automated text search method was employed to search for these keywords in the syndrome differentiation analysis report. When a specific keyword is found, the corresponding syndrome element is considered present in that EHR. We further verified the obtained syndrome elements manually to ensure accuracy. As a result, each EHR may be associated with 0 to 16 different syndrome elements (see the third panel in Figure A1b for an example). This showcases the complex relationship between syndrome elements and syndromes.

A.1.4 Diagnostic Coverage

To further evaluate the diagnostic coverage of the NSTSPC dataset, we systematically examined both the Western Medicine (WM) diagnostic records and TCM disease entities for each EHR in the dataset. Given that the original WM diagnoses were recorded as unstructured free text, we implemented a keyword based semantic mapping strategy to categorize these records into standard ICD-11 chapters (e.g., mapping terms such as 'stomach' to *Chapter 13 Diseases of the digestive system*). To intuitively visualize the hierarchical cross-domain relationships, we constructed a three-layer Sankey

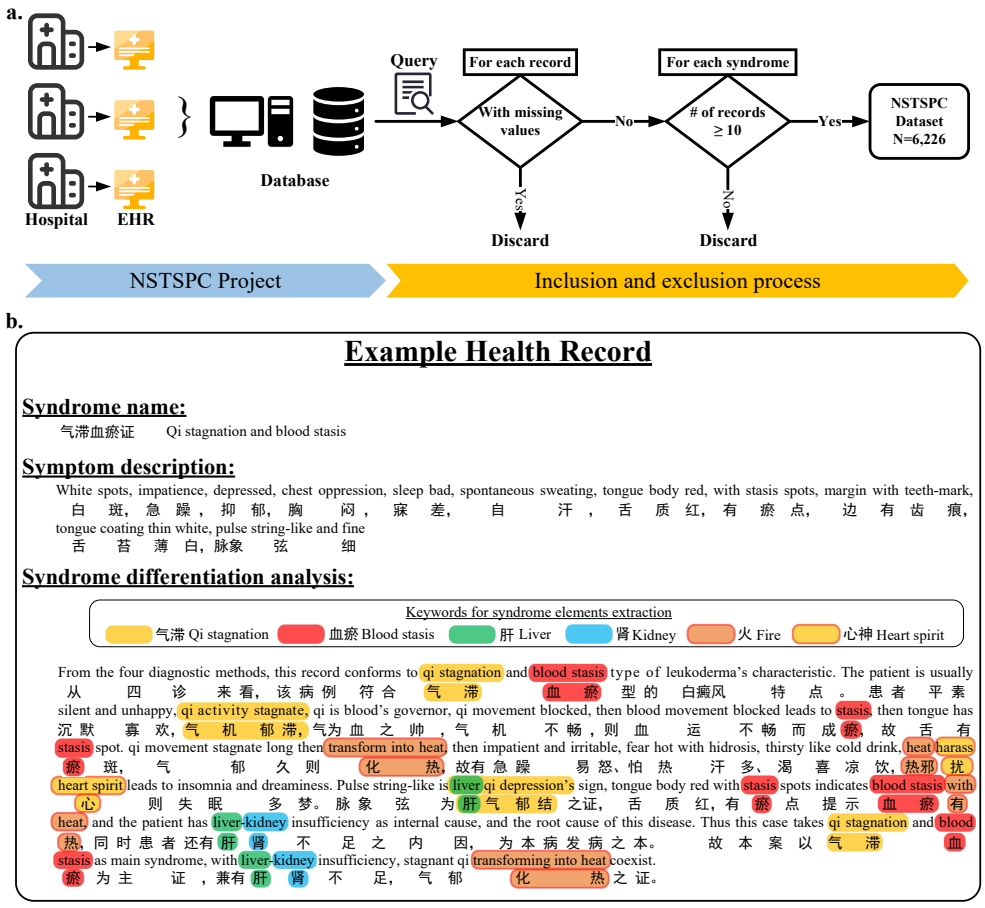


Fig. A1: Data collection process. a) Demonstration of the collection, as well as the inclusion and exclusion criteria of our NSTSPC dataset. b) An example of health record of the NSTSPC dataset. The keywords in the syndrome differentiation analysis report used in the syndrome element extraction are emphasized using different colors.

diagram (Figure A2), in which the most prevalent TCM disease entities are labeled individually while the long-tail distribution of less frequent diseases is aggregated into an "Other TCM Diseases" category. In the Sankey diagram, a continuous flow connecting the Western Medicine diagnosis in the left layer, the TCM disease entity in the middle layer, and the TCM syndrome differentiation in the right layer signifies that a patient was simultaneously diagnosed with these three conditions within the

same EHR. As illustrated in Figure A2, the dataset demonstrates extensive diagnostic coverage, spanning a wide spectrum of major physiological systems including the digestive, circulatory, respiratory, endocrine, and nervous systems. The flow visualization explicitly validates the TCM principle of "One Syndrome, Multiple Diseases", revealing that specific syndromes are associated with diverse Western pathological sources and TCM disease entities. Moreover, the substantial patient flow through the "Other TCM Diseases" category highlights that the training data encompasses a significant long-tail distribution of rare clinical conditions, indicating that the proposed model possesses strong generalizability across complex, heterogeneous real-world settings. Collectively, these analyses empirically validate the comprehensive diagnostic coverage of the NSTSPC dataset, confirming that it captures a sufficiently broad range of clinical patterns to support robust syndrome differentiation.

A.2 Independent test dataset

All the EHRs in the ITD are extracted from TCM clinical guidelines issued by CACM including *Clinical Guidelines for Diagnosis and Treatment of Internal Diseases in TCM*, *Guidelines for Diagnosis and Treatment of Common Internal Diseases In Chinese Medicine*, *Guidelines for Diagnosis and Treatment of Common Diseases of Surgery In Traditional Chinese Medicine*, *Guidelines for Diagnosis and Treatment of Common Diseases of Gynecology In Traditional Chinese Medicine*, *Guidelines for Diagnosis and Treatment of Common Diseases of Pediatrics In Traditional Chinese Medicine*, *Guidelines for Diagnosis and Treatment of Common Diseases of Orthopedics and Traumatology In Traditional Chinese Medicine*, *Guidelines for Diagnosis and Treatment of Common Diseases of Spinal Orthopedics In Traditional Chinese Medicine*, *Guidelines for Diagnosis and Treatment of Common Diseases of Ophthalmology In Traditional Chinese Medicine*, *Guidelines for Diagnosis and Treatment of Common Diseases of Otolaryngology In Traditional Chinese Medicine*, *Guidelines for*

