Acupuncture and Herbal Medicine

Large language models in traditional Chinese medicine: a systematic review

Zhe Chen[1,2,3], Hui Wang[1], Chengxian Li[1,2], Chunxiang Liu[1], Fengwen Yang[1], Dong Zhang[1], Alice Josephine Fauci[4,5,*], Junhua Zhang[1,2,*]

[1]Evidence-based Medicine Center, Tianjin University of Traditional Chinese Medicine, Tianjin, China; [2]Haihe Laboratory of Modern Chinese Medicine, Tianjin, China; [3]National Key Laboratory of Chinese Medicine Modernization, Tianjin University of Traditional Chinese Medicine, Tianjin, China; [4]Italian National Institute of Health, Rome, Italy; [5]Istituto Superiore di Sanità, Centro Eccellenza Clinica, Qualità e Sicurezza delle Cure, Rome, Italy

Abstract

Objective: Generative artificial intelligence (AI) technology, represented by large language models (LLMs), has gradually been developed for traditional Chinese medicine (TCM); however, challenges remain in effectively enhancing AI applications for TCM. Therefore, this study is the first systematic review to analyze LLMs in TCM retrospectively, focusing on and summarizing the evidence of their performance in generative tasks.

Methods: We extensively searched electronic databases for articles published until June 2024 to identify publicly available studies on LLMs in TCM. Two investigators independently selected and extracted the related information and evaluation metrics. Based on the available data, this study used descriptive analysis for a comprehensive systematic review of LLM technology related to TCM.

Results: Ten studies published between 2023 and 2024 met our eligibility criteria and were included in this review, including 40% LLMs in the TCM vertical domain, 40% containing TCM data, and 20% honoring the TCM contribution, with a foundational model parameter range from 1.8 to 33 billion. All included studies used manual or automatic evaluation metrics to evaluate model performance and fully discussed the challenges and contributions through an overview of LLMs in TCM.

Conclusions: LLMs have achieved significant advantages in TCM applications and can effectively address intelligent TCM tasks. Further in-depth development of LLMs is needed in various vertical TCM fields, including clinical and fundamental research. Focusing on the functional segmentation development direction of generative AI technologies in TCM application scenarios to meet the practical needs-oriented demands of TCM digitalization is essential.

Keywords: Generative artificial intelligence, Intelligence clinical applications, Large language model, Systematic review, Traditional Chinese medicine

Graphical abstract: http://links.lww.com/AHM/A152.

Introduction

Large language models (LLMs), as representations of artificial intelligence (AI), offer new methodologies and technologies for interpreting and applying clinical support, particularly in natural language processing (NLP) for generative tasks in traditional Chinese medicine (TCM) and classification in healthcare settings[1–3]. LLMs can interact with progression and learn features in human language by leveraging extensive large datasets with pre-training. They have gained significant milestones in generating, interpreting, and responding to achieve expressive and inferential capabilities for medical AI[1,4–5]. Based on clinical applications and individualized clinical decision-making powered by LLMs, generative AI (GAI) technology has been used to handle complex TCM tasks by expanding capabilities and achieving superior performance, accelerating the digital and intelligent construction of TCM[2–3,6].

The TCM theoretical framework, with its unique concepts such as the Zang-Fu theory, eight-principle syndrome differentiation, and holistic diagnostics, presents a distinctive challenge for LLMs accustomed to modern medicine paradigms[7–8]. The TCM professional knowledge base (a pre-trained corpus reflecting the practice and principles of TCM) is a prerequisite for building vertical-domain LLMs that contain diagnostic methods, therapeutic concepts, clinical rules of medication, Chinese patent medicine, TCM compounds, relevant targets, and signaling pathways in line with TCM characteristics[6,8–9]. Many studies have explored integrating LLM-related technologies to enhance AI-aided decision support in the TCM domain and achieve multitasking collaboration in TCM diagnostic methods,

*Corresponding author. Junhua Zhang, E-mail: zjhtcm@foxmail.com; Alice Josephine Fauci, E-mail: alice.fauci@iss.it.

Received 4 September 2024 / Accepted 21 January 2025

How to cite this article: Chen Z, Wang H, Li CX, Liu CX, Yang FW, Zhang D, Fauci AJ, Zhang JH. Large language models in traditional Chinese medicine: a systematic review. Acupunct Herb Med 2025;5(1):57–67. DOI: 10.1097/HM9.0000000000000143

Copyright © 2025 Tianjin University of Traditional Chinese Medicine. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

personalized and specialized treatments, and basic research explorations[2–3,10].

From improving model accuracy to personalized treatment regimens and further facilitating the exploration of underlying mechanisms, the integration of TCM features into LLMs with fine-tuning and prompt templates could revolutionize the existing TCM diagnostic and therapeutic modes and even exert a paradigmatic impact on the intelligent research and development of novel TCM pharmaceuticals[11–13]. However, LLMs face numerous obstacles, such as fine-tuning in specific domains and establishing assessment metrics that can accurately reflect a model's performance in the context of TCM[10–12]. TCM, as the holistic healthcare paradigm centered on "patient-symptom," uses TCM-specific syndrome differentiation for diagnostic discrimination in clinical practice, which requires LLMs based on formidable "computational capacity combined with algorithms" to demonstrate an in-depth understanding beyond NLP for TCM features[3,14–15].

A lack of comprehensive summary exists regarding LLMs in TCM that meets the needs of researchers, patients, and clinicians for GAI-based TCM, nor can the potential factors and assessment systems driving TCM research using AI be fully elucidated. This systematic review aims to synthesize current perspectives and practices on the capabilities of LLMs associated with TCM tasks for the first time, highlighting the latest developments and applications. Through a comprehensive performance analysis, we assessed the impact of LLMs on TCM and provided insights into future policies and practices for GAI in TCM.

Methods

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement with a transparent, complete, and accurate methodology[16]. We reviewed and analyzed LLMs by screening the literature, focusing on their status in TCM, and registered our protocol with CRD42024563281 in the International Prospective Register of Systematic Reviews (PROSPERO) (https://www.crd.york.ac.uk/PROSPERO/).

