AI Medicine
https://www.sciltp.com/journals/aim

Article

Large Language Models in Traditional Chinese Medicine: A Short Survey and Outlook

Ruyi Zhang *, Dingcheng Tian and Yu Wang

College of Medicine and Biological Information Engineering, Northeastern University, Chuangxin Road, Shenyang 110016, China
* Correspondence: 2390160@stu.neu.edu.cn

How To Cite: Zhang, R.; Tian, D.; Wang, Y. Large Language Models in Traditional Chinese Medicine: A Short Survey and Outlook. *AI Medicine* **2025**, *2*(1), 3. https://doi.org/10.10.53941/aim.2025.100003.

Received: 5 May 2025
Revised: 12 June 2025
Accepted: 13 June 2025
Published: 24 June 2025

Abstract: Traditional Chinese medicine (TCM) not only embodies a wealth of medical knowledge and cultural value but also exerts a profound influence on the global health domain. However, due to the antiquity, complexity, and uniqueness of its content, traditional computer-aided diagnosis methods struggle to harness its potential fully. In recent years, with the rapid development of artificial intelligence technology, large language models (LLMs) have gradually been applied to the field of TCM due to their outstanding natural language processing capabilities. This paper investigates the current applications of LLMs in TCM, covering various aspects such as clinical assistance, education assessment, and knowledge mining. We first review the construction of datasets related to TCM, including corpus data from TCM literature, medical case records, and clinical data. Next, we explore the specific applications of LLMs in integrating TCM knowledge, supporting clinical decision-making, and assessing education, analyzing their performance in improving efficiency and optimizing resources. Finally, we discuss the challenges faced by LLMs in TCM applications, such as data scarcity, model adaptability, and standardization of evaluation, and provide an outlook on future research and development prospects.

Keywords: large language model; traditional Chinese medicine; clinical Chinese medicine; knowledge graph

1. Introduction

With the rapid development of artificial intelligence technology, large language models (LLMs), as a natural language processing technology based on deep learning, have demonstrated remarkable application potential in various fields in recent years [1]. By training on vast amounts of text data, LLMs are capable of comprehensively understanding, generating, and processing human language, significantly advancing the automation of tasks such as information processing and knowledge discovery [2]. However, in the field of traditional Chinese medicine (TCM), a medical discipline with a long history and a unique theoretical system, the application of LLMs remains in its exploratory stage. As a discipline with thousands of years of practical experience, TCM's knowledge system is primarily embedded in textual sources, such as ancient medical texts, case records, and modern clinical data [3]. The complexity of these textual materials, the archaic nature of the language, and the distinctiveness of the theoretical framework [4] pose significant challenges for traditional data processing methods in effectively extracting and utilizing the valuable knowledge they contain. Consequently, the introduction of LLMs into the field of TCM not only accelerates the digitization and structuring of TCM knowledge but also provides robust technical support for modern research, clinical practice, and educational training in TCM.

The core advantage of LLMs lies in their powerful semantic modeling and reasoning capabilities, enabling them to extract key information from unstructured data and transform it into computable structured representations [5]. In the context of TCM, LLMs can analyze ancient medical texts and case records to identify symptom descriptions, syndrome patterns, and treatment plans. By integrating TCM theories such as yin-yang and the five elements [6], as well as syndrome differentiation and treatment, LLMs can generate auxiliary diagnostic suggestions. This capability is particularly crucial for handling the polysemy and diversity inherent in TCM texts. Moreover, TCM diagnosis and treatment heavily rely on the experience and subjective judgment of physicians [7]. The semantic understanding and knowledge integration capabilities of LLMs can systematize this experience, thereby enhancing the efficiency and consistency of clinical decision-making. More importantly, LLMs enable the integration of dispersed knowledge resources and the construction of a unified digital platform. This not only facilitates the large-scale inheritance of TCM but also promotes research on the integration of Chinese and Western medicine through comparative analysis with modern medical data. As a result, LLMs provide new perspectives and tools for smart healthcare and the globalization of TCM.

This paper aims to comprehensively explore the current application status, challenges, and future development directions of LLMs in the field of TCM. The rest of the paper is structured as follows: Section 2 introduces the importance of TCM datasets in the application of large language models and provides statistics on the public datasets collected from the reviewed papers. Section 3 elaborates on the specific applications of LLMs in the field of TCM, including clinical decision support, educational assessment and assistance, and knowledge integration, and provides statistics on the actual effects and potential of these applications. Section 4 delves into the technical challenges faced by LLMs in TCM applications, such as data scarcity, model adaptability, and the standardization of evaluation systems, and proposes corresponding solutions. Section 5 concludes the paper.