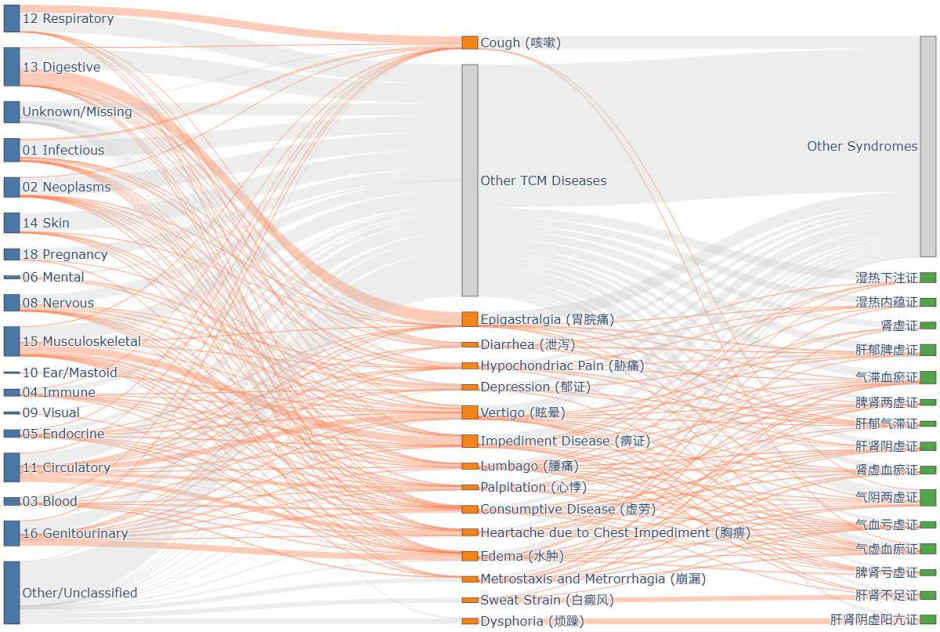


Fig. A2: Diagnostic coverage and pattern flow visualization with a three-layer Sankey diagram illustrating the hierarchical diagnostic distribution. Left: Patients categorized by ICD-11 System (Blue) based on semantic mapping of western medicine diagnoses. Middle: TCM disease entities, with the most prevalent diseases labeled individually (Orange) and the long-tail distribution aggregated into "Other TCM Diseases" (Gray). Right: The distribution of TCM Syndromes (Green). The diagram demonstrates the extensive coverage across multiple body systems and the complex mapping patterns between Western diagnoses and TCM syndromes.

Diagnosis and Treatment of Common Diseases of Dermatology In Traditional Chinese Medicine and *Guidelines for Diagnosis and Treatment of Common Diseases of Coloproctology In Traditional Chinese Medicine*. Specifically, these guidelines list typical syndromes and provide one example of common symptoms for each syndrome, which can be regarded as the symptom descriptions. Therefore, we can treat the syndrome name and the corresponding symptom description as a single EHR. The symptom descriptions are thoroughly reviewed by experts to ensure that they are typical, standardized, and accurate, resulting in a high-quality dataset for the validation purpose. Note that the clinical guidelines do not include syndrome differentiation

analysis reports. We cannot extract true labels of syndrome elements from the ITD. Therefore, the ITD can only be used as an independent testing dataset.

A total of 1,057 EHRs, corresponding to the 161 syndromes in NSTSPC dataset, were collected in ITD. Each EHR contains a symptom description as model input and syndrome name as the diagnosis result. The syndrome names were also standardized according to the standard GB/T 16751.2-2021.

We highlight that the guidelines provide only a few typical records for each syndrome. Consequently, the distribution of the ITD is significantly different from that of the NSTSPC dataset, posing a great challenge to the generalizability of our AIAD method trained on the NSTSPC dataset. To illustrate this, we also partitioned the ITD into the head, middle, and tail subgroups, according to the syndromes in each subgroup of the NSTSPC dataset. If share a similar distribution, each subgroup for ITD should approximately account for 1/3 of all the EHRs in ITD. However, the pie chart in Figure 2b shows that the proportions of the three subgroups in ITD are 15.9%, 36.0%, and 48.1%, respectively. This indicates the significant distributional differences between the two datasets. If our AIAD method, trained on the NSTSPC dataset, also performs well on the ITD, we may conclude that the method exhibits good generalizability and can be applied in practice.

Appendix B Lists of syndromes and syndrome elements

Table B1: List of the included 161 syndrome names in Chinese and English. The English translations are collected from GB/T 16751.2-2021.

ID	Syndrome Name (ZH)	Syndrome Name (EN)
1	气滞血瘀证	Qi Stagnation and Blood Stasis
2	肝郁脾虚证	Liver Depression and Spleen Deficiency
3	气阴两虚证	Qi and Yin Deficiency

Continued on next page

Table B1 – Continued from previous page

ID	Syndrome Name (ZH)	Syndrome Name (EN)
4	湿热下注证	Dampness and Heat Pouring Downwards
5	肝肾不足证	Liver and Kidney Insufficiency
6	肝肾阴虚阳亢证	Yin Deficiency and Yang Hyperactivity in Liver and Kidney
7	气虚血瘀证	Qi Deficiency with Blood Stasis
8	湿热内蕴证	Internal Accumulation of Dampness and Heat
9	肝肾阴虚证	Yin Deficiency in Liver and Kidney
10	肾虚证	Kidney Deficiency
11	肾虚血瘀证	Blood Stasis due to Kidney Deficiency
12	气血亏虚证	Qi and Blood Depletion
13	脾肾两虚证	Dual Deficiency of Spleen and Kidney
14	肝郁气滞证	Qi Stagnation due to Liver Depression
15	脾肾亏虚证	Spleen and Kidney Insufficiency
16	脾虚湿困证	Spleen Deficiency with Dampness Encumbrance
17	脾肾阳虚证	Yang Deficiency in Spleen and Kidney
18	痰热壅肺证	Phlegm and Heat Congesting in the Lung
19	肝阳上亢证	Liver Yang Hyperactivity
20	脾胃虚证	Spleen and Stomach Deficiency
21	热毒炽盛证	Intense Exuberant Heat Toxin
22	肝胆湿热证	Dampness and Heat in the Liver and Gallbladder
23	肾阴虚证	Kidney Yin Deficiency
24	肝郁肾虚证	Liver Depression and Kidney Deficiency
25	阴虚火旺证	Yin Deficiency with Effulgent Fire
26	肝胃不和证	Disharmony of Liver and Stomach
27	肝气郁结证	Liver Qi Depression
28	湿热瘀阻证	Dampness and Heat Stagnation and Obstruction
29	心脾两虚证	Dual Deficiency of the Heart and Spleen
30	阴虚热盛证	Yin Deficiency with Exuberant Heat
31	湿热蕴结证	Dampness and Heat Accumulation and Binding