Eligibility criteria

We included studies describing GAI technology using LLMs in TCM (including TCM clinical decision support in diagnosis and treatment, homage to renowned ancient TCM scientists in the medical domain, and following the TCM concept). We reported the original data and results to illustrate prediction performance. All the included studies required explicit naming of the LLMs.

The exclusion criteria were as follows: NLP without LLMs; LLMs without TCM tasks, homages, or concepts; no vertical-domain adjustments based on LLMs; functional testing for LLMs in certain topics related to TCM; unavailable full text; duplicate publications; non-generative pre-trained models; other pathways to publish TCM LLMs without original data; no reported results of model performance; and publication types in articles such as reviews, commentaries, views, opinions, and letters to the editor.

Search strategy

We conducted a search strategy using keywords developed by specialists to identify studies that reflected the development and validation of LLMs for applications in TCM methods or concepts. We targeted preprints and peer-reviewed publications reporting the emergence and application of LLMs in TCM, published up to June 2024. The systematic search focused on the following databases: CNKI, Wanfang, VIP, SinMed, PubMed, Cochrane Library, Embase, and Web of Science to find relevant peer-reviewed articles, conference proceedings, and preprints without language restrictions and checked the references of all selected manuscripts to explore more relevant studies. The detailed search strategy and PRISMA 2020 checklist are included in Supplementary Files 1 and 2, http://links.lww.com/AHM/A152, including variations in keywords for "traditional Chinese medicine" and "large language models."

Selection criteria and data collection process

References identified from the search were independently screened by two investigators (Zhe Chen and Hui Wang) for titles and abstracts using pre-specified inclusion criteria after removing duplicates. The final selection scanned all potential studies for full-text review to identify studies that met the eligibility criteria. A third researcher resolved any inconsistencies.

The following detailed items regarding the specific LLMs in TCM were extracted from the characteristics and related features, including the first author, year of publication, foundational model, status, institutions, open-source situation, purpose, applications of LLMs, limitations, and conclusions. We extracted evaluation metrics to assess the performance of the LLMs with automatic and manual evaluations. When there were missing data or serious problems, we contacted the first or corresponding author.

Evaluation metrics

When evaluating the capability of LLMs, it is necessary to comprehensively consider their performance in multiple dimensions to ensure stability when measuring the performance of the model. We used diverse evaluation criteria and systems to define two main categories of evaluation metrics: automatic and manual. Automatic evaluation metrics, calculated automatically based on the model structure, have been widely applied in the preliminary assessment and development phase to reflect quantitative performance, including accuracy, Bilingual Evaluation Under Study (BLEU), General Language Understanding Evaluation (GLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Distinc, et al. Manual evaluation metrics, which require experts to analyze and evaluate the outputs generated by LLMs, involve several specific evaluation dimensions, and the final evaluation results are primarily based on subjective

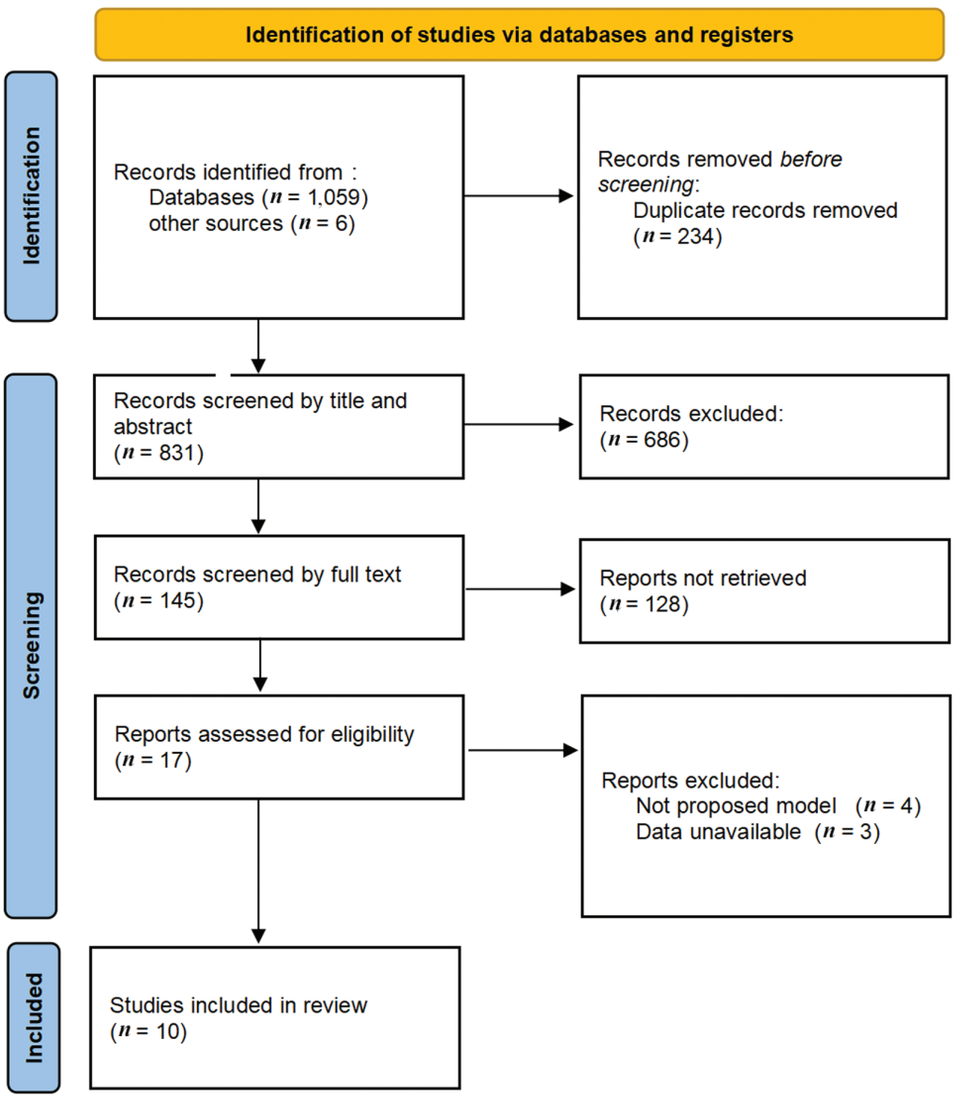


Figure 1. Summary of search and selection process.

human feedback in aspects such as safety, fluency, professionalism, semantic similarity, and others.

