scientific data

Check for updates

OPEN

DATA DESCRIPTOR

TCMEval-SDT: a benchmark dataset for syndrome differentiation thought of traditional Chinese medicine

Zhe Wang[1,2,5], Meng Hao[3,5], Suyuan Peng[3], Yuyan Huang[4], Yiwei Lu[3], Keyu Yao[3], Xiaolin Yang[1]✉ & Yan Zhu[3]✉

This paper presents a large publicly available benchmark dataset (TCMEval-SDT) for the thought process involved in syndrome differentiation in traditional Chinese medicine (TCM). The dataset consists of 300 TCM syndrome diagnosis cases sourced from the internet, classical Chinese medical texts, and medical records from hospitals, with metadata adhering to the Findable, Accessible, Interoperable, and Reusable (FAIR) principles. Each case has been annotated and curated by TCM experts and includes medical record ID, clinical data, explanatory summary, TCM syndrome, clinical information, and TCM pathogenesis, to support algorithms or models in emulating the diagnostic process of TCM clinicians. To provide a comprehensive description of the TCM syndrome diagnosis process, we summarize the diagnosis into four steps: (1) clinical information extraction, (2) TCM pathogenesis reasoning, (3) TCM syndrome reasoning, and (4) explanatory summary. We have also established validation criteria to evaluate their ability in TCM clinical diagnosis using this dataset. To facilitate research and evaluation in syndrome diagnosis of TCM, the TCMEval-SDT dataset is made publicly available under the CC-BY 4.0 license.

Background & Summary

Traditional Chinese Medicine (TCM) plays a significant role in the treatment and prevention of diseases and is an important part of the world's traditional medicine[1,2]. For example, artemisinin's effectively treat polycystic ovarian syndrome (PCOS) by mediating the LONP1-CYP11A1 interaction, leading to decreased androgen synthesis[3]. An herbal-based injection has been demonstrated to be effective in reducing 28-day mortality in patients with sepsis[4]. Bianzheng Lunzhi (Syndrome Differentiation and Treatment) is a core component of the theoretical framework of TCM. This personalized diagnostic and therapeutic approach involves a comprehensive analysis of various factors, including the patient's specific disease, constitution, and environmental conditions, to determine the most appropriate treatment plan. Bianzheng Lunzhi represents the fundamental strategy and methodology of clinical practice in TCM[5,6].

Over the last few decades, artificial intelligence (AI) has seen rapid advancements in diverse industries. AI is increasingly demonstrating its potential in the medical field, with AI algorithms and models achieving significant results in disease diagnosis, drug discovery, patient car[7]. To objectively evaluate the performance of these AI algorithms and models, several benchmark datasets are currently being used. For example, DigestPath[8] is utilized to assess gastrointestinal pathology detection algorithms, and MultiMedQA serves as the benchmark for evaluating medical questions[9].

The diagnostic procedure in TCM clinical practice is different from that of Western medicine in that it diagnoses not only disease but also syndrome. The process of diagnosing a disease contains medical history

[1]Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences; School of Basic Medicine, Peking Union Medical College, Beijing, 100005, China. [2]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, 04103, Germany. [3]Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing, 100700, China. [4]Institute of Basic Theory for Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing, 100700, China. [5]These authors contributed equally: Zhe Wang, Meng Hao. ✉e-mail: yangxl@pumc.edu.cn; zhuyan166@126.com

Item	Sub-items	Score
Patient information	Demographic information: gender, age, etc.	0-1
	Main symptoms: brief and highlight the main symptoms	0-1
	Medical history: past medical history, family genetic history, and present medical history	0-1
	Previous interventions related to the disease	0-1
Clinical findings	Physical examination	0-1
	The four diagnostic methods in traditional Chinese medicine.	0-1
Timeline	Depict important dates and times related to the onset and progression in the case.	0-1
Diagnostic Evaluation	Diagnosis basis	0-1
	Diagnostic reasoning	0-1
	Diagnostic conclusion	0-1

Table 1. Details of TCM Medical Record Quality Assessment Scale.

collection, physical examination, medication use and laboratory tests. However, for diagnosing syndrome, there are no specialized benchmark for the process of syndrome differentiation. Existing benchmark datasets mostly focus on answering basic TCM knowledge questions, such as TCM Bench[10], or on evaluating syndromes derived from case analysis, such as TCM-SD[11]. However, these benchmark datasets do not cover the reasoning process of TCM syndrome diagnosis.

To address the above problems, this study first summarizes the TCM syndrome diagnosis into four steps. (1) clinical information extraction; (2) pathogenesis reasoning; (3) syndrome reasoning; and (4) explanatory summary. Based on this framework, we annotated and curated the TCM medical records. To this end, we have developed TCMEval-SDT, a benchmark dataset specifically designed to evaluate the ability of algorithms or models in TCM clinical diagnosis through syndrome differentiation. Our study aims to advance the development of algorithms or models capable of syndrome differentiation thinking in TCM, such as enabling large language models (LLMs) to think or reason like TCM clinicians during syndrome differentiation using Chain-of-Thought (CoT)[12] based on TCMEval-SDT. Ultimately achieve automated diagnosis in the field of TCM. This study has three main objectives:

1. To present a large TCM syndrome diagnosis dataset with the metadata that comply with Findable, Accessible, Interoperable, and Reusable (FAIR) principles[13]. For example, medical record ID (DE0087751), clinical Data (DE0087752), clinical information (DE0087755), TCM pathogenesis (DE0087756), TCM syndrome (DE0087757) and explanatory summary (DE0087753).
2. To establish evaluation metrics and allow users to evaluate their answers for performance assessment.
3. To invite users to submit new data to collaboratively build and reuse a benchmark dataset for syndrome diagnosis in TCM, aiming to improve the reusability of data and the overall quality of TCM assessment datasets.