Continued on next page

Table B1 – Continued from previous page

ID	Syndrome Name (ZH)	Syndrome Name (EN)
32	瘀血阻络证	Stasis Obstructing Collaterals
33	寒邪犯肺证	Pathogenic Cold Invading the Lung
34	肝气犯胃证	Liver Qi Invading the Stomach
35	湿热中阻证	Dampness and Heat Obstructing the Middle
36	痰瘀互结证	Phlegm and Stasis Obstructing the Internal
37	气血两虚证	Dual Deficiency of Qi and Blood
38	肝胃郁热证	Stagnant Heat in the Liver and Stomach
39	胃阴虚证	Stomach Yin Deficiency
40	肝肾亏虚证	Liver and Kidney Depletion
41	风热犯肺证	Wind and Heat Invading the Lung
42	肝郁化火证	Stagnated Liver Qi Transforming Into Fire
43	脾胃阳虚证	Yang Deficiency in Spleen and Stomach
44	热邪壅肺证	Pathogenic Heat Obstructing the Lung
45	肾气虚证	Kidney Qi Deficiency
46	中气不足证	Middle Qi Insufficiency
47	脾虚证	Spleen Deficiency
48	痰热搏结证	Combat and Binding of Phlegm and Heat
49	寒湿阻络证	Cold and Dampness Obstructing Collaterals
50	痰热闭肺证	Phlegm and Heat Blocking the Lung
51	脾胃湿热证	Dampness and Heat in the Spleen and Stomach
52	阴虚阳亢证	Yin Deficiency and Yang Hyperactivity
53	血热妄行证	Frenetic Movement of Blood due to Heat
54	肺脾气虚证	Lung and Spleen Qi Deficiency
55	痰热阻肺证	Phlegm and Heat Obstructing the Lung
56	肾阴阳两虚证	Dual Deficiency of Kidney Yin and Yang
57	血瘀证	Blood Stasis
58	经气不利证	Inhibited Flow of Meridian Qi
59	肺肾两虚证	Dual Deficiency of the Lung and Kidney

Continued on next page

Table B1 – Continued from previous page

ID	Syndrome Name (ZH)	Syndrome Name (EN)
60	风寒湿痹阻证	Wind and Cold and Dampness Impediment
61	血虚风燥证	Wind Dryness due to Blood Deficiency
62	气虚证	Qi Deficiency
63	肝脾不调证	Liver and Spleen Disharmony
64	脾虚气滞证	Qi Stagnation due to Spleen Deficiency
65	血热风燥证	Blood Heat with Wind Dryness
66	肾精亏虚证	Kidney Essence Depletion
67	肾阳虚证	Kidney Yang Deficiency
68	气阴亏虚证	Qi and Yin Depletion
69	肝郁血虚证	Blood Deficiency due to Liver Depression
70	风热证	Wind and Heat
71	肝热血瘀证	Blood Stasis due to Liver Heat
72	阴虚血热证	Yin Deficiency and Blood Heat
73	湿热阻痹证	Dampness and Heat Obstruction and Impediment
74	肝郁血瘀证	Blood Stasis due to Liver Depression
75	血热证	Blood Heat
76	寒凝血瘀证	Cold Congealing and Blood Stasis
77	风热外袭证	Wind and Heat Attacking the External
78	肾气不足证	Kidney Qi Deficiency
79	痰湿蕴肺证	Phlegm and Dampness Accumulating in the Lung
80	肺经风热证	Wind and Heat in the Lung Meridian
81	风寒阻络证	Wind and Cold Obstructing Collaterals
82	寒热错杂证	Cold and Heat Complex
83	脾虚痰湿证	Phlegm Dampness due to Spleen Deficiency
84	脾气虚证	Spleen Qi Deficiency
85	痰湿内蕴证	Phlegm and Dampness Retain in the Internal
86	痰湿中阻证	Phlegm and Dampness Obstructing the Middle
87	湿热瘀滞证	Dampness and Heat Stagnation

Continued on next page

Table B1 – Continued from previous page

ID	Syndrome Name (ZH)	Syndrome Name (EN)
88	冲任瘀阻证	Thoroughfare Vessel and Conception Vessel Obstruction
89	心肾不交证	Heart and Kidney Failure in Interaction
90	火毒证	Fire Toxin
91	寒湿痹阻证	Cold and Dampness Impeding and Obstruction
92	心阴不足证	Heart Yin Insufficiency
93	痰气证	Phlegm and Qi
94	肝旺脾虚证	Liver Effulgence and Spleen Deficiency
95	痰热蕴结证	Accumulation and Binding of Phlegm and Heat
96	痰浊中阻证	Middle Obstruction of Turbid Phlegm
97	阴虚血瘀证	Yin Deficiency and Blood Stasis
98	瘀闭血脉证	Stasis Blocking the Blood Vessels
99	肺阴虚证	Lung Yin Deficiency
100	脾虚湿热证	Spleen Deficiency with Dampness and Heat
101	风湿痹阻证	Wind and Dampness Impediment and Obstruction
102	瘀血痹阻证	Impediment and Obstruction due to Blood Stasis
103	血热壅滞证	Blood Heat Congestion and Stagnation
104	气血瘀滞证	Qi and Blood Stasis and Stagnation
105	肝郁痰凝证	Phlegm Congealing due to Liver Depression
106	血热瘀滞证	Blood Stasis and Stagnation due to Heat
107	风热感冒证	Common Cold due to Wind and Heat
108	痰瘀阻络证	Phlegm and Stasis Obstructing the Collaterals
109	肝风内动证	Liver Wind Stirring Inside
110	脾肺气虚证	Qi Deficiency in Spleen and Lung
111	脾肾气虚证	Qi Deficiency in Spleen and Kidney
112	火毒入络证	Fire Toxin Entering Collaterals
113	风水相搏证	Wind and Edema Combat
114	风湿热痹证	Wind-Dampness with Heat Impediment
115	风痰阻络证	Wind and Phlegm Blocking Collaterals

Continued on next page

Table B1 – Continued from previous page

ID	Syndrome Name (ZH)	Syndrome Name (EN)
116	气滞痰阻证	Qi Stagnation with Phlegm Obstruction
117	风邪犯肺证	Pathogenic Wind Invading the Lung
118	肝火上炎证	Liver Fire Flaming Up
119	热证	Heat
120	风寒痹阻证	Wind and Cold Impediment and Obstruction
121	风阳上越证	Wind Yang Floating Upwards
122	肺燥阴虚证	Lung Dryness with Yin Deficiency
123	湿热浸淫证	Dampness and Heat Invasion and Spreading
124	亡阳证	Yang Prostration
125	肺胃热盛证	Exuberant Heat in the Lung and Stomach
126	风热犯表证	Wind and Heat Invading the Surface
127	痰瘀闭阻证	Phlegm and Stasis Obstruction
128	正虚邪恋证	Healthy Qi Deficiency and Pathogenic Factors Lingering
129	风寒袭肺证	Wind and Cold Attacking the Lung
130	瘀血内结证	Internal Binding of Blood Stasis
131	寒凝经络证	Cold Coagulation in Meridians and Collaterals
132	下焦湿热证	Dampness and Heat in the Lower Energizer
133	湿热壅滞证	Dampness and Heat Congestion and Stagnation
134	中气下陷证	Middle Qi Sinking
135	气滞湿阻证	Qi Stagnation and Dampness Obstruction
136	肾虚湿热证	Dampness and Heat due to Kidney Deficiency
137	痰热内扰证	Phlegm and Heat Harassing the Internal
138	气滞证	Qi Stagnation
139	热毒壅结证	Heat Toxin Congestion and Binding
140	湿毒蕴结证	Dampness Toxin Accumulation and Binding
141	阴虚肠燥证	Yin Deficiency with Intestinal Dryness
142	肝经湿热证	Dampness and Heat in the Liver Meridian
143	痰火扰神证	Phlegm and Fire Harassing the Spirit