Data synthesis and statistical analysis

We conducted a qualitative analysis of the performance results and fundamental features of LLMs in TCM by systematically reviewing original studies based on current applications within the AI domain. In this comprehensive overview, we did not perform a statistical analysis because of the variability of LLMs with inconsistent availability (diversity and complexity of data, unified measurement standards, heterogeneity of studies, and the purpose of the applications); therefore, we focused on summarizing and synthesizing to support this descriptive report.

Results

Study selection and characteristics

After removing duplicates, we screened 1,059 records from multiple public databases using our search, of which 831 studies were identified with potential eligibility, and six studies were added through additional searches. After title and abstract review, 145 studies met the criteria for full-text review, of which 128 were excluded owing to a lack of relevance in TCM and/or AI-based generative tasks, and 17 were assessed for eligibility. After excluding studies that did not meet our eligibility criteria, 10 studies were included in the systematic review (Figure 1).

Characteristics of included studies

All the included studies used generative interactive responses for clinical diagnosis and treatment decision support[6,8–9,17–23]. Ten LLMs reported different types of TCM studies, including TCM vertical domain (40%), Chinese medicine containing TCM data (40%), and Chinese medicine honoring the contribution of TCM (20%). All the above studies on TCM LLMs were published between 2023 and 2024, with six studies published in 2023 and four published in 2024. Eighty percent of the studies were conducted through multi-unit collaborations between different institutions, including two multi-region studies, whereas only two LLMs were based on independent research and development.

Table 1

Basic characteristics of LLMs in TCM

LLMs	First author	Types	Publication date	Title	Institution	Developer with P&R	Publish
TCM-GPT	Yang[17]	LLMs in TCM	2023	TCM-GPT: Efficient Pre-training of Large Language Models for Domain Adaptation in Traditional Chinese Medicine	Multi-regions, Multi-units (Beijing University of Posts and Telecommunications University College London)	N	Preprint
QiBo	Zhang[8]	LLMs in TCM	2024	Qibo: A Large Language Model for Traditional Chinese Medicine	Multi-units (Tianjin University Tianjin University of Traditional Chinese Medicine TianDaZhiTu)	Y	Preprint
MedChatZH	Tan[6]	LLMs in TCM	2024	MedChatZH: A tuning LLM for traditional Chinese medicine consultations	Multi-regions, Multi-units (East China University of Science and Technology Shanghai Artificial Intelligence Laboratory Shanghai Jiao Tong University The University of Sydney)	Y	Peer review
CPMI-ChatGLM	Liu[9]	LLMs in TCM	2024	CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions	Multi-units (Anhui University of Traditional Chinese Medicine China Academy of Chinese Medical Sciences)	Y	Peer review
Ming-MOE	Liao[18]	LLMs in CM (containing TCM)	2023	MING-MOE: Enhancing Medical Multi-Task Learning in Large Language Models with Sparse Mixture of Low-Rank Adapter Experts	Multi-units (Shanghai Jiao Tong University Fudan University Shanghai Artificial Intelligence Laboratory)	N	Preprint
HuatuoGPT	Zhang[19]	LLMs in CM (containing TCM)	2023	HuatuoGPT, toward Taming Language Model to Be a Doctor	Multi-units (Shenzhen Research Institute of Big Data The Chinese University of Hong Kong)	N	Preprint
IvyGPT	Wang[20]	LLMs in CM (containing TCM)	2023	IVYGPT: INTERACTIVE CHINESE PATHWAY LANGUAGE MODEL IN MEDICAL DOMAIN	Multi-units (Macao Polytechnic University Opera inc Shanghai Jiao Tong University)	N	Preprint
Zhongjing-LLaMa	Yang[21]	LLMs in CM (containing TCM)	2023	Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue	Independent research and development (Zhengzhou University)	N	Preprint
HuoTuo (BenTsao)	Wang[22]	LLMs in CM (honoring TCM)	2023	HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge	Independent research and development (Harbin Institute of Technology)	N	Preprint
Bianque	Chen[23]	LLMs in CM (honoring TCM)	2023	BianQue: Balancing the Questioning and Suggestion Ability of Health LLMs with Multi-turn Health Conversations Polished by ChatGPT	Multi-units (South China University of Technology Guangdong Women and Children Hospital South China University of Technology Pazhou Lab)	Y	Preprint

CM: Chinese medicine; LLMs: Large language models; N: No; P&R: Pharmacy and Medicine; TCM: Traditional Chinese medicine; Y: Yes.

Reporting on the background of the developers showed that 40% of the studies had developers with pharmacies and medicines, and only one of them was developed by TCM University as the first organization. All the studies included in this review were published and disseminated as articles. At the time of publication, only two studies had been published after peer review, while the remaining studies were published as preprints (Table 1).

Performance summary

All the studies that met the inclusion criteria were based on an open-source foundational model for constructing TCM LLMs (Table 2). Among them, six studies were based on the LLaMa model (QiBo, MedChatZH, HuatuoGPT, IvyGPT, Zhongjing-LLaMa, HuoTuo); MedChatZH and HuatuoGPT were based on the Baichuan model; CPMI-ChatGLM and Bianque were based on the ChatGLM model; TCM-GPT was based on the BLOOM model; and Ming-MOE was based on the Qwen model. The parameter range of the foundational model is from 1.8 to 33 billion, with IvyGPT (33 B) including LLMs with the largest parameters and Ming-MOE (1.8 B) with the smallest. The parameter values of 1.8 B, 4 B, 14 B, and 33 B involved four LLMs with the foundational model (Qwen 1.5 in 1.8 B, 4 B, and 14 B versions and LLaMA-33 B) reported in two studies. In contrast, the 6 B, 7 B, and 13 B parameter values were related to two, six, and three studies, respectively. Regarding the availability of data and codes, only 20% of the studies were closed-source, whereas the remaining 80% were open-source studies.