Methods

In this study, the medical records were processed by TCM-Experts, ensuring that all medical records underwent anonymization. A rigorous quality assurance process was implemented to ensure the privacy, accuracy, and reliability of the collected medical records. Subsequently, 300 medical records were selected through manual screening. These records were annotated using Baibu Knowledge Engine[14,15], a corpus Tool in the field of TCM that supports automatic annotation, human-machine combined annotation, and manual annotation modes for entity and relation annotation, to construct a comprehensive and systematically organised dataset for TCM syndrome diagnosis.

Data collection. The medical records were sourced from a self-built database established by our team, curated by experts from the Institute of Information on Traditional Chinese Medicine-China Academy of Chinese Medical Sciences, the Institute of Basic Theory for Chinese Medicine-China Academy of Chinese Medical Sciences, and senior TCM students. The data were collected from diverse sources, such as the China National Knowledge Infrastructure (CNKI, https://www.cnki.net), Wanfang data (https://www.wanfangdata.com.cn), classical Chinese medical texts and medical records from hospitals.

The data were first screened by TCM experts according to the following standards: (1) Complete medical record, including information such as clinical data and clinical experience, etc.; (2) Cases of common diseases. Cases of rare diseases and duplicate cases were excluded. To evaluate the quality of TCM medical records, we developed a TCM Medical Record Quality Assessment Scale (as shown in Table 1) based on the CARE guidelines and TCM expert opinions. This scale comprises ten sub-items, including patient information, clinical findings, timeline, and diagnostic evaluation, to systematically assess the quality of TCM case data. Evaluation results are categorized as "clearly described" "not clearly described" and " not described" with corresponding scores of 1, 0.5, and 0, respectively[16,17]. The TCM expert group assessed the quality of the manually screened cases using this scale, excluding cases with scores lower than 6 and including those with scores of 6 or higher.

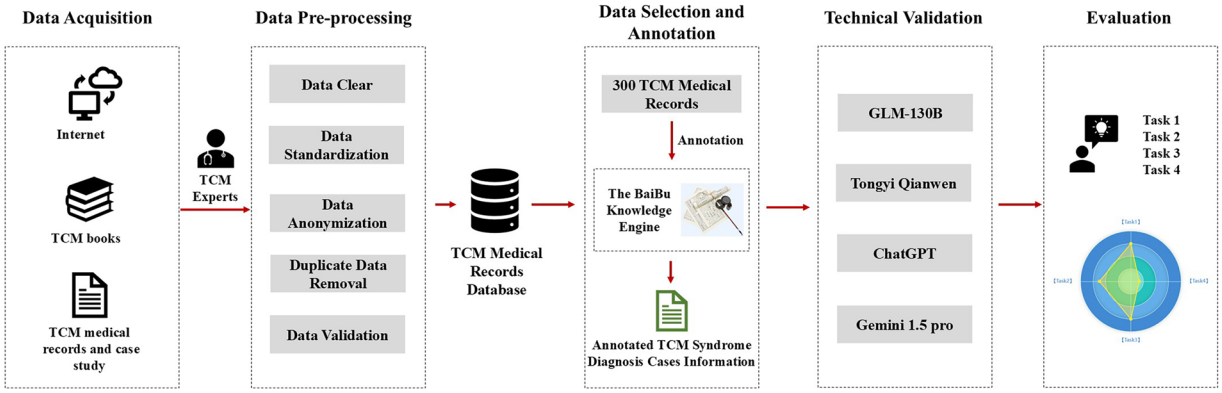

Fig. 1 Overview of the data processing workflow and evaluation for the TCMEval-SDT benchmark dataset. The TCM syndrome diagnosis cases sourced from the internet, classical Chinese medical texts, and hospital medical records. The original medical records underwent data preprocessing, including data cleaning, anonymization, and the removal of duplicates, before being stored in a database. From this database, 300 cases meeting specific criteria, such as non-rare cases, were selected. These cases were then annotated and curated by TCM experts using the Baibu knowledge engine. Finally, validation was performed using publicly available LLMs, including GLM-130B, Tongyi Qianwen, ChatGPT, and Gemini 1.5 Pro. ***Note***. TCM = traditional Chinese medical; LLMs = large language models.

Metadata Name	Metadata ID/ID	Metadata URL	Description
Medical Record ID	DE0087751	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087751	Unique identifier for each medical report.
Clinical Data	DE0087752	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087752	Contains basic patient information, including chief complaint, medical history, physical examination findings, and diagnostic test results.
Explanatory Summary	DE0087753	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087753	A summary provided by the TCM clinician, reflecting their interpretation of the patient's diagnosis, treatment approach, and observed changes in condition.
Syndrome Differentiation	DE0087754	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087754	Based on TCM theory, the syndrome is determined through a comprehensive analysis of information gathered using the four diagnostic methods (inspection, listening/smelling, inquiry, and palpation).

Table 2. FAIR-compliant metadata of medical record in TCMEval-SDT dataset.