2. Materials and Methods

2.1. Detailed Design and Implementation of LLMs in TCM Clinical Decision Support

The application of LLMs in TCM involves a multi-step process designed to leverage the unique characteristics of TCM data and theory. This process encompasses data collection, preprocessing, model pre-training, fine-tuning, deployment, and evaluation.

Figure 1 illustrates the workflow of LLMs in TCM. The application of LLMs in TCM for clinical decision support involves a streamlined process starting with data collection from diverse sources like ancient texts, clinical records, and structured datasets, which often require expert translation and handling of unstructured formats. This data is then preprocessed through cleaning, tokenization, and annotation to identify key entities and relationships using tailored natural language processing (NLP) techniques. General-purpose LLMs are selected and pre-trained on TCM-specific corpora to adapt to domain concepts such as Yin-Yang and syndrome differentiation, followed by fine-tuning with task-specific datasets and efficient methods like LoRA to optimize performance for applications like diagnosis or prescription generation. The models are deployed on local or cloud platforms, integrated into clinical or educational tools for real-time use, and evaluated using task-specific metrics like accuracy and F1-score, validated against expert annotations or clinical outcomes to ensure reliability and safety.

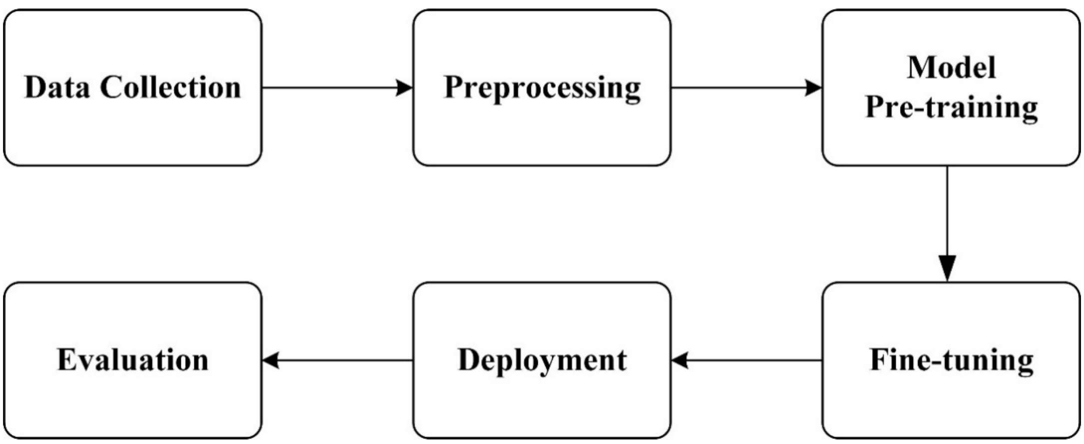

Figure 1. The workflow of LLMs in TCM.

2.2. Dataset

In the development of artificial intelligence, particularly large language models, the quality and scale of datasets are critical determinants of model performance [8]. In the field of traditional Chinese medicine, the unique theoretical system and thousands of years of practical experience underscore the heightened importance of high-quality datasets. TCM knowledge is derived from diverse sources; however, the collection and organization of these data present significant challenges. First, ancient TCM texts are written in classical Chinese, which is obscure and difficult to understand, requiring translation and annotation by experts to convert them into machine-readable structured data. Second, traditional medical case records are often documented in handwritten or oral forms, lacking standardized formats, which makes data extraction and digitization time-consuming and labor-intensive. While modern clinical data are structured to some extent, issues related to patient privacy and ethical concerns limit their widespread availability and sharing. Furthermore, TCM diagnosis and treatment heavily rely on the experience and subjective judgment of physicians, leading to variations in syndrome differentiation and treatment approaches among practitioners. This variability results in inconsistent data annotation standards, further complicating dataset construction. Additionally, TCM data exhibit significant class imbalances. For instance, medical cases and prescriptions for common conditions, such as colds or spleen-stomach disorders, are relatively abundant, whereas records of rare or complex diseases are exceedingly scarce. Such imbalances may cause models to overfit to common cases during training, thereby reducing their generalization ability for less frequent scenarios. These challenges render the construction of TCM datasets a complex, systematic endeavor, necessitating coordinated support from both technological and resource perspectives.