Manual and automatic evaluation metrics were included to assess model performance. Manual evaluation metrics are mainly based on assessments by experts or AI feedback involving the SPF metrics in QiBo (safety: Win 38%–95%, tie 3%–29%, loss 1%–33%; professionalism: Win 39%–96%, tie 1%–33%, loss 3%–38%; fluency: Win 32%–96%, tie 2%–37%, loss 2%–35%) and Zhongjing-LLaMa (safety: Win 26%–99%, tie 0%–29%, loss 1%–46%; professionalism and fluency: Win 40%–98%, tie 1%–27%, loss 4%–33%); SUS metrics in CPMI-ChatGLM (safety: 2.88; usability: 2.78; smoothness: 2.950) and HuoTuo (safety: 2.88; usability: 2.12; smoothness: 2.47). Other manual evaluation metrics included consistent scores for TCM with AI feedback of 8.6 and 8.8 in HuatuoGPT and a semantic similarity score of 93.58 in IvyGPT.

Numerous automatic evaluation metrics focus on assessing a model's performance in classification tasks (accuracy and F1 score) and natural language generation tasks (BLEU, GLEU, and ROUGE).

The performance of the automatic evaluation metrics indicated that two studies reported accuracy in medicine (TCM-GPT for TCM diagnosis: 0.264; Ming-MOE-1.8 B: 41.58, Ming-MOE-4 B: 50.31, Ming-MOE-7 B: 57.03, Ming-MOE-14 B: 63.2) and in TCM examinations (TCM-GPT: 0.29; Ming-MOE-1.8 B: 33.96; 4 B: 45; 7 B: 49.58; 14 B: 59.79). Only Ming-MOE reported an F1 score of 1.8 B (65.4), 4 B (69.48), 7 B (71.82), and 14 B (72.85).

In the NLP tasks, the performance results of the automatic evaluation metrics were as follows: QiBo achieved ROUGE-L values of 0.72, 0.61, and 0.64 in handling NLP for TCM-NER, TCM-RP, and TCM-SD, respectively. MedChatZH, HuatuoGPT, and Bianque reported BLEU-1 (56.14; 25.05; 13.4725), BLEU-2 (32.14; 13.07; 8.895), BLEU-3 (17.58; 7.39; 6.6025), BLEU-4 (9.17; 4.28; 5.03), GLEU (10.32; 8.13; not reported), ROUGE-1 (35.99; 27.63; 19.46), ROUGE-2 (10.31; 7.28; 4.71), and ROUGE-L (21.77; 17.67; 17.0425). For the auxiliary diagnosis and treatment of Chinese patent medicine, CPMI-ChatGLM achieved BLEU-4 (0.7641), ROUGE-1 (0.8188), ROUGE-2 (0.7738), ROUGE-L (0.8107), and BART scores (-2.4786) for CPM recommendations. Distinct metrics were reported only by HuatuoGPT, with Distinct-1 at 0.73 and Distinct-2 at 0.93. Bianque defined the model's proactive questioning ability as a new metric to measure the performance of questions as 0.6525.

Descriptive overview of LLMs in TCM

The challenges, contributions, and limitations of LLMs in TCM across the included studies varied in terms of differences and similarities (Table 3). Among the studies in the challenge, three studies identified a lack of domain knowledge in line with the TCM characteristics, which failed to effectively summarize TCM diagnosis and treatment data, thereby impacting the enhancement of computational efficiency. Three studies suggested that LLMs could not be prescribed like clinical doctors and lacked the ability for multi-turn communication and understanding; two studies suggested that the absence of medical knowledge or low-parameter, low-quality datasets limits the inferential capabilities of LLMs; one pointed out the current inability to generate detailed instructions for the use of traditional CPMs; one considered the annotation of medical knowledge to be insufficient; and one study highlighted that Chinese medical LLMs perform poorly and are prone to hallucinations.

All the included studies discussed the contributions of LLMs to medical tasks as follows:

TCM-GPT uses a new domain adaptation method in TCM (TCMDA), utilizing TCM-Corpus-1B and LoRA to enhance TCM model performance through domain adaptation and underscore the importance of specialized pre-training[17].

QiBo outperforms larger open-source models with a TCM-specific corpus and SFT pre-training and sets TCM benchmarks[8].

MedChatZH proposed a GAI for TCM inquiries and outperformed baseline models with a specialized corpus, emphasizing ethical medical practice and recommending qualified physician guidance[6].

CPMI-ChatGLM, the first domain-specific LLM in CPM, utilizes a high-quality dataset for instruction tuning and publicly releases the CPMI dataset as a TCM resource[9].

MING-MOE was the first MOE model capable of handling diverse medical tasks, and other medical tasks performed at a superior level[18].

HuatuoGPT, the first RLAIF-based medical LLM, used real and distilled data and outperformed existing open-source LLMs, even those most similar to ChatGPT (GPT-3.5-turbo)[19].

IvyGPT, which integrates supervised training, reward, and reinforcement learning, offers a 33 B parameter model with low computational power and assesses it against other LLMs using high-quality, real-life medical dialogue datasets[20].