Continued on next page

Table B1 – Continued from previous page

ID	Syndrome Name (ZH)	Syndrome Name (EN)
144	湿热蕴毒证	Dampness and Heat and Toxin Accumulation
145	气血痹阻证	Qi and Blood Impediment and Obstruction
146	营卫不和证	Nutrient Qi and Defense Qi Disharmony
147	痰浊阻肺证	Turbid Phlegm Obstructing the Lung
148	风燥伤肺证	Wind and Dryness Damaging the Lung
149	大肠湿热证	Dampness and Heat in the Large Intestine
150	肝胆火热证	Fire-Heat in the Liver and Gallbladder
151	痰热证	Phlegm and Heat
152	风水泛滥证	Wind Edema Flooding
153	肝经郁热证	Heat Stagnation In Liver Meridian
154	心血瘀阻证	Heart Blood Stasis Obstruction
155	痰饮证	Phlegm And Fluid
156	寒湿凝滞证	Cold And Dampness Coagulation And Stagnation
157	寒盛痹痛证	Impediment Marked By Pain Dominated By Cold
158	肺热阴虚证	Yin Deficiency Due To Lung Heat
159	血热风盛证	Blood Heat With Exuberant Wind
160	湿毒证	Dampness Toxin
161	湿热证	Dampness And Heat

Table B2: List of the included 51 syndrome element names in Chinese and English. The English translations are collected from GB/T 16751.2-2021 and *WHO international standard terminologies on traditional Chinese medicine*.

ID	Name	Name (EN)	ID	Name	Name (EN)
1	脾	Spleen	27	气逆	Qi Counterflow
2	肝	Liver	28	阳亢	Yang Hyperactivity
3	肾	Kidney	29	动风	Wind Stirring
4	火	Fire	30	食	Food
5	湿	Dampness	31	膀胱	Bladder
6	血瘀	Blood Stasis	32	血热	Blood Heat
7	胃	Stomach	33	出血	Hemorrhage
8	经络	Meridians and Collaterals	34	精髓亏	Essence and Marrow Deficiency
9	气虚	Qi Deficiency	35	胞宫	Uterus
10	阴虚	Yin Deficiency	36	表	External
11	气滞	Qi Stagnation	37	燥	Dryness
12	心	Heart	38	津液亏	Fluid and Humor Depletion
13	风	Wind	39	饮	Fluid
14	肺	Lung	40	气陷	Qi Sinking
15	痰	Phlegm	41	暑	Summer-Heat
16	毒	Toxin	42	亡阳	Yang Prostration
17	肌肤	Skin	43	半表半里	Neither Exterior nor Interior
18	寒	Cold	44	虫	Worm
19	血虚	Blood Deficiency	45	小肠	Small Intestine
20	心神	Heart Spirit	46	气闭	Qi Block
21	气不固	Qi Insecurity	47	气脱	Qi Collapse
22	阳虚	Yang Deficiency	48	血寒	Blood Cold
23	水	Water	49	精室	Essence Chamber
24	骨	Bone	50	血脱	Blood Prostration
25	胆	Gallbladder	51	亡阴	Yin Prostration
26	大肠	Large Intestine			

Appendix C Technical details of the BERT model

BERT is one of the most popular architectures in natural language processing (NLP) and achieved state-of-the-art performance in various tasks when it was first introduced

[47, 48]. The architecture of BERT consists of an embedding block, 12 transformer encoder blocks [49], and a linear layer.

Embedding block. For an input sentence (i.e. symptom description) of length L, the embedding block first adds a word [CLS] to the beginning of the sentence and a word [SEP] to the end, resulting in a new sentence of length $L+2$. It then assigns a unique 768-dimensional real-valued vector to each word in the sentence, producing a sequence of real-valued vectors with a shape of $(L+2)\times 768$, denoted as $\mathbf{X}_{Emb} \in \mathbb{R}^{(L+2)\times 768}$. Let $Emb(\cdot)$ denotes the embedding block, the aforementioned procedure can be described as:

$$X_{Emb} = Emb(\text{symptom description}). \tag{C1}$$

Transformer encoder block. In this block, X_{Emb} is passed through the 12 transformer encoder blocks sequentially, while keeping the dimension unchanged in this process. Specifically, for an input sequence $X_{in} \in \mathbb{R}^{(L+2)\times 768}$ and an output sequence $X_{out} \in \mathbb{R}^{(L+2)\times 768}$, a transformer encoder block, denoted by $TE(\cdot)$, can be represented as:

$$X_{out} = TE(X_{in}). \tag{C2}$$

Since the sequence lengths of X_{in} and X_{out} are the same, we can treat them as a pair of sequences, where the vector in the output is a transformed representation of the vector at the same position in the input. For example, the first vector in the output of each block can be considered as the representation of the first vector in the input, and ultimately, the representation of the word [CLS] in the input.

Linear layer. A linear layer with input dimension M and output dimension N can project an input vector $\mathbf{x}_{in} \in \mathbb{R}^M$ to an output vector $\mathbf{x}_{out} \in \mathbb{R}^N$. Let $LN(\cdot)$ denotes the linear layer, then a linear layer is defined as:

$$\mathbf{x}_{out} = LN(\mathbf{x}_{in}) = \mathbf{W}\mathbf{x}_{in} + \mathbf{b}, \tag{C3}$$

where $\mathbf{W} \in \mathbb{R}^{N \times M}$ and $\mathbf{b} \in \mathbb{R}^N$. Usually, there is an activation layer placed after the linear layer in deep learning. An activation layer simply applies a non-linear function to each element in the output of the linear layer.

In BERT, the first vector of the output of the transformer encoder block, i.e., the vector corresponding to the word [CLS], will be passed through a linear layer with both M and N equal to 768, and an activation layer with $\tanh(\cdot)$ being the activation function, to obtain the final pooled output. Let $\mathbf{x}^{PO} \in \mathbb{R}^{768}$ denote the pooled output, then the overall BERT model can be described as:

$$\mathbf{x}^{PO} = BERT(\text{symptom description}) \tag{C4}$$

$$= \tanh\left(LN\left(\underbrace{TE(\ldots TE}_{12}(Emb(\text{symptom description}))\right)\right). \tag{C5}$$

Appendix D Evaluation Metrics and Experiment Setup

D.1 Top-K accuracy

Top-K accuracy is a widely adopted metric in multi-class classification tasks. It is an intuitive evaluation metric that measures the proportion of sample points with the correct label among the top K predicted labels with the highest probabilities. Higher top-K accuracy indicates better classification performance. Given our dataset, which contains a large number of classes (161 syndromes) and each EHR corresponds to only one syndrome, top-K accuracy is particularly suitable [50, 51]. Mathematically, the top-K accuracy is defined as:

$$\text{top-K accuracy} = \frac{\text{\# EHRs with true label in top-K predicted labels}}{\text{sample size}}, \tag{D6}$$

where we set $K = 1, 3, 5, 10, 15$ in our experiments.