Data pre-processing and anonymization. The preprocessing workflow for the medical records is shown in Fig. 1. The first step involves anonymizing each medical record by permanently removing identifiable information, such as patient ID and name, to protect patient privacy. The second step entails cleaning and organizing the data by removing duplicate or null data and standardizing the medical records. The FAIR principles serve as foundational guidelines for data sharing and reuse. To support these goals, we designed metadata for medical records in our study that comply with the FAIR principles. We shared the metadata of the TCMEval-SDT dataset on the CDE Portal (https://cdeportal.bmicc.cn), a public metadata registration and management platform, to facilitate the design and management of metadata for similar future projects (as shown in Table 2). We organized unstructured data, including TXT, PDF, Word, and HTML files, into structured data according to metadata requirements, and then assigned a unique identifier to each medical record. Finally, we constructed a benchmark database for syndrome diagnosis, named TCMEval-SDT.

Data selection and annotation. The diagnosis of syndromes in TCM is inherently multidimensional, involving a comprehensive evaluation of the interactions between a patient's physiological, pathological, and environmental factors. For theoretical analysis and practical guidance, we have summarized the TCM syndrome diagnosis process into four steps, as illustrated in Fig. 2.

(1) **Clinical Information Extraction**: emulating TCM clinicians in obtaining clinical information from the patient's medical data.
(2) **Pathogenesis Reasoning**: Inferring TCM pathogenesis from relevant clinical information.
(3) **Syndrome Reasoning**: Inferring TCM syndromes from relevant TCM pathogenesis.
(4) **Explanatory Summary**: Summarizing clinical experiences and insights from TCM clinicians.

Entity and relation for medical record. We selected 300 medical records and employed the Baibu Knowledge Engine to annotate them according to the aforementioned steps. The annotated entities and their relations are shown in Tables 3, 4.

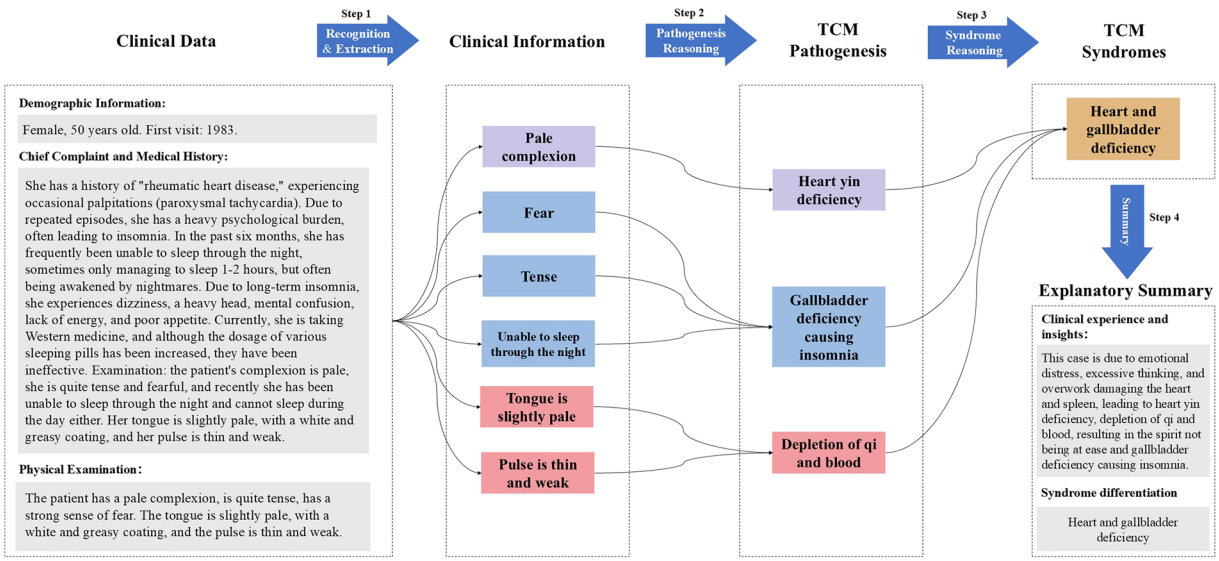

Fig. 2 Key steps for syndrome diagnosis of TCM. The figure illustrates the four key steps in TCM syndrome diagnosis. On the left side, the processed clinical data is shown, including the patient's demographic information, chief complaint, medical history, and physical examination. First step, through recognition and extraction, the patient's clinical information is obtained. Based on this clinical information, the corresponding TCM pathogenesis is inferred. Then, the TCM pathogenesis is used to inferred the relevant TCM syndromes. Finally, an explanatory summary is provided, emulating the process TCM clinicians follow for syndrome diagnosis. ***Note***. TCM = traditional Chinese medical.

Annotation guidelines.