Despite the significant challenges in constructing TCM datasets, the emergence of publicly available datasets has injected new vitality into this field, markedly advancing the development of artificial intelligence in TCM. Public datasets not only provide researchers with valuable training resources but also facilitate performance comparisons among different models and the validation of academic findings. For instance, the TCMPT Dataset [9] integrates 1522 ancient TCM texts, modern textbooks, and clinical data, comprising a total of 160 million tokens. This dataset covers multiple aspects of TCM theory and practice, offering a comprehensive knowledge base for model training. The CMExam Dataset [10] focuses on TCM disease classification and clinical department information, supporting model applications in tasks such as knowledge-based question answering and diagnostic reasoning. Additionally, the MedChatZH Dataset [11] includes preprocessed corpora from over 1000 TCM books and 764,000 medical dialogue instruction data, specifically designed for scenarios such as TCM consultation and medication recommendations. The open sharing of these datasets has not only enhanced the training efficiency and predictive accuracy of models but also laid a foundation for standardized research in TCM artificial intelligence. Table 1 shows the public TCM datasets. Table 2 shows examples of the datasets.

Table 1. Traditional Chinese Medicine Datasets.

Dataset	Remarks	Website
TCMPT Dataset [9]	Contains 160 million tokens, including 1522 ancient Chinese medicine books, textbooks, case records, pharmacopeias, package inserts, and public clinical data, integrating theoretical and practical TCM knowledge.	https://github.com/TCMAI-BJTU/LingdanLLM (accessed on 23 June 2025)
TCPMD Dataset [9]	Includes 200,000 single-turn dialogues, generating standardized Q&A about TCPM indications, ingredients, and contraindications based on structured data from pharmacopeias and package inserts. Also includes 1599 multi-turn dialogues designed with a TCM clinical diagnostic inquiry logic, covering symptom follow-ups, syndrome diagnosis, and safety consultations for medication recommendations.	
SSHPR Dataset [9]	Consists of 39,122 outpatient records with patient chief complaints, illness histories, and corresponding herbal prescriptions, focusing on spleen and stomach disease diagnosis and treatment scenarios.	
TCM-QA-datasets [12]	A Chinese medicine-focused resource designed for knowledge and reasoning tasks includes three question types: 574 single-choice questions, 131 multiple-choice questions, and 97 true or false questions.	https://github.com/yizhen-buaa/TCM-QA-datasets (accessed on 23 June 2025)
CMExam Dataset [10]	Partially covers disease classifications related to traditional Chinese medicine, such as traditional medicine conditions in ICD-11, clinical departments of traditional Chinese medicine, and medical disciplines including traditional Chinese medicine and traditional Chinese pharmacy.	https://github.com/williamliujl/CMExam (accessed on 23 June 2025)
EMPEC [13]	Includes specialized exam questions for Traditional Chinese Medicine Practitioners, covering subjects like basic and clinical Chinese medicine, to evaluate large language models' knowledge in TCM and their cross-professional healthcare capabilities.	https://github.com/zhehengluoK/eval_empec (accessed on 23 June 2025)
MedChatZH dataset [11]	Includes preprocessed corpus from over 1000 traditional Chinese medicine books and 764,000 cleaned medical dialogue instruction data, designed to train dialogue models for improved performance in TCM consultations, symptom analysis, and medication recommendations.	https://huggingface.co/datasets/tyang816/MedChatZH (accessed on 23 June 2025)

Table 1. *Cont.*

Dataset	Remarks	Website
CMB Dataset [14]	Includes multiple-choice questions from TCM-related qualification exams and clinical diagnostic questions based on real cases, derived from public medical exams, textbooks, and clinical teaching materials, covering basic theories of Traditional Chinese Medicine, formula classification, pattern differentiation and treatment, and other content.	https://github.com/FreedomIntelligence/CMB (accessed on 23 June 2025)
ShennongAlpha [15]	Contains knowledge of over 14,000 Natural Medicinal Materials, sourced from authoritative references such as the Chinese Pharmacopoeia and standardized into a structured knowledge base. It includes detailed information such as Systematic Nomenclature, chemical components, flavor/nature/meridian tropism, functions and indications, applicable diseases, and pharmacological effects.	https://shennongalpha.westlake.edu.cn (accessed on 23 June 2025)
HDI Dataset [16]	Includes 7591 herbal medicine-drug interaction (HDI) records after cleaning. It involves 1424 drugs, 681 herbal medicines, and 43,517 natural products. The content includes drug SMILES sequences, herbal medicine names, and their corresponding natural product information, used for constructing models to predict herbal medicine-drug interactions.	https://github.com/sisyyuan/HDI (accessed on 23 June 2025)
HBot Dataset [17]	This dataset is a document-level entity and relation annotation dataset in the Traditional Chinese Medicine (TCM) domain, containing 270 documents, 5996 entities, and 4685 relations. It covers entities such as TCM acupoints, diseases, symptoms, treatment methods, and their interrelationships.	https://gitee.com/plabrolin/interactive-3d-acup (accessed on 23 June 2025)
CMeEE dataset [18]	Includes 15,000 training samples, 5000 validation samples, and 3000 test samples. The content covers 9 types of medical entities, including body composition, disease, medical examination item, medical department, symptom, medical equipment, medical procedure, drug, and microorganism.	https://github.com/bucm-tcm-tool/NewTuning/ (accessed on 23 June 2025)

2.3. Metrics

P@K (Precision at K): This metric measures the proportion of correctly recommended items (e.g., herbs) among the top K recommendations. In TCM prescription tasks, it indicates how accurately the model suggests relevant herbs.