Zhongjing-LLaMa, a novel medical LLM with process-oriented training, surpassed Chinese medical models across

Table 2

Model performance of LLMs in TCM tasks

LLMs	Foundational model	Availability	Manual evaluation metrics	Automatic evaluation metrics
TCM-GPT	BLOOM-7B	Closed-source	NR	Accuracy in medicine (0.264) Accuracy in TCM Examination (0.29)
QiBo	Chinese-LLaMA-7B/13B	Closed-source	Safety (Win 38%–95%, tie 3%–29%, loss 1%–33%) Professionalism (Win 39%–96%, tie 1%–33%, loss 3%–38%) Fluency (Win 32%–96%, tie 2%–37%, loss 2%–35%)	ROUGE-L: TCM-NER (0.72) TCM-RP (0.61) TCM-SD (0.64)
MedChatZH	Baichuan-7B Ziya-LLaMA-7B-Reward	Open source	NR	BLEU-1 (56.14) BLEU-2(32.14) BLEU-3 (17.58) BLEU-4 (9.17) GLEU (10.32) ROUGE-1 (35.99) ROUGE-2 (10.31) ROUGE-L (21.77)
CPMI-ChatGLM	ChatGLM-6B	Open source	Safety (2.88) Usability (2.78) Smoothness (2.95)	BLEU-4 (0.7641) ROUGE-1 (0.8188) ROUGE-2 (0.7738) ROUGE-L (0.8107) BART Score (–2.4786)
Ming-MOE	Qwen1.5-1.8B/4B/7B/14B	Open source	NR	Accuracy in medicine (1.8B: 41.58; 4B: 50.31; 7B: 57.03; 14B: 63.2) Accuracy in TCM Examination (1.8B: 33.96; 4B: 45; 7B: 49.58; 14B: 59.79) F1 score (1.8B: 65.4; 4B: 69.48; 7B: 71.82; 14B: 72.85)
HuatuoGPT	Baichuan-7B Ziya-LLaMA-13B-v1	Open source	Consistent scores for TCM with AI Feedback (8.6 and 8.8)	BLEU-1 (25.05) BLEU-2 (13.07) BLEU-3 (7.39) BLEU-4 (4.28) GLEU (8.13) ROUGE-1 (27.63) ROUGE-2 (7.28) ROUGE-L (17.67) Distinct-1 (0.73) Distinct-2 (0.93)
IvyGPT	LLaMA-33B	Open source	Semantic similarity Score (93.58)	NR
Zhongjing-LLaMa	Ziya-LLaMA-13B-v1	Open source	Safety (Win 26%–99%, tie 0%–29%, loss 1%–46%) Professionalism and Fluency (Win 40%–98%, tie 1%–27%, loss 4%–33%)	NR
HuoTuo (BenTsao)	LLaMA-7B	Open source	Safety (2.88) Usability (2.12) Smoothness (2.47)	NR
Bianque	ChatGLM-6B	Open source	NR	BLEU-1 (13.4725) BLEU-2 (8.895) BLEU-3 (6.6025) BLEU-4 (5.03) ROUGE-1 (19.46) ROUGE-2 (4.71) ROUGE-L (17.0425) Model's Proactive Questioning Ability (0.6525)

BLEU: Bilingual Evaluation Under Study; GLEU: General Language Understanding Evaluation; LLMs: Large language models; NR: Not report; ROUGE: Recall-Oriented Understudy for Gisting Evaluation; TCM: Traditional Chinese medicine; TCM-NER: The entity recognition task of TCM prescriptions; TCM-RP: TCM reading comprehension quiz pair construction; TCM-SD: Syndrome differentiation in TCM.

dimensions and matched ChatGPT in specific domains using a multidimensional CMtMedQA dataset[21].

Huo Tuo: The first open-source Chinese biomedical LLM integrated structured and unstructured medical knowledge for accuracy and proposed a manual evaluation metric for evaluating security, usability, and smoothness[22].

Bianque: A proposed healthy LLM with balanced Q&A capabilities was validated using the BianQue Corpus, showing excellent multi-round questioning proficiency[23].

The summary of the limitations of the included LLMs indicated that the LLM response did not guarantee stability and accuracy, which could have been used for physician assistance or to offer personalized suggestions and should not have been directly applied to clinical decision-making. In application scenarios involving multimodal tasks, the existing LLMs in the TCM domain are only capable of handling NLP-based medical knowledge questions and answers and cannot meet the clinical diagnostic needs for the recognition and processing of images, medical imaging, and patient physiological signals. The performance of some LLMs was constrained by the quality of the training data and its volume. Performance was influenced by the original training texts, culture, and context, which limited the models' knowledge depth and breadth, further affecting the quality of the output results (MedChatZH, CPMI-ChatGLM, and IvyGPT). The information provided by generative techniques may include misleading elements, leading to significant ethical and moral risks (eg, HuatuoGPT and IvyGPT). Furthermore, when engaging with LLMs in a question-and-answer format, there is the potential for models to acquire user information actively, raising concerns and risks related to personal privacy (Bianque).

Discussion

Because of their real-time interactive responses to user queries, LLMs have become indispensable to AI applications in TCM clinical practice. We conducted the first comprehensive systematic review of the performance and applications of existing studies on LLMs in TCM, focusing on generative tasks, to explore their potential prospects for TCM AI decision-support technology. A descriptive analysis of the results revealed that studies of LLMs for TCM applications were predominantly published between 2023 and 2024, focusing on multi-institutional collaborations in most studies. A lack of leadership and participation from TCM research institutions in LLMs exists, which does not effectively ensure that the development of LLMs in the vertical domain of TCM conforms to the characteristics of TCM clinical diagnosis and treatment.

We observed that most studies made their open-source information publicly available to enhance data and algorithm openness, transparency, and reproducibility. Although most studies have been published in a preprint form, which can accelerate data sharing, this may lead to errors or incomplete data. Compared with peer review studies, they may not guarantee the quality of research and scientific acceptability. Among the included studies, only four LLMs were developed specifically for the TCM vertical field, focusing on constructing TCM-specific databases, data fine-tuning based on existing open-source models, and TCM data feedback technology based on expert adjustments. Although these vertically trained models have achieved certain results and performance improvements in generative medical decision-making tasks for TCM, their clinical credibility cannot be guaranteed. The remaining LLMs were trained in the general Chinese medical field, among which four included TCM datasets, and two were named after ancient physicians (Hua Tuo and Bian Que) in honor of their contributions to medicine, indicating the clinical demand for technological research and development in the vertical field of TCM.