(1) We classified the clinical information into two types: relevant information and irrelevant information. Relevant information refers to critical clinical information that significantly influences the diagnostic process, while irrelevant information refers to clinical information that does not impact the diagnosis. The annotated entities include only the relevant information in the TCM syndrome diagnosis process. For example, belching (clinical information) – stomach qi upward (TCM pathogenesis) – liver and stomach disharmony (syndrome). Irrelevant information, such as "red tongue with white coating" is excluded from the annotation scope as it does not directly influence this diagnostic process.
(2) It is essential that the annotated entities must be as comprehensive as possible. For example, in "painful distension behind the sternum and in the epigastric region", the entire phrase must be annotated to prevent loss of critical information by annotating only "painful".
(3) Inferential relationships exist between clinical information and TCM pathogenesis, and also between TCM pathogenesis and TCM syndromes. For example, extracting clinical information such as "belching" and "depressed state" leads to the inference of TCM pathogenesis, including "stomach qi upward" and "liver-qi stagnation". Integrating these pathogenic indicators results in the identification of TCM syndromes like "liver and stomach disharmony".
(4) In this study, the annotation task adheres to a specific rule for long mentions where multiple entities are connected: each entity with independent significance is annotated separately. For example, in the phrase "painful distension behind the sternum and in the epigastric region, burning sensation behind the sternum, sensation of obstruction when swallowing, accompanied by belching and nausea", the annotation was conducted as follows: "painful distension behind the sternum and in the epigastric region", "burning sensation behind the sternum", "sensation of obstruction when swallowing" accompanied by "belching" and "nausea". This approach ensures that each meaningful entity is properly annotated based on its individual significance.

Example of clinical records annotation through the Baibu Knowledge Engine. Figure 3 illustrates an example of a TCM record annotated using the Baibu Knowledge Engine. TCM experts annotate the clinical Information, TCM pathogenesis, TCM syndrome, and its relations.

Example of the thought process design in syndrome differentiation. Figure 4 illustrates the detailed design of the thought process in syndrome differentiation. TCM experts extract clinical information and infer TCM pathogenesis based on the clinical data. The inferred pathogenesis is then used to deduce the corresponding syndromes. This process emulates the specific reasoning steps employed by TCM clinicians during syndrome differentiation, providing AI algorithms and models with detailed steps to emulate this reasoning process.

Entity Type	Definition	Example
Clinical Information	During the syndrome diagnosis process in TCM, the TCM clinicians extract clinical information from the patient's clinical data, including physical signs and medical history, etc.	Pale complexion, Fear
TCM Pathogenesis	In TCM clinical diagnosis, the mechanism of the onset, development and progression of each disease is described in detail, including the nature of the disease, the location of the disease, the situation of the disease, the changes in the qi and blood of the internal organs, and its prognosis. In the process of identification and diagnosis, the TCM practitioner infers the nature of the disease and key pathological changes through step-by-step reasoning based on the patient's clinical information.	Heart yin deficiency, Depletion of qi and blood
TCM Syndrome	The name of the disease as classified within the TCM syndrome classification system.	Heart and gallbladder deficiency

Table 3. Annotated entity of medical records.

Relation Type	Relation Name	Definition
Clinical Information -TCM Pathogenesis	TCM Pathogenesis reasoning	Gradually inferring the pathogenesis from the patient's clinical information.
TCM Pathogenesis – TCM Syndromes	TCM Syndrome reasoning	Gradually inferring the syndromes from the patient's pathogenesis.

Table 4. Annotated relation of entity in medical records.

Data evaluation. After the data annotation process was completed, a quality assessment was performed on the 300 medical records used in this study. Each medical record was thoroughly annotated to ensure the completeness and accuracy of the case information. Additionally, to reduce potential biases introduced by incomplete information, all data records were required to contain no missing values. Finally, to maintain the representativeness of the sample, rare medical records were excluded. The final statistics of all TCM medical records, classified according to the ICD-11 for Mortality and Morbidity Statistics (https://icd.who.int/en) are shown in Table 5. All 300 annotated medical records satisfied the aforementioned selection criteria.

Data Records

TCMEval-SDT benchmark dataset is available for access and download on Figshare[18], provided under the CC-BY 4.0 license. A total of 300 medical records were incorporated into TCMEval-SDT to create this benchmark dataset. The data were divided into training (n = 200), testing (n = 50), and validation sets (n = 50) following a 4:1:1 ratio. To aid the algorithm or model in performing diagnosis, four subtasks were designed for each case: (1) data extraction; (2) pathogenesis reasoning; (3) syndrome reasoning; and (4) explanation summary. Pathogenesis reasoning and syndrome reasoning were formatted as multiple-choice questions with ten options to assess the model's diagnostic reasoning ability. The multiple-choice questions on pathogenesis and syndrome reasoning are generated by a Python script (generate_multiple_choice_options.py), which is available on Figshare[18]. This script first collects the annotated pathogenesis and syndrome data from the TCMEval-SDT dataset, randomizes the options, and creates option lists for pathogenesis and syndrome. Finally, the script generates the ten options multiple-choice questions by selecting the correct options based on medical records, along with randomly selected options from the pathogenesis and syndrome lists. The Python script (evaluate.py) used for technical validation is also available on Figshare[18].

The TCMEval-SDT dataset includes three JSON files: (1) Train_TCM_Data_v1.json containing 200 cases, (2) Test_TCM_Data_v1.json containing 50 cases, and (3) Validation_TCM_Data_v1.json containing 50 cases. Table 6 provides an overview of the metadata for the dataset. In TCMTval-SDT, we have designed metadata in accordance with the FAIR principles. All information can be accessed via the CDE Portal, laying the foundation for scientific data sharing in the field of TCM, and supporting more researchers in using and developing TCMTval-SDT. It is important to note that the Test_TCM_Data_v1.json file and Validation_TCM_Data_v1.json file do not include the information of pathogenesis reasoning, syndrome reasoning and its correct answer options.

Technical Validation

In this chapter, we first introduce the criteria for evaluating answers. To validate and evaluate TCMEval-SDT, we selected four publicly available LLMs and randomly selected 50 medical records from the training set (n = 200). Using these records, we constructed zero-shot prompts to compare the TCM syndrome diagnosis capabilities of different LLMs.