D.2 NDCG@K

NDCG is a common evaluation metric in information retrieval [52, 53]. In our experiments, it counts for the rank of the true label in the classification result. A higher NDCG implies a higher rank of the true label and a better model performance. Given the top-K labels with the highest classification probabilities, the definition of NDCG@K is

$$NDCG@K = \frac{DCG@K}{IDCG@K}, \tag{D7}$$

where

$$DCG@K = \sum_{i=1}^{K} \frac{I_i}{\log_2(i+1)}, \tag{D8}$$

where I_i is an indicator function that equals 1 if the ith label is the true label, and 0 otherwise, and IDCG represents the ideal DCG among all possible DCGs. In our case, $IDCG = 1$ because the best DCG is achieved when the model assigns the highest classification probability to the true label. The range of NDCG is $[0, 1]$. NDCG equals 1 if our model assigns the highest classification probability to the true label for all the records, and NDCG equals 0 if the true label is not within the top K classification probabilities for any record. Therefore, NDCG closer to 1 indicates a better classification performance. In our experiments, we choose $k = 3, 5, 10, 15$. We omit $k = 1$ because, in this case, the value of NDCG is equal to the value of top-1 accuracy.

D.3 More Discussion on the Evaluation Metrics

In this study, we employed Top-K accuracy and NDCG@K (K=1,3,5,10,15) to evaluate the performance of the proposed AIAD-A2 method. While Top-1 accuracy is a standard metric in machine learning, its clinical utility in this context is limited by the inherent complexity of clinical practice. The primary function of an AI diagnostic

tool, such as AIAD-A2, is not to replace the physician's final decision, but to augment their diagnostic process by generating a reliable list of potential syndromes for purposes of differential diagnosis. Given the challenge of distinguishing among 161 distinct syndromes, relying solely on Top-1 predictions carries a substantial risk of diagnostic omission. Instead, the core value of our system lies in its ability to narrow the diagnostic search space. By consistently capturing the true diagnosis within the Top-15 candidates (Top-15 accuracy $\approx 80\%$), the model effectively filters out over 90% of irrelevant conditions, transforming a complex 161-class classification problem into a manageable review of a focused candidate list. This high accuracy within the Top-K set ensures that the correct syndrome is presented for physician consideration, while favorable NDCG scores indicate that these diagnoses are prioritized to align with efficient clinical workflows.

While the ultimate goal of any clinical decision support system is to achieve perfect diagnostic precision (i.e., 100% Top-1 accuracy), the performance of AIAD-A2 remains highly competitive when benchmarked against established standards, particularly given the constraints in our dataset. With 6,226 samples distributed across 161 classes (approximately 38 samples per class), this task presents a significant few-shot learning challenge. Despite this, our results compare favorably to existing methods. Existing studies have shown that widely used online tools typically achieve a Top-1 accuracy of 34%, Top-3 of 51%, and Top-20 of 58% [54], while similar tools in Australia report Top-1 and Top-10 accuracies of 36% and 58%, respectively [55]. In contrast, our model's ability to achieve approximately 80% coverage within the Top-15 demonstrates promising prospects for clinical application. This suggests that, even though perfect Top-1 prediction remains the long-term objective, the proposed method currently provides valuable diagnostic support by effectively capturing the correct syndrome within a clinically manageable list.

Table E3: P-values for top-K accuracy and NDCG@K differences in syndrome differentiation performance between AIAD-A1, AIAD-A2, and Baseline.

H_0	P-value of top-K Accuracy				
	Top-1	Top-3	Top-5	Top-10	Top-15
AIAD-A1 $\leq$ Baseline	0.596	0.140	0.038	0.018	0.223
AIAD-A2 $\leq$ Baseline	0.002	0.001	<0.001	<0.001	<0.001
AIAD-A2 $\leq$ AIAD-A1	0.001	0.002	<0.001	<0.001	<0.001
H_0	P-value of NDCG@K				
	NDCG@1	NDCG@3	NDCG@5	NDCG@10	NDCG@15
AIAD-A1 $\leq$ Baseline	-	0.253	0.112	0.033	0.100
AIAD-A2 $\leq$ Baseline	-	0.001	<0.001	<0.001	<0.001
AIAD-A2 $\leq$ AIAD-A1	-	0.002	<0.001	<0.001	<0.001

D.4 Experimental Setup

The whole dataset is first evenly split into 5 subsets. Then in each fold, 1 subset is used for test and the remaining 4 subsets are used for model training. In the training phase, we randomly leave out 10% of the training sample as a validation set. The models are implemented using Python (ver. 3.8.13), Pytorch (ver. 2.1.2) and transformers (ver. 4.33.3). When tuning the BERT model, we run 10 epochs with batch size equals to 8. We adopt the AdamW optimizer [56] with a learning rate of 1e-3 and a linear scheduler.

Appendix E Paired T-test Results

Results for paired t-test is shown in Table E3.

Appendix F Verification of the Mediation Role of Syndrome Elements

This section validates the mediation role of syndrome elements. Before investigating the mediation mechanism in depth, we first validated the predictive performance of the proposed AIAD-A2 against single path baselines (the symptom only Baseline and the syndrome element only AIAD-A1), as depicted in Figure 4. Furthermore, to ensure the

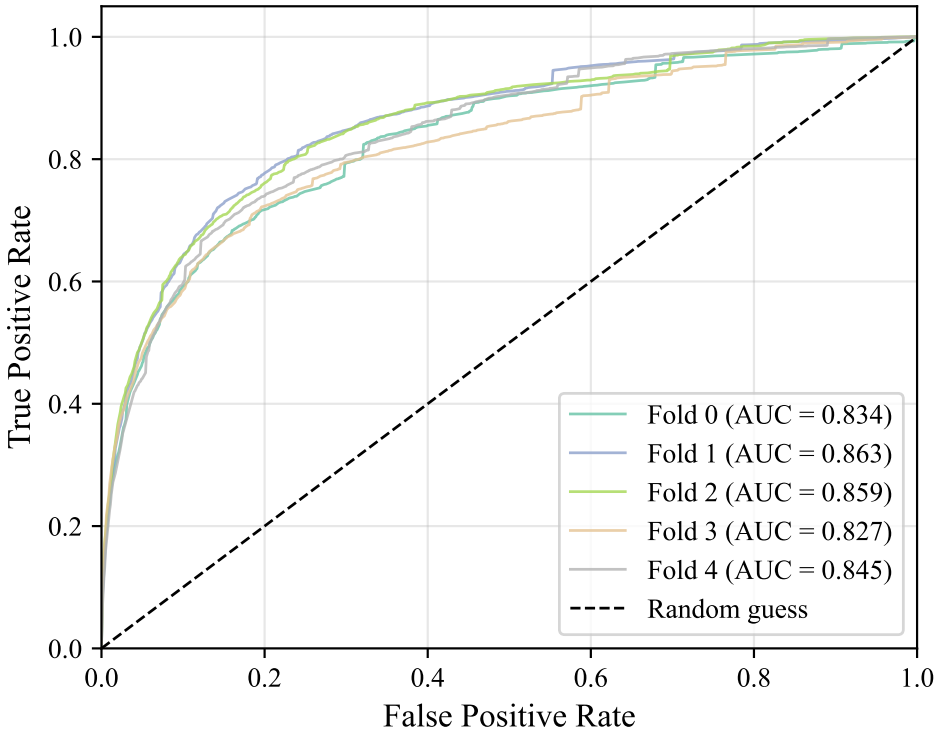


Fig. F3: The average ROC curves of all the syndrome elements for each fold.