R@K (Recall at K): This measures the proportion of relevant items that are included in the top K recommendations. It assesses the model's ability to retrieve all relevant herbs for a given condition.

F@K (F1-score at K): The harmonic means of P@K and R@K, providing a balanced measure of precision and recall.

BLEU (Bilingual Evaluation Understudy): Used in text generation tasks (e.g., consultation dialogues), BLEU measures the similarity between generated text and reference text based on n-gram overlap.

ROUGE (Recall-Oriented Understudy for Gist Evaluation): Also, for text generation, ROUGE focuses on recall, measuring how much of the reference text is captured in the generated text.

Faithfulness, Answer Relevance, Context Recall: These metrics are used in question-answering systems to evaluate the accuracy, relevance, and completeness of the model's responses.

Metrics like P@K and R@K are crucial as they directly relate to the clinical utility of the recommendations. High precision ensures that suggested herbs are relevant, while high recall ensures that important herbs are not missed. BLEU and ROUGE are standard for evaluating the quality of generated responses in consultation scenarios. However, in TCM, additional metrics like clinical accuracy or user satisfaction may be necessary to fully assess performance. For tasks like knowledge graph construction, metrics such as F1-score for entity and relation extraction are essential to ensure the accuracy of the structured knowledge. These metrics provide a quantitative basis for evaluating LLMs in TCM but should be complemented with qualitative assessments by TCM experts to ensure clinical relevance and safety.

3. Current Applications in Traditional Chinese Medicine

3.1. Large Language Models as Clinical Decision Support Tools

Large Language Models provide intelligent support for the diagnosis and treatment of Traditional Chinese Medicine by processing and analyzing complex clinical data, such as symptom descriptions, pulse records, and medical history texts. Their core functions include semantic understanding, pattern recognition, and reasoning generation, enabling the transformation of unstructured patient information into structured, actionable diagnostic suggestions or treatment plans. For instance, LLMs can parse patients' chief complaints and pulse descriptions, integrate TCM theories with modern natural language processing techniques, such as syndrome differentiation and treatment, identify potential syndrome patterns, and generate corresponding treatment ideas. This technological integration not only significantly enhances the efficiency and accuracy of clinical decision-making but also provides new possibilities for the design of personalized treatment plans. By analyzing patients' constitutional characteristics, medical history, and lifestyle habits, LLMs can output diagnostic and treatment suggestions that better meet individual needs. At the same time, they can extract patterns from ancient TCM texts and modern clinical data, assisting doctors in quickly identifying key information in complex cases and avoiding subjective judgments based solely on experience.

In recent years, several research works have explored the application of LLMs in clinical decision-making in TCM, focusing on technological and methodological innovations. Hua et al. [9] developed the Lingdan system based on the Baichuan2-13B-Base model, utilizing datasets from ancient TCM texts, textbooks, and clinical cases for pre-training and fine-tuning to optimize the model's performance in clinical reasoning and prescription recommendations. Yang et al. [19] proposed the TCM domain adaptation method, which involves constructing a TCM domain corpus and performing domain-adaptive pre-training on the BLOOM-7B model to enhance its adaptability in TCM examination and diagnostic tasks. Zhu et al. [20] developed the model named Zhongjing based on LLaMA2, fine-tuned it with a TCM question-answering dataset to enable operation on local devices, and designed the TCMEval evaluation method to assess the model's professionalism. Tan et al.'s [11] MedChatZH system, based on the Baichuan-7B model, optimized the generation of TCM consultation dialogues through pre-training on TCM literature and fine-tuning with dialogue data. Table 2 summarizes the applications of large language models in clinical decision support for TCM.

Table 2. Large Language Models in Clinical Decision Support for Traditional Chinese Medicine.