Answer evaluation scheme. For the responses generated by LLMs, we have developed evaluation criteria tailored to each of the four tasks.

Task 1 Clinical information extraction:

$$S_c = \frac{|A \cap B|}{|A|} \tag{1}$$

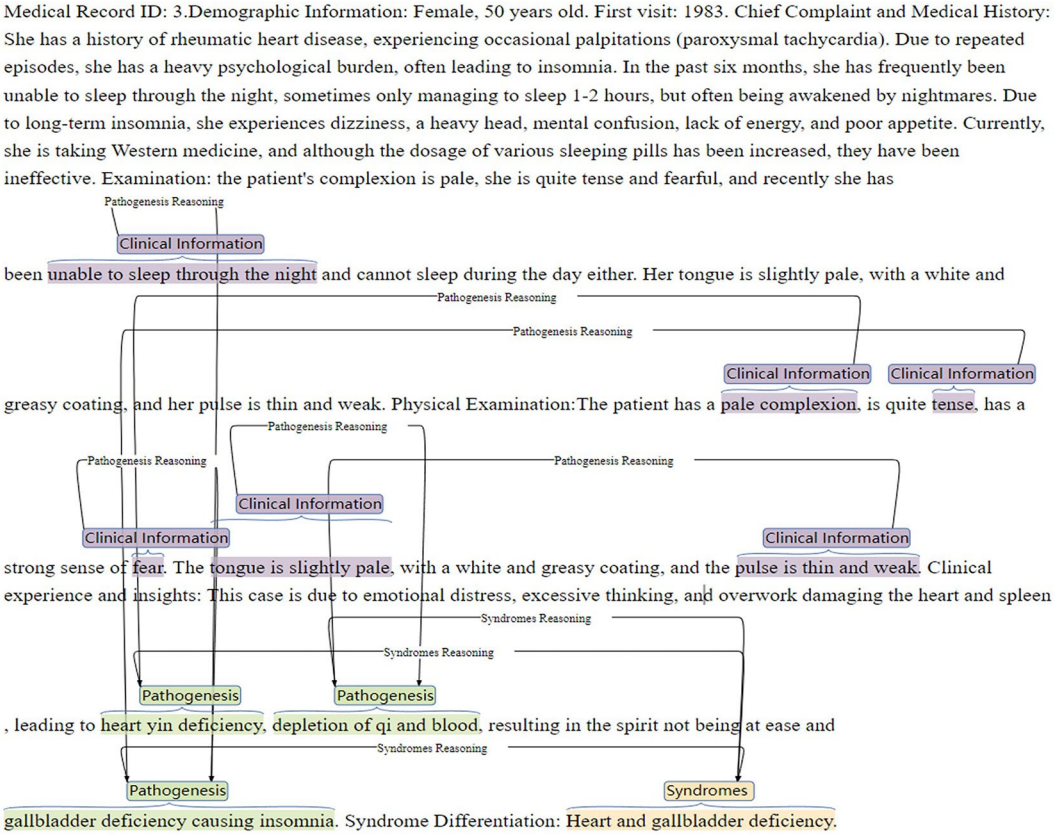

Fig. 3 Example of annotation for TCM clinical records. ***Note***. TCM = traditional Chinese medical.

where S_c is the score of Task 1; $|A|$ is the number of clinical information for medical record extracted by TCM Expert; $|A \cap B|$ is the number of intersections of clinical information for medical record extracted by TCM Expert and clinical information for medical record extracted by the LLMs

Task 2 Pathogenesis reasoning:

$$S_p = \frac{|A \cap B|}{|A| + |\overline{A} \cap B|} \tag{2}$$

where S_p is the score of Task 2; A is the set of correct answers of TCM pathogenesis; B is the set of answers selected of TCM pathogenesis by LLMs; $|A \cap B|$ is the number of options selected correctly by LLMs; $|A|$ is the number of correct answers of TCM pathogenesis; $|\overline{A} \cap B|$ is the number of options selected incorrectly by LLMs.

Task 3 Syndrome reasoning

$$S_s = \frac{|A \cap B|}{|A| + |\overline{A} \cap B|} \tag{3}$$

where S_s is the score of Task 3; A is the set of correct answers of TCM syndrome; B is the set of answers selected of TCM syndrome by LLMs. $|A \cap B|$ is the number of options selected correctly by LLMs; $|A|$ is the number of correct answers of TCM syndrome; $|\overline{A} \cap B|$ is the number of options selected incorrectly by LLMs.

Task 4 Explanatory summary:

$$S_r = ROUGE\_L(X, Y) \tag{4}$$

where S_r is the score of Task 4, calculated based on ROUGE-L[19]; X is the generated text; Y is the reference text.

Final score S_f for LLMs in TCM syndrome diagnosis task is:

$$S_f = \omega_1 S_c + \omega_2 S_p + \omega_3 S_s + \omega_4 S_r \tag{5}$$

where $\omega_1 = 0.2$, $\omega_2 = 0.3$, $\omega_3 = 0.4$ *and* $\omega_4 = 0.1$ are the weights assigned to each task score.