The quality of the datasets and model parameters was critical to the performance; however, most of the LLMs in TCM were published based on open-source models with smaller parameters, and only one model reached 33 billion parameters. Universities lead the vast majority of existing LLMs in TCM with computer-related majors and may have received academic support in algorithms. Therefore, TCM professionals, computer technology researchers, and technology companies should collaborate to develop in-depth GAI technologies that conform to the characteristics of TCM and provide more robust, rational, and responsible LLMs. The GAI technology in TCM will compensate for the diagnostic and treatment differences faced by clinical researchers and physicians, change the existing clinical practice model of TCM, and promote TCM standardization.

With the continuous development of AI, performance assessments have become crucial for models[24]. After synthesizing and reviewing the information for all indicators, we found that most LLMs used automatic or manual evaluation metrics to assess model performance. In automatic evaluation metrics, the accuracy of the predicted results is primarily observed as an indicator that directly reflects the performance of the classified data. To summarize the manual evaluation metrics, multiple metrics based on NLP and professional assessments were reported, and the subjective performance assessment of the LLMs in TCM reached an ideal state. After synthesizing the overall predictive results, we found that none of the assessment metrics were adjusted or optimized according to the characteristics of TCM. Therefore, it is necessary to construct new metrics that conform to TCM to judge the results of generative technology and improve the existing assessment metric system standards for TCM GAI technology.

The use of GAI to assist in clinical decision-making in clinical practice could reduce the clinical workload and human error; however, compared with enhancing clinical diagnostics and treatment performance, it is still subject to ethical restrictions[25–27]. Special attention must be paid to user privacy and clinical ethics[28–29]. Publicly available LLMs are restricted to ensure the rationality and morality of the output results in clinical decision-making. TCM characteristics tend to be based on a holistic view of syndrome differentiation and treatment, which captures patient privacy during multiple rounds of dialogue[8,14,30]. The future of TCM clinical decision-making assistance support will require LLMs to fully understand user needs in Q&A while ensuring data privacy and security.

Table 3

Comprehensive statement in challenges, contributions, and limitations

LLMs	Challenges	Contributions	Limitations
TCM-GPT	Lack of domain knowledge, unique objectives, computational efficiency, and effectiveness in TCM domains	1. TCMDA, a new domain adaptation method for TCM, was proposed. 2. The TCM-specific corpus (TCM-Corpus-1B) was constructed. 3. Domain adaptation was performed using the LoRA. 4. Multi-level evaluation was conducted to verify the model performance improvement. 5. The importance of domain-specific pre-training was emphasized	NR
QiBo	The essential difference between TCM theories and modern medicine, and the lack of specialized corpus resources.	1. "Qibo" LLM was constructed for TCM. 2. The model implemented pre-training by SFT with excellent performance. 3. The performance outperforms other open-source models with more parameters. 4. A high-quality training corpus for the TCM domain was constructed. 5. Created the Qibo-benchmark to evaluate and standardize TCM model performance.	1. Not guarantee all responses were accurate. 2. Advised to treat the information generated with caution and consult a professional for users. 3. Unable to handle complex multi-modal medical tasks (medical images and patient physiological signals).
MedChatZH	Lack strong generalization in traditional Chinese medicine consultations	1. Proposed MedChatZH, a generative AI-based dialogue system for inquiries, which performed well in TCM dialogs. 2. Formed a pre-trained corpus and constructed a high-quality dataset combining general and medical dialogs. 3. MedChatZH outperformed other baseline models on several evaluation metrics. 4. MedChatZH combined technological innovation with ethical responsibility in medical practice, to improve the safety and compliance of model applications, with the recommendation to use model outputs under the guidance of a qualified physician.	1. Reliance on translated texts may potentially affect the quality of the model output. 2. There were subtle cultural or contextual mismatches. 3. The inability to effectively conduct dialogs in languages other than Chinese, limited the global applicability.
CPMI-ChatGLM	Lack of emphasis on generating detailed usage instructions for Chinese patent medicine	1. CPMI-ChatGLM was the first domain-specific LLM in CPMI. 2. Use a high-quality CPMI dataset to perform instruct-tuning of foundation model. 3. Constructing and publicly releasing the first CPMI dataset, and use it as a valuable resource for TCM.	1. Relatively small parameter sizes and data sizes may lead to errors (the inclusion of English characters in the generated Chinese text). 2. Improving model performance that may be affected by data diversity.
Ming-MOE	Due to the inherent complexity and diversity of medical tasks, limiting the annotation of specific tasks during inference in real-world applications.	1. A MOE-based medical large-scale language model (MING-MOE) was proposed, which had reached the industry leading level in medical multi-task learning and was the first MOE model capable of handling diverse medical tasks. 2. Performance comparisons were made with other models, and the experimental results emphasized the superior performance of the MING-MOE model on medical tasks.	NR
HuatuoGPT	ChatGPT performed poorly in the medical field, especially in Chinese. Due to ethical and safety concerns, ChatGPT refused to diagnose and prescribe medication. ChatGPT could not function like a doctor. ChatGPT would generate hallucinations due to its auto-regressive nature.	1. HuatuoGPT was the first medical LLM to enable Reinforcement Learning with Augmented Inverse Framework based medical LLMs to use both real and distilled data. 2. It underwent the first systematic evaluation of the medical language model. 3. Manual evaluation showed that HuatuoGPT outperformed existing open-source LLMs and ChatGPT (GPT-3.5-turbo), with its performance being most similar to that of doctors.	Misleading information generated by generative technologies could create serious risks and ethical issues in the biomedical field.