learned syndrome elements capture meaningful pathological semantics, we evaluated the AIAD-A2's accuracy on the syndrome element diagnosis task. The model achieved a mean AUC of 0.85 ± 0.01 on *Symptom* $\rightarrow$ *Syndrome Element* (see Figure F3), confirming that the learned mediator representations are semantically valid and clinically aligned. While the superior performance of the proposed AIAD-A2 demonstrates its empirical effectiveness, it remains critical to verify the extent to which the model genuinely leverages the syndrome elements as a causal mediator. Traditional mediation analysis methods, such as the Sobel test [57], typically assume linear relationships and operate on scalar variables. These assumptions are incompatible with deep neural networks, which rely on complex non-linear interactions and high-dimensional latent

representations. To rigorously address this issue, we employ causal mediation analysis [58, 59] tool in causal inference utilizing the interchange interventions technique [60, 61].

F.1 Causal Mediation Analysis

To quantify the mediation role of the syndrome element, we define the causal effect of symptoms (X) on the syndrome differentiation (Y) through two pathways: the direct path ($X \to Y$) and the indirect path ($X \to$ syndrome element (M) $\to Y$). We estimate two key quantities, the Natural Indirect Effect (NIE) and the Natural Direct Effect (NDE), to measure the causal mediation effect.

NIE measures the contribution of the indirect path. It is estimated by observing the change in model output when the input X is held fixed, but the intermediate representation M is intervened upon by replacing it with a counterfactual representation M' derived from a different sample. **NDE** measures the contribution of the direct path. It is estimated by intervening on the input X (replacing it with X'), while forcing the intermediate representation M to remain consistent with the original input X. If both NIE and NDE are significant, we can verify the mediating role of syndrome elements.

Due to the multi-class nature of Y in our problem, direct subtraction using discrete model outputs (labels) is ill-defined. Therefore, we measure the NIE and NDE using the differences in model performance. The operational formulas are given by:

$$NIE = \mathcal{P}(X, M) - \mathcal{P}(X, M') \quad \text{and} \quad NDE = \mathcal{P}(X, M) - \mathcal{P}(X', M), \tag{F9}$$

where $\mathcal{P}$ denotes the performance metric on the test set. We implement the NIE and NDE interventions using the interchange interventions technique, where X' and M' are drawn from randomly shuffled samples in the test set.

F.2 Experimental Results

We performed intervention experiments on the test set. The performance changes under causal interventions are summarized in Table F4. We have the following observations:

- **Baseline Performance**: AIAD-A2 significantly outperforms both the symptom only model Baseline and the syndrome element or mediator only model AIAD-A1, indicating a synergistic effect.
- **NIE Intervention**: When the correct syndrome element representation M is replaced with M' (cutting the indirect path), AIAD-A2's performance drops but stabilizes at a level comparable to the baseline.
- **NDE Intervention**: When the correct symptom input X is replaced with X' while retaining the correct syndrome element representation M (cutting the direct path), the performance drops dramatically, falling significantly below the level of the mediator only AIAD-A1.

These results confirm a partial mediation mechanism where the direct path ($X \rightarrow Y$) acts as a foundational baseline, ensuring robust diagnostic capability even when the mediator path is disrupted. Complementing this, the indirect path ($X \rightarrow M \rightarrow Y$) functions as a critical semantic booster, enabling the model to surpass the baseline by leveraging high-level pathological insights. Taken together, these findings provide robust empirical evidence validating the syndrome elements as significant mediators in the diagnostic process of syndrome differentiation.

Appendix G Technical Details of the Saliency Map

To elucidate the diagnostic rationale of the proposed AIAD-A2 model, we utilize saliency maps, a technique implemented in state-of-the-art interpretability frameworks such as Google's Learning Interpretability Tool [62], for local explanation. Saliency

Table F4: Causal mediation analysis results using top-K accuracy (%) as the performance measurement.

Model	Top 1	Top 3	Top 5	Top 10	Top 15
Baseline	25.73	42.89	52.15	65.19	73.30
AIAD-A1	25.65	43.09	52.95	66.18	73.76
AIAD-A2	29.32	48.57	58.29	71.50	79.24
NIE Intervened	25.95	43.56	52.68	65.94	73.90
NIE	3.38	5.01	5.61	5.57	5.35
P-value	<0.01	<0.01	<0.01	<0.01	<0.01
NDE Intervened	3.87	11.40	17.51	29.30	38.49
NDE	25.45	37.17	40.78	42.20	40.75
P-value	<0.01	<0.01	<0.01	<0.01	<0.01

Note: 'Intervened' indicates model performance under the intervention. NIE and NDE are differences in performance between the corresponding intervened model and AIAD-A2 according to Equation (F9).

maps visualize the influence of different parts of the input features, which refer to individual tokens (words) in the symptom description in our case, on the model's prediction for the primary syndrome.

The underlying methodology employs gradient-based token-wise salience to quantify how much each word contributes to the final classification [63]. In our study, for an input symptom sequence, let S_y be the model's predicted score (logit or probability) for the target syndrome y. The salience M_i of the i-th token is calculated as the magnitude of the gradient of S_y with respect to the token's input embedding e_i:

$$M_i = ||\frac{\partial S_y}{\partial e_i}||, \tag{G10}$$

where $||\cdot||$ refers to the norm operation.

In our visualization, the background of each word is colored based on its calculated salience. A higher color intensity represents a larger gradient, indicating that the model's prediction is highly sensitive to that specific word. Intuitively, words with high salience are those whose presence significantly drives the model's attention. In other words, their removal would lead to a substantial drop in the probability of the predicted syndrome. This approach allows us to perform an interactive inspection

of the diagnostic logic, verifying whether the model correctly focuses on key TCM symptoms, such as "stasis spots" on the tongue in our example, to reach its conclusion.

Appendix H Additional Results

H.1 Targeting a smaller group of syndromes

We have demonstrated the performance of our proposed method in the scenario of complex multitudinous syndrome differentiation. Therefore, the number of syndromes in this paper is much larger compared to the existing works on syndrome differentiation. The syndrome differentiation targeting a smaller subset of syndromes is another typical scenario in clinical practice. To test our proposed methods in this scenario, we truncated the test sets in both the NSTSPC and the ITD datasets to retain the top 20 and 40 syndromes with the largest number of records. We evaluated our models on these truncated test sets.