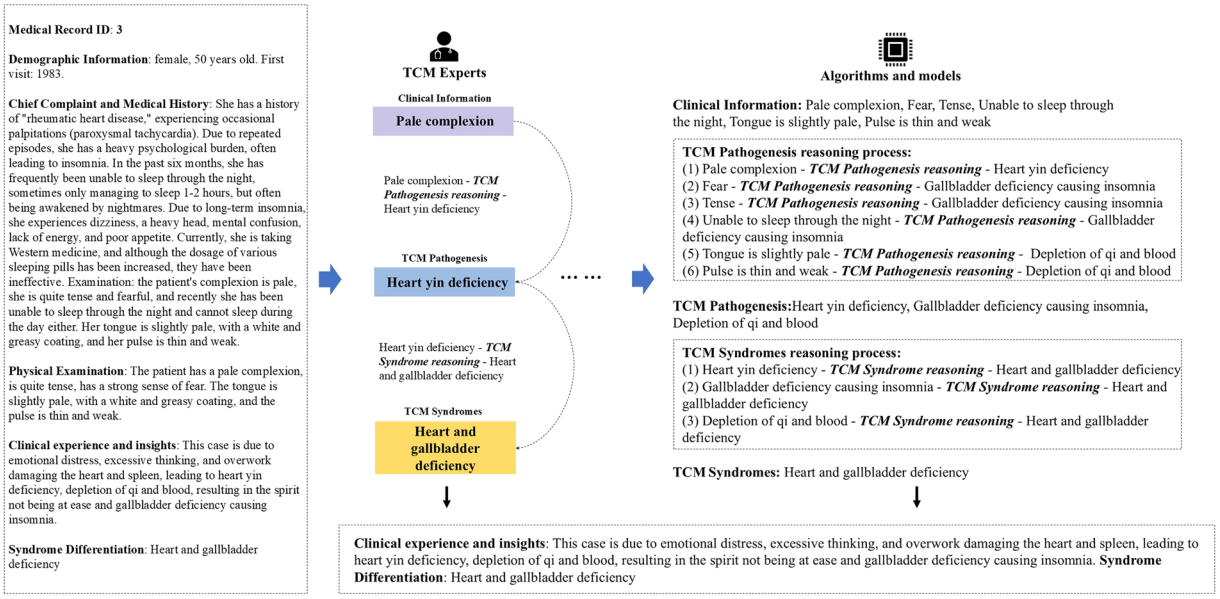

Fig. 4 Example of the thought process design in syndrome differentiation. The left side of the figure shows the patient's clinical data. Based on this data, TCM experts annotate and provide specific guided reasoning steps for the algorithms or models on the right side. Algorithms or models can follow these steps in a step-by-step reasoning process, thereby emulating the detailed procedure by TCM clinicians in syndrome diagnosis. ***Note***. TCM = traditional Chinese medical.

Type	URL	Count
Organ system disorders (TM1)	https://icd.who.int/browse/2024-01/mms/en#637432379	123
Other body system disorders (TM1)	https://icd.who.int/browse/2024-01/mms/en#111874651	96
Qi, blood and fluid disorders (TM1)	https://icd.who.int/browse/2024-01/mms/en#1055520234	4
Mental and emotional disorders (TM1)	https://icd.who.int/browse/2024-01/mms/en#1350805389	25
External contraction disorders (TM1)	https://icd.who.int/browse/2024-01/mms/en#672879010	13
Childhood and adolescence associated disorders (TM1)	https://icd.who.int/browse/2024-01/mms/en#386528385	39

Table 5. Statistics of 300 TCM medical records classified by disease according to the ICD-11.

Design of experiment and result analysis. In this study, we selected four publicly available LLMs: ChatGPT[20], Gemini 1.5-pro[21], ChatGLM-130B[22], and Tongyi Qianwen[23]. For each medical record, we designed zero-shot prompts. The validation process consisted of two steps: (1) testing the selected LLMs via API calls and manual queries; (2) manually organizing the responses from the LLMs. We queried the dataset (n = 50) using both API calls and manual questioning, initially verifying whether the responses from the LLMs adhered to the required format. Subsequently, TCM experts reviewed the responses to identify any null values or formatting inconsistencies.

We employed evaluation scripts to assess the responses generated by the LLMs, as illustrated in Fig. 5. Overall, ChatGLM-130B demonstrated the best performance, achieving the highest total weighted score of 24.7378, followed by Gemini 1.5-pro and ChatGPT with weighted scores of 23.1816 and 21.4753, respectively. ChatGLM-130B performed excellently in Task 1 and Task 2 (see Fig. 5b,c), with weighted scores of 6.2112 and 7.98. For Task 3 (see Fig. 5d) and Task 4, Gemini 1.5-pro demonstrated superior performance, with weighted scores of 9.5067 and 1.4765, and ChatGLM-130B performance was slightly inferior. During the experiments, we observed that ChatGLM-130B and Gemini 1.5-pro demonstrated notable proficiency in TCM diagnostic tasks, achieving commendable scores across all four sub-tasks.

Usage Notes

The TCMEval-SDT benchmark dataset is available for download and review on Figshare[18]. This dataset was created to assess the capabilities of algorithms and models in the diagnosis of TCM syndromes. It has been meticulously curated and annotated by TCM experts and includes the following components: Medical Record ID, Medical Data, Explanatory Summary, Syndrome Differentiation, Clinical Information, TCM Pathogenesis, TCM Syndrome, Options of TCM Pathogenesis, Options of TCM Syndrome, Answers of TCM Pathogenesis, and Answers of TCM Syndrome.