(*Continued*)

Table 3
(Continued)

LLMs	Challenges	Contributions	Limitations
IvyGPT	Small parameters and the lack of high-quality data conforming to real doctor–patient scenarios, restrict the generalization abilities of LLM.	1. IvyGPT consisted of three components: supervised training, rewarding models, and reinforcement learning, which allowed large models with 33 billion parameters to be trained on devices with low computational power. 2. Providing a high-quality dataset containing validated real-life doctor–patient dialog scenarios. 3. Comparing and evaluating IvyGPT with other LLMs in the medical domain.	1. Difficulties in accessing medical data limited the breadth and depth of the model's knowledge. 2. The model's limited ability to personalize and understand context led to inadequate consideration of individual patients. 3. Issues in ethics and accountability needed to be seriously addressed.
Zhongjing-LLaMa	Lack of the ability to initiate questioning and multiple rounds of comprehension like a doctor and was unable to align answers with the intent of the specialist.	1. A proposed novel medical LLMs (Zhongjing-LLaMa) with process-oriented training from pre-training, SFT to RLHF. 2. A multi-round medical database dataset (CMtMedQA) was constructed based on many cases from several medical sectors, including physician consultations. 3. Annotation rules and evaluation criteria were developed covering three dimensions and nine different capabilities. 4. It was demonstrated through multiple experiments that the Zhongjing outperformed previous top Chinese medical models on all dimensions and matched ChatGPT in specific domains.	Since inaccurate data could not guarantee the accuracy of all responses, users needed to seek help from medical experts when using the generated information.
HuoTuo (BenTsao)	LLMs were not optimally implemented in the biomedical tasks because medical expertise was required with response.	1. The first open-source Chinese biomedical LLM that used knowledge-based instruction data. 2. Ensures model accuracy by integrating structured and unstructured medical knowledge to construct domain-specific knowledge. 3. Security, usability, and smoothness were considered, and SUS, a new metric for evaluating LLMs, was proposed.	The accuracy of LLM-generated responses could not be guaranteed and should not have been considered a substitute for professional medical advice.
Bianque	The limited information in a single-round response from users resulted in insufficient personalization and targeting of generation.	1. A healthy LLM with balanced questioning and hinting capabilities was proposed. 2. A fine-tuned large-scale multi-round health dialog datasets (BianQue Corpus) included balanced proportional questions and hints. 3. Empirical results demonstrated excellent ability in multi-round questioning.	1. There was no guarantee of the accuracy for text generated by LLMs, and need to enhance mechanisms for checking and correcting errors in generated health advice. 2. When LLM had the ability to initiate questions, it might have asked about privacy-related issues and risks.

LLMs: Large language models; LORA: Low-rank adaptation; NR: Not report; RLHF: Reinforcement Learning from Human Feedback; SFT: Supervised Fine-Tuning; SUS: Security, Usability, and Smoothness; TCM: Traditional Chinese medicine.

Although recognized medical applications of LLMs are developing in TCM, there are still potential obstacles[3,12]. In clinical practice, interactive communication between doctors and patients cannot be replaced by AI technology[31]. The quality and credibility of the results from generative technology would directly affect the degree of acceptance of the recommendations generated by users. GAI technology, represented by LLMs, must address interpretability, a nontransparent decision-making problem caused by the "black box" of AI models[32–33]. Public disclosure of algorithms and databases can only illustrate the structure of the model and the source of data but cannot effectively ensure that the output results cannot be challenged by evidence-based research[1,34]. When LLMs trained on massive datasets perform tasks, the rationality of the output results is as important as that of the model[1,4]. Complementary exploration of LLMs is necessary to be compatible with user acceptance and feedback systems, forming a cooperative approach between humans and AI for clinical diagnostics and treatment assistance decision-making[35]. In the field of TCM, using LLMs can effectively address existing issues, such as evidence transformation and data processing, and intelligent technologies can aid in the development of TCM clinical decision-making[6,36]. As there is no evaluation system for AI in TCM, we need to develop a system for evaluating GAI technology that meets TCM's characteristics of TCM. A standardized AI evaluation framework for TCM LLMs is required to ensure their reliability, stability, and transparency.

During the development of this systematic review, we did not conduct a quality assessment of the included studies, primarily because of significant differences in the evaluation standards for generative models, such

as LLMs. In the future, a set of general methodological quality assessment standards for AI clinical decision support for generative technology can be established, thereby facilitating the standardized design and transparent implementation of LLMs.

This study has several limitations in its review of LLMs in TCM, mainly focusing on the following: Due to the limitations of the design and data of the included studies, we were unable to explore the overall combined results of LLMs in TCM, mainly reflected in those models published in non-article forms and led to the results of this study being based on partial studies due to a lack of comprehensiveness. Most of the studies included in this review were published in preprint form, and the lack of rigorous peer review may reduce confidence in our research conclusions. Despite pre-training with existing massive data to learn features, they still show insufficient capability when faced with complex TCM diagnostic characteristics or rare diseases, primarily because of the lack of corresponding high-quality datasets and real clinical information. Although all studies discussed the challenges and contributions solved by LLMs and unanimously emphasized the importance of the intelligent medical field, they still had deficiencies in TCM diagnosis and treatment compared with human performance. Although LLMs perform better on specific tasks, they cannot replace the knowledge obtained through clinical diagnostic experience. In addition, the composition of the clinical data and language imbalances may have affected the quality of the datasets. LLMs in TCM are subject to regional restrictions that further affect their global sharing and openness. In this review, all included studies were developed for clinical decision support, and none were used to explore the development of GAI for basic TCM research on mechanisms and multimodal development.

Conclusion and perspective

LLMs have great potential in TCM, especially in physician-patient communication and NLP tasks, which could accelerate the integration of TCM-specific concepts with existing medical models. Despite their satisfactory performance and human verification, the development of LLM for TCM faces challenges in terms of clinical acceptance and the impact of reasonable explanations. LLMs in TCM require high-quality, specialized knowledge to enhance the comprehension capabilities of the models. Therefore, the rationality and accuracy of the generated results should be approached with caution in clinical practice. Establishing representative data in the TCM vertical field and refining the GAI system and framework with TCM characteristics to achieve interactions, open-source code, weights, and data are needed.

Conflict of interest statement

The authors declare no conflict of interest.