The results of top-K accuracy are given in Table H5. Both the AIAD-A1 and AIAD-A2 methods demonstrate improved performance over the baseline model on the two datasets. Taking the top-5 accuracy as an example, on the NSTSPC dataset, the baseline achieves an accuracy of 82.58% on the top 20 syndromes and 73.10% on the top 40 syndromes. AIAD-A1 achieves an accuracy of 83.99% on the top 20 syndromes and 74.25% on the top 40 syndromes. AIAD-A2 achieves an accuracy of 84.38% on the top 20 syndromes and 76.09% on the top 40 syndromes. On the ITD dataset, the baseline achieves an accuracy of 76.88% on the top 20 syndromes and 55.55% on the top 40 syndromes. AIAD-A1 achieves an accuracy of 87.19% on the top 20 syndromes and 67.42% on the top 40 syndromes. AIAD-A2 achieves an accuracy of 87.66% on the top 20 syndromes and 76.52% on the top 40 syndromes.

These results indicate that when targeting a smaller subset of common syndromes, our proposed methods, especially AIAD-A2, achieve high accuracy.

Table H5: Top-K accuracy (%) on truncated test samples for the first 20 syndromes and the first 40 syndromes with the largest sample sizes.

Model	Top-1	Top-3	Top-5	Top-10	Top-15
	First 20 Syndromes				
Baseline	50.15	73.08	82.58	95.12	98.88
AIAD-A1	49.93	73.83	83.99	95.09	99.13
AIAD-A2	51.70	75.38	84.38	95.85	99.13
Baseline (ITD)	29.30	63.52	76.88	90.31	97.03
AIAD-A1 (ITD)	43.12	78.44	87.19	94.22	98.83
AIAD-A2 (ITD)	52.58	79.69	87.66	95.31	98.59
	First 40 Syndromes				
Baseline	38.84	62.18	73.10	86.39	92.69
AIAD-A1	38.57	62.67	74.25	87.65	93.51
AIAD-A2	42.33	65.68	76.09	88.24	93.83
Baseline (ITD)	17.34	40.00	55.55	73.48	86.29
AIAD-A1 (ITD)	24.41	52.58	67.42	85.74	93.79
AIAD-A2 (ITD)	29.80	57.50	76.52	90.39	95.35

Note: '(ITD)' indicates that the model was tested on the ITD dataset.

H.2 Sensitivity analyses

In this section, we conduct three sensitivity analyses for the AIAD models. First, we test whether our results are sensitive to the feature extraction NLP model. We replace the BERT model with various state-of-the-art NLP models, including RoBERTa [64], ALBERT [65], and MacBERT [66]. At the same time, we also compare with other state-of-the-art deep learning or transformer-based models for TCM diagnosis, including **TCM-Bert** [67], **BERT-BiL-Att** [68] and **TCM-BERT-CNN** [69], to further contextualize performance gains. The second sensitivity analysis experiments on two state-of-the-art LLMs, GPT-4o [70] and Qwen2.5-72B [71], to test whether these general-purpose LLMs can outperform the developed AIAD models. The last sensitivity analysis evaluates the sensitivity of our results to the setting of loss function. Since the main target here is syndrome differentiation, it may not be fully adequate to assign equal weights for the loss functions of syndrome differentiation and syndrome element diagnosis. Therefore, we introduce a hyperparameter α here to the loss

function defined in Equation (7) as follows:

$$\mathcal{L}_{MTL} = \mathcal{L}_s + \alpha \mathcal{L}_e, \tag{H11}$$

where α controls the contribution of syndrome element diagnosis to the overall loss function.

The results for the sensitivity analyses are summarized in Table H6. We have the following findings. First, with different NLP models as the feature extraction module, the AIAD-A2 method consistently outperforms the baseline model and AIAD-A1. These results confirm the robustness of the developed AIAD-A2 method and the mediating role of syndrome elements in syndrome differentiation. Second, the two LLMs exhibit certain levels of capability in syndrome differentiation under the zero-shot setting, with GPT-4o achieving a Top-15 accuracy of 47.60% and Qwen2.5-72B achieving 52.03%. However, our AIAD-A2 method substantially outperforms both LLMs, underscoring the importance of fine-tuning with domain-specific data for syndrome differentiation. Third, as observed, while TCM-Bert achieves results similar to the vanilla Bert, BERT-BiL-Att and TCM-BERT-CNN outperform the vanilla BERT baseline by benefiting from domain adaptation or complex feature extractors. However, they still underperform our proposed AIAD-A2. This context highlights the specific advantage of our approach in that while the compared state-of-the-art models primarily focus on enhancing textual representation, AIAD-A2 goes a step further by explicitly modeling the intrinsic diagnostic logic via syndrome elements as mediators. This demonstrates that integrating medical knowledge driven structural priors is more effective than pure representation learning for this complex task. Finally, our AIAD-A2 are relatively robust to the choice of α and an appropriate selection of α can further enhance model performance. For instance, when α is reduced from 1 to 0.5, the Top-15 accuracy of our AIAD-A2 method (with BERT as the NLP model) improves from 79.24% to 79.73%.

Table H6: Top-K accuracy (%) for different base NLP models and different settings of α. "Baseline" denotes only using base NLP model for syndrome differentiation. "A1" denotes the intermediate implementation and "A2" denotes the mediator implementation. The best result for each base model is bolded, and the second best result for each base model is underlined.