However, the released dataset has several limitations. Currently, the dataset is relatively small in size, for example, it contains only four medical record related Qi, blood and fluid disorders,. In the future, we plan to

Metadata Name	Metadata ID/ID	Metadata URL	Description
Medical Record ID	DE0087751	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087751	Unique identifier for each medical report.
Clinical Data	DE0087752	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087752	Includes the patient's demographic Information, chief complaint, medical history, and physical examination.
Explanatory Summary	DE0087753	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087753	The summary provided by the TCM clinician is based on personal insights and experiences related to the patient's diagnosis, treatment, and changes in condition.
Syndrome Differentiation	DE0087754	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087754	Based on TCM theory, the syndrome is determined through a comprehensive analysis of information gathered using the four diagnostic methods (inspection, listening/smelling, inquiry, and palpation).
Clinical Information	DE0087755	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087755	During the syndrome diagnosis process, TCM clinicians extract clinical information from the patient's data, including physical signs and medical history.
TCM Pathogenesis	DE0087756	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087756	Provides a detailed description of the mechanisms of disease occurrence, development, and changes, including disease nature, location, progression, changes in qi and blood, and prognosis.
TCM Pathogenesis Reasoning	DE0088288	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0088288	Gradually inferring the pathogenesis from the patient's clinical information.
TCM Syndrome	DE0087757	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087757	The name of the disease as classified within the TCM syndrome classification system.
TCM Syndrome Reasoning	DE0088289	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0088289	Gradually inferring the syndromes from the patient's pathogenesis.
Options of TCM Pathogenesis	DE0087758	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087758	Available options for TCM pathogenesis.
Options of TCM Syndrome	DE0087759	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087759	Available options for TCM Syndrome.
Answers of TCM Pathogenesis	DE0087760	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087760	Correct options for TCM pathogenesis.
Answers of TCM Syndrome	DE0087761	https://cdeportal.bmicc.cn/cde/detail?data_element_id=DE0087761	Correct options for TCM syndrome.

Table 6. Metadata for clinical record in the TCMEval-SDT dataset.

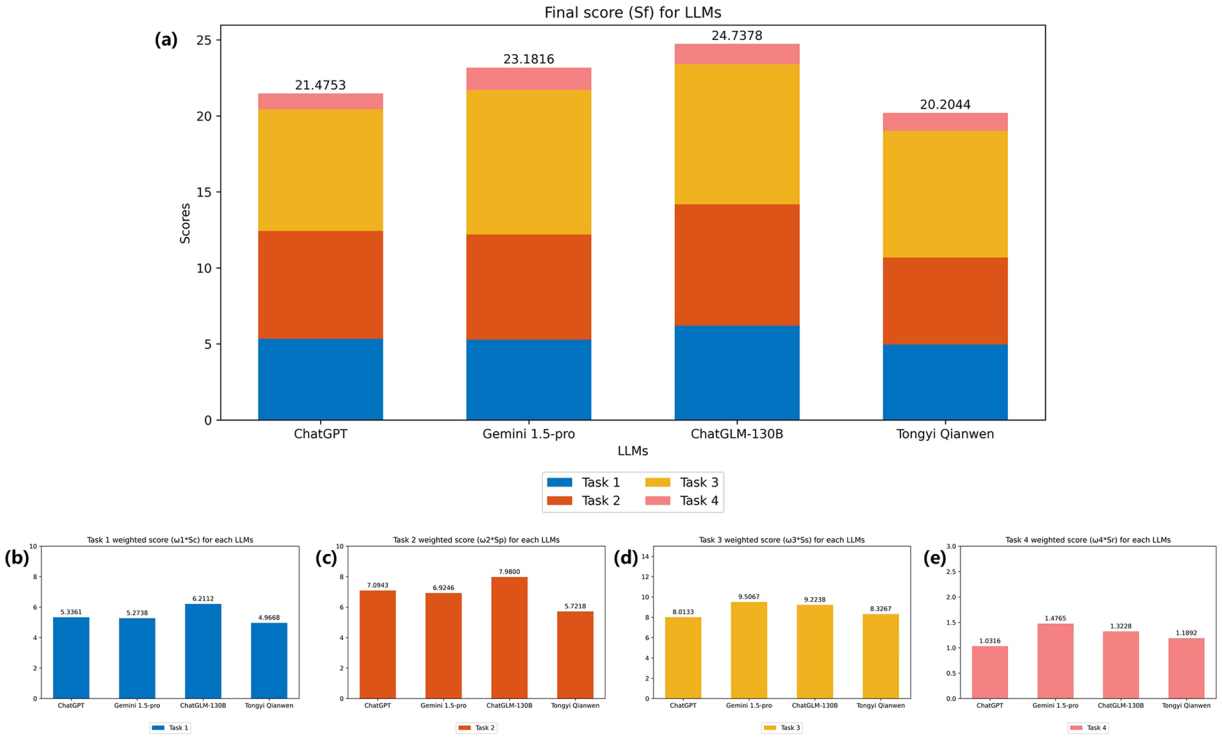

Fig. 5 Performance of LLMs on the TCMEval-SDT benchmark dataset(n = 50). Fifty clinical records from the training data were selected for validation, either through API or manual inquiry, and the results were statistically analysed and all scores were calculated based on task weights:
Task $1(\omega_1 = 0.2)$, Task $2(\omega_2 = 0.3)$, Task $3(\omega_3 = 0.4)$, Task $4(\omega_4 = 0.1)$. The findings indicate that ChatGLM-130B achieved the highest overall performance with a total weighted score of 24.74, excelling in Task 1 and Task 2 with weighted scores of 6.21 and 7.98, respectively. Gemini 1.5 Pro performed best in Task 3 and Task 4, with weighted scores of 9.51 and 1.48, respectively. ***Note***. TCM = traditional Chinese medical; LLMs = large language models.

include additional medical records and gradually expand the overall size of the dataset to ensure a more balanced distribution of disease types. Additionally, we aim to incorporate rare disease cases from TCM to develop a more specialized diagnostic dataset. We invite enthusiasts to join our community in enhancing this syndrome diagnosis benchmark dataset and contribute to the advancement of scientific data sharing and reuse.