Funding

This work was supported by the National Multidisciplinary Innovation Team of Traditional Chinese Medicine (ZYYCXTD-D-202204), China Postdoctoral Science Foundation (2023M742627), Postdoctoral Fellowship Program of CPSF (GZC20231928), Foundation of State Key Laboratory of Component-based Chinese Medicine (CBCM2023201).

Author contributions

Zhe Chen and Junhua Zhang conceived the manuscript. Zhe Chen drafted the manuscript, analyzed the data, and interpreted the review. Zhe Chen and Hui Wang filtered the articles and performed data extraction. Chunxiang Liu and Chengxian Li summarized the models and performed the assessment. Fengwen Yang, Dong Zhang, Alice Josephine Fauci, and Junhua Zhang provided critical version of the manuscript. All authors contributed to the revision of the manuscript and approved the final manuscript.

Ethical approval of studies and informed consent

Not applicable.

Acknowledgments

None.

Data availability

All data can be available from the corresponding author with a reasonable request.

References

[1] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023;29(8):1930–1940.
[2] Zhu L, Mou W, Lai Y, et al. Language and cultural bias in AI: comparing the performance of large language models developed in different countries on traditional Chinese medicine highlights the need for localized models. *J Transl Med* 2024;22(1):319.
[3] Wang Z, Li K, Ren Q, Yao K, Zhu Y. Traditional Chinese medicine formula classification using large language models. Paper presented at: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 5–8 Dec. 2023. 4647–4654.
[4] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023;620(7972):172–180.
[5] Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature* 2023;616(7956):259–265.
[6] Tan Y, Zhang Z, Li M, et al. MedChatZH: a tuning LLM for traditional Chinese medicine consultations. *Comput Biol Med* 2024;172:108290.
[7] Cyranoski D. Why Chinese medicine is heading for clinics around the world. *Nature* 2018;561(7724):448–450.
[8] Zhang HY, Wang X, Meng ZP, et al. Qibo: a large language model for traditional Chinese medicine. *arXiv preprint arXiv:240316056* 2024.
[9] Liu C, Sun K, Zhou Q, et al. CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. *Sci Rep* 2024;14(1):6403.
[10] Mondal H, Komarraju S, Sathyanath D, et al. Assessing the capability of large language models in naturopathy consultation. *Cureus* 2024;16(5):e59457.
[11] Li YZ, Huang SH, Qi JX, et al. Exploring the comprehension of ChatGPT in traditional Chinese medicine knowledge. *arXiv preprint arXiv:240309164* 2024.
[12] Zhang Y, Hao Y. Traditional Chinese medicine knowledge graph construction based on large language models. *Periodical* 2024;13(7):1395.
[13] Yu P, Song K, He F, et al. TCMD: a traditional Chinese medicine QA dataset for evaluating large language models. *arXiv preprint arXiv:240604941* 2024.

[14] Chen Z, Zhang D, Liu C, et al. Traditional Chinese medicine diagnostic prediction model for holistic syndrome differentiation based on deep learning. *Integr Med Res* 2024;13(1):101019.
[15] Yue W, Wang X, Zhu W, et al. TCMBench: a comprehensive benchmark for evaluating large language models in traditional Chinese medicine. *arXiv preprint arXiv:240601126* 2024.
[16] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
[17] Yang G, Liu X, Shi J, et al. TCM-GPT: efficient pre-training of large language models for domain adaptation in traditional Chinese medicine. *Comput Methods Programs Biomed* 2024;6:100158.
[18] Liao Y, Jiang S, Wang Y, et al. MING-MOE: enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts. *arXiv preprint arXiv:240409027* 2024.
[19] Zhang H, Chen J, Jiang F, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:230515075* 2023.
[20] Wang R, Duan Y, Lam C, et al. Ivygpt: interactive Chinese pathway language model in medical domain. *arXiv preprint arXiv:2307.10512* 2023.
[21] Yang S, Zhao H, Zhu S, et al. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *arXiv preprint2023:arXiv:2308.03549* 2023.
[22] Wang H, Liu C, Xi N, et al. Tuning llama model with Chinese medical knowledge. *arXiv preprint 2023;arXiv:230406975*.
[23] Chen Y, Wang Z, Xing X, et al. Bianque: balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint 2023;arXiv:231015896*.
[24] Aydın Temel F, Cagcag Yolcu O, Turan NG. Artificial intelligence and machine learning approaches in composting process: a review. *Bioresour Technol* 2023;370:128539.
[25] Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI copilot for human pathology. *Nature* 2024;634(8033):466–473.
[26] Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Netw Open* 2024;7(3):e240357.
[27] Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. *Int J Surg* 2024;110(6):3701–3706.
[28] Martin KD, Zimmermann J. Artificial intelligence and its implications for data privacy. *Curr Opin Psychol* 2024;58:101829.
[29] Price WN 2nd, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019;25(1):37–43.
[30] Wang Y, Shi X, Li L, et al. The impact of artificial intelligence on traditional Chinese medicine. *Am J Chin Med* 2021;49(6):1297–1314.
[31] Sauerbrei A, Kerasidou A, Lucivero F, et al. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Med Inform Decis Mak* 2023;23(1):73.
[32] Garrett BL, Rudin C. Interpretable algorithmic forensics. *Proc Natl Acad Sci USA* 2023;120(41):e2301842120.
[33] Chakraborty C, Bhattacharya M, Islam MA, et al. ChatGPT indicates the path and initiates the research to open up the black box of artificial intelligence. *Int J Surg* 2023;109(12):4367–4368.
[34] Liu X, Gong T. Artificial intelligence and evidence-based research will promote the development of traditional medicine. *Acupunct Herb Med* 2024;4(1):134–135.
[35] Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. *NPJ Digital Med* 2023;6(1):210.
[36] Bao YF, Ding HK, Zhang ZH, et al. Intelligent acupuncture: data-driven revolution of traditional Chinese medicine. *Acupunct Herb Med* 2023;3(4):271–284.