Base model	Type	α	Top1	Top3	Top5	Top10	Top15
BERT	Baseline	–	25.73	42.89	52.15	65.19	73.30
	TCM-Bert	–	25.26	42.63	52.08	64.92	73.32
	BERT-BiL-Att	–	26.73	43.39	53.56	67.47	75.61
	TCM-BERT-CNN	–	27.16	45.37	54.73	67.26	74.28
	A1	1	25.65	43.09	52.95	66.18	73.76
		0.5	25.58	44.08	53.53	66.06	73.95
		0.1	23.70	40.11	49.25	62.10	70.44
		0.05	20.92	36.00	45.15	58.32	66.19
		0.01	10.95	20.27	25.86	37.09	45.25
	A2	1	<u>29.32</u>	<u>48.57</u>	<u>58.29</u>	<u>71.50</u>	<u>79.24</u>
		0.5	**29.73**	**49.60**	**59.06**	**72.74**	**79.73**
		0.1	28.75	47.05	56.88	70.30	77.91
		0.05	27.53	45.75	54.95	68.66	76.85
		0.01	26.20	43.38	52.79	67.06	74.74
RoBERTa	Baseline	–	21.88	35.86	45.64	58.92	67.21
	A1	1	26.18	44.34	53.78	66.22	73.70
		0.5	25.86	44.07	54.18	65.72	73.56
		0.1	19.99	33.89	41.87	54.64	62.01
		0.05	13.50	25.16	32.72	46.30	54.31
		0.01	3.44	10.80	16.43	27.96	37.00
	A2	1	<u>29.75</u>	<u>49.11</u>	<u>59.06</u>	<u>72.50</u>	<u>79.96</u>
		0.5	**29.88**	**49.77**	**60.03**	**73.39**	**80.69**
		0.1	26.87	45.65	55.05	68.03	75.67
		0.05	24.18	41.04	51.02	64.20	72.15
		0.01	22.24	37.39	46.57	60.64	68.77
ALBERT	Baseline	–	19.19	32.95	41.32	54.74	63.20
	A1	1	7.63	16.10	22.94	35.09	43.79
		0.5	14.26	27.21	33.96	46.82	54.92
		0.1	10.12	22.19	29.23	41.74	50.63
		0.05	17.90	30.59	39.24	52.60	60.61
		0.01	12.05	21.43	28.88	40.37	49.47
	A2	1	8.59	17.92	24.43	36.91	45.89
		0.5	18.62	33.76	42.46	55.80	63.99
		0.1	**27.59**	**46.90**	**56.50**	**69.52**	**77.32**
		0.05	18.49	32.12	39.75	52.54	60.16
		0.01	<u>21.84</u>	<u>37.93</u>	<u>46.06</u>	<u>58.65</u>	<u>66.41</u>
MacBERT	Baseline	–	24.64	40.88	50.28	64.10	72.13
	A1	1	25.81	44.00	54.01	66.74	75.06
		0.5	25.44	43.37	53.42	66.28	74.15
		0.1	23.33	39.64	49.41	62.18	70.07
		0.05	19.34	33.72	43.33	57.33	64.62
		0.01	7.59	16.59	22.11	32.66	41.12
	A2	1	<u>29.61</u>	<u>49.20</u>	<u>59.03</u>	<u>72.83</u>	<u>79.79</u>
		0.5	**30.17**	**49.41**	**59.18**	**72.95**	**80.04**
		0.1	28.45	47.27	56.57	70.78	78.33
		0.05	26.89	45.27	55.19	68.52	76.42
		0.01	26.15	42.69	52.58	66.52	74.19
LLM	GPT-4o	–	13.73	27.44	34.66	44.20	47.60
	Qwen2.5-72B	–	15.39	29.49	37.47	48.27	52.03

Table I7: List of Abbreviations.

ID	Abbreviation	Full Name
1	AI	Artificial Intelligence
2	AIAD	Artificial Intelligence Aided Diagnosis
3	BERT	Bidirectional Encoder Representations from Transformers
4	CACM	China Association of Chinese Medicine
5	CAD	Computer Aided Diagnosis
6	EHR	Electronic Health Record
7	ITD	Independent Test Dataset
8	LLM	Large Language Model
9	MTL	Multi-task Learning
10	NLP	Natural Language Process
11	NSTSPC	National Science and Technology Support Program of China
12	TCM	Traditional Chinese Medicine
13	WM	Western Medicine

Appendix I Abbreviations

See Table I7 for list of abbreviations used in this work.

References for Appendix

[41] Li Z, Zeng L, Liang Z, Zhang Y, Chen Q, Fu W. The development of Chen's flying needle based on the traditional Chinese Medicine: A needle-inserting technique created by a distinguished veteran doctor of TCM. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops. IEEE; 2012. p. 444–447.

[42] Liu Z, Chen S, Cai J, Zhang E, Lan L, Zheng J, et al. Traditional Chinese medicine syndrome-related herbal prescriptions in treatment of malignant tumors. Journal of Traditional Chinese Medicine. 2013;33(1):19–26.

[43] Wu J, Guo W, Tang Y, Han A, Yang B, Zhang D, et al. A study of TCM master Yan Zhenghua's medication rule in prescriptions for digestive system diseases based on Apriori and complex system entropy cluster. Journal of Traditional Chinese Medical Sciences. 2015;2(4):241–247.

[44] Yang F, Jin H, Qi H. Reasoning research on vague information based on case-based reasoning and fuzzy-based reasoning in traditional Chinese medicine diagnosis. In: Proceedings of 2012 2nd International Conference on Computer Science and Network Technology. IEEE; 2012. p. 1813–1817.

[45] Zhao Fy, Zhang Zy, Xia Xj, Hong YF, Zhao Yx, Yan Hx, et al. Memory response to manual acupuncture in chronic insomniacs: evidence from event-related potentials. Journal of Acupuncture and Tuina Science. 2018;16(6):382–388.

[46] Tang Y, Li Z, Yang D, Fang Y, Gao S, Liang S, et al. Research of insomnia on traditional Chinese medicine diagnosis and treatment based on machine learning. Chinese Medicine. 2021;16:1–21.

[47] Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics. 2021;8:842–866.

[48] Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems. 2023;3:121–154.

[49] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[50] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012;25.

[51] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020;.

[52] Seltzer ML, Droppo J. Multi-task learning in deep neural networks for improved phoneme recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2013. p. 6965–6969.

[53] Järvelin K, Kekäläinen J. IR evaluation methods for retrieving highly relevant documents. In: ACM SIGIR Forum. vol. 51. ACM New York, NY, USA; 2017. p. 243–250.

[54] Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. bmj. 2015;351.

[55] Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. Medical Journal of Australia. 2020;212(11):514–519.

[56] Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:171105101. 2017;.

[57] Sobel ME. Asymptotic confidence intervals for indirect effects in structural equation models. Sociological methodology. 1982;13:290–312.

[58] Pearl J. Direct and indirect effects. In: Probabilistic and causal inference: the works of Judea Pearl; 2022. p. 373–392.

[59] Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychological methods. 2010;15(4):309.

[60] Vig J, Gehrmann S, Belinkov Y, Qian S, Nevo D, Singer Y, et al. Investigating gender bias in language models using causal mediation analysis. Advances in neural information processing systems. 2020;33:12388–12401.

[61] Geiger A, Lu H, Icard T, Potts C. Causal abstractions of neural networks. Advances in Neural Information Processing Systems. 2021;34:9574–9586.

[62] Tenney I, Wexler J, Bastings J, Bolukbasi T, Coenen A, Gehrmann S, et al. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. arXiv preprint arXiv:200805122. 2020;.

[63] Li J, Chen X, Hovy E, Jurafsky D. Visualizing and understanding neural models in NLP. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016. p. 681–691.

[64] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019;.

[65] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:190911942. 2019;.

[66] Cui Y, Che W, Liu T, Qin B, Wang S, Hu G. Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:200413922. 2020;.

[67] Yao L, Jin Z, Mao C, Zhang Y, Luo Y. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. Journal of the American Medical Informatics Association. 2019;26(12):1632–1636.

[68] Hui Y, Du L, Lin S, Qu Y, Cao D. Extraction and classification of tcm medical records based on bert and bi-lstm with attention mechanism. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2020. p. 1626–1631.

[69] Chen Z, Zhang D, Liu C, Wang H, Jin X, Yang F, et al. Traditional Chinese medicine diagnostic prediction model for holistic syndrome differentiation based on deep learning. Integrative Medicine Research. 2024;13(1):101019.

[70] Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, et al. Gpt-4o system card. arXiv preprint arXiv:241021276. 2024;.

[71] Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, et al. Qwen2 Technical Report. arXiv preprint arXiv:240710671. 2024;.