Code availability

The Python scripts used in the curation and the technical validation are available on Figshare[18] and GitHub (https://github.com/zhuyan166/TCMEval/tree/main/evaluation/TCMEval-SDT).

Received: 10 October 2024; Accepted: 6 March 2025;
Published online: 13 March 2025

References

1. Guo, C. *et al*. Exploring the Mechanism of Action of Canmei Formula Against Colorectal Adenoma Through Multi-Omics Technique. *Front Cell Dev Biol* **9**, 778826, https://doi.org/10.3389/fcell.2021.778826 (2021).
2. Zhang, M. *et al*. Semen Cassiae Extract Improves Glucose Metabolism by Promoting GlUT4 Translocation in the Skeletal Muscle of Diabetic Rats. *Front Pharmacol* **9**, 235, https://doi.org/10.3389/fphar.2018.00235 (2018).
3. Liu, Y. *et al*. Artemisinins ameliorate polycystic ovarian syndrome by mediating LONP1-CYP11A1 interaction. *Science* **384** https://doi.org/10.1126/science.adk5382 (2024).
4. Liu, S. *et al*. Effect of an Herbal-Based Injection on 28-Day Mortality in Patients With Sepsis: The EXIT-SEP Randomized Clinical Trial. *JAMA Intern Med* **183**, 647–655, https://doi.org/10.1001/jamainternmed.2023.0780 (2023).
5. Hu, L. *et al*. A Systematic Study of Mechanism of Sargentodoxa cuneata and Patrinia scabiosifolia Against Pelvic Inflammatory Disease With Dampness-Heat Stasis Syndrome via Network Pharmacology Approach. *Front Pharmacol* **11**, 582520, https://doi.org/10.3389/fphar.2020.582520 (2020).
6. Xu, H. *et al*. A comprehensive review of integrative pharmacology-based investigation: A paradigm shift in traditional Chinese medicine. *Acta Pharm Sin B* **11**, 1379–1399, https://doi.org/10.1016/j.apsb.2021.03.024 (2021).
7. Rong, G., Mendez, A., Bou Assi, E., Zhao, B. & Sawan, M. Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering* **6**, 291–301, https://doi.org/10.1016/j.eng.2019.08.015 (2020).
8. Da, Q. *et al*. DigestPath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis* **80**, 102485, https://doi.org/10.1016/j.media.2022.102485 (2022).
9. Singhal, K. *et al*. Large language models encode clinical knowledge. *Nature* **620**, 172–180, https://doi.org/10.1038/s41586-023-06291-2 (2023).
10. Yue, W. *et al*. TCMBench: A Comprehensive Benchmark for Evaluating Large Language Models in Traditional Chinese Medicine. arXiv preprint arXiv:2406.01126 (2024).
11. Mucheng, R. *et al*. in *Proceedings of the 21st Chinese National Conference on Computational Linguistics*. 908–920.
12. Wei, J. *et al*. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 (2022).
13. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
14. Wang, Z., Liu, L., Yao, K., Wang, J. & Zhu, Y. in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 3727–3732.
15. Zhang, S., Wang, Z., Yao, K., Liu, L. & Zhu, Y. in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 4718–4725.
16. Z Tao. Recommendation on classification and quality evaluation for case reports in traditional Chinese medicine format: taking COVID-19 for incidence. Modern Chinese Clinical Medicine, **30**, 17–20 (In Chinese) (2023).
17. Riley, D. S. *et al*. CARE guidelines for case reports: explanation and elaboration document. *Journal of clinical epidemiology* **89**, 218–235 (2017).
18. Zhu, Y. TCMEval-SDT. *figshare* https://doi.org/10.6084/m9.figshare.27184596.v4 (2024).
19. Lin, C.-Y. in *Text summarization branches out*. 74–81.
20. OpenAI. *Introducing ChatGPT*, https://openai.com/blog/chatgpt (2022).
21. Google. *Gemini*, https://gemini.google.com/ (2023).
22. Zeng, A. *et al*. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414* (2022).
23. Bai, J. *et al*. Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

Acknowledgements

This work was supported by Beijing Natural Science Foundation (7252253, 7254504), National Natural Science Foundation of China (82174534), Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (No. CI2021A05306), National Chinese Medicine Examination 2023 Scientific Research Project (TB2023008) and CAMS Innovation Fund for Medical Sciences (CIFMS, No.2021-I2M-1-057). We acknowledge the use of GPU and High-performance Computing Platform at the Center for Bioinformatics. Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College.

Author contributions

Data curation, M.H., Y.L., Y.H., S.P., X.Y. and Y.Z.; Resources, M. H., Y.L., Y.H., K.Y.; Software, Z.W.; Project administration, Y.Z.; Validation, Z.W., M.H., K.Y.; Writing - original, Z.W. and M.H.; Writing - review & editing, P.S., L.H, X.Y. and Y.Z., all authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.Y. or Y.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025