Paper	Object	Foundation Model	Result
[9]	Clinical reasoning and prescription recommendation in TCM	Baichuan2-13B-Base	P@20 = 63.52%, R@20 = 56.08%, F@20 = 58.83%,
[21]	Herbal prescription recommendation in TCM	ChatGPT (text-embedding-ada-002), GCN	P@20 = 13.0%, R@20 = 32.5%
[22]	TCM interactive consultation	GPT-3.5, and GPT-4.0	Achieved high humanistic care scores
[23]	TCM diagnosis and treatment for rheumatoid arthritis	LLaMA-7B	ACC = 54%
[11]	TCM consultation dialogue	Baichuan-7B	BLEU-1 = 56.14, ROUGE-1 = 35.99
[24]	Mitigate bias against TCM in LLMs	Llama model integrated with RAG	Confidence scores = 0.68, Bias Detection Rate = 31%
[25]	Predict TCM herbal prescriptions	Chinese-Alpaca-Plus-7B	F1-score = 80.31%, NMSE = 0.0604
[26]	Question-answering in TCM epidemic prevention and treatment	ChatGLM	BLEU-4 = 44.02, ROUGE-L = 61.10, METEOR = 59.40
[27]	Addressing challenges like theoretical differences from modern medicine and a scarce specialized corpus	Chinese LLaMA-7B/13B	Average Subjective Win Rate = 63%, Objective Accuracy = 58%
[20]	Develop a locally deployable LLM for TCM	LLaMA2	Rouge-1 = 61.31, and BLEU = 22.07
[19]	Propose a domain-specific adaptation approach, TCM Domain Adaptation (TCMDA)	BLOOM-7B	Accuracy Improvements = 17.4% (TCM examination), 12.3% (TCM diagnosis)

3.2. Large Language Models as Tools for Educational Assessment and Assistance

Large Language Models can serve as an auxiliary tool in Traditional Chinese Medicine education by analyzing learners' knowledge mastery, providing personalized question-and-answer support, and simulating practice scenarios to assist learners in their studies. Their primary functions include parsing specialized TCM questions, generating targeted learning recommendations, and assessing learners' comprehension and application abilities. For instance, LLMs can identify deficiencies in learners' TCM knowledge based on their performance in answering questions and provide corresponding suggestions for improvement. This technology can, to some extent, alleviate issues in traditional TCM education, such as insufficient feedback or limited resources, thereby helping learners acquire knowledge more efficiently.

Yue et al. [28] constructed a comprehensive benchmark dataset named TCMBench, which collects questions from the TCM Practitioner Qualification Examination, covering TCM basic theory and clinical practice. Using this dataset, they tested the performance of LLMs in the field of TCM knowledge. The results indicated that while LLMs demonstrated certain capabilities in specific tasks, their overall performance still has significant room for improvement. This work laid the foundation for evaluating the applicability of LLMs in TCM education. Peng et al. [29] focused on assessing the performance of GPT-4 and mainstream Chinese LLMs in the TCM graduate entrance examination, such as Ernie Bot and ChatGLM. They found that some models performed well in examination scenarios, highlighting the potential of LLMs in TCM educational assessment. This work suggests that LLMs can serve as auxiliary tools for educational assessment, offering practical value. Li et al. [12] created the TCM-QA dataset, which includes single-choice, multiple-choice, and true/false questions. Based on this dataset, they tested the capabilities of GPT-3.5-turbo. The study revealed that the model performed strongly in true/false questions but was weaker in tasks requiring complex reasoning, such as multiple-choice questions. This work uncovered the varying abilities of LLMs in handling different types of TCM questions. Table 3 summarizes the applications of LLMs in educational assessment and assistance in TCM.

Table 3. Large Language Models in Educational Assessment and Assistance for Traditional Chinese Medicine.

Paper	Evaluation Object	Foundation Model	Result
[30]	Chinese Master's Degree Entrance Examination	ChatGPT3.5, and GPT-4	Both models scored above the admission threshold
[29]	Chinese Postgraduate Examination for TCM	4 LLMs	Ernie Bot (146/300) and ChatGLM (136.5/300) exceeded the passing threshold (117)
[31]	Taiwan's TCM licensing examination	3 ChatGPT versions	Only GPT-4o passed with an overall accuracy of 62.29%
[28]	A comprehensive benchmark in TCM	6 LLMs	LLMs showed potential but failed to meet passing thresholds
[12]	A newly constructed TCM-QA dataset	GPT-3.5-turbo	Achieved the highest precision in true/false questions
[14]	A comprehensive Chinese medical benchmark integrating TCM and Western medicine	9 LLMs	HuatuoGPT-II leading medical models (76.82% average)
[13]	A large-scale Chinese benchmark	17 LLMs	GPT-4 led with over 75% accuracy, but struggled in specialized fields (e.g., TCM Practitioners)
[32]	Named entity recognition for 6 entity types across 3 domains	6 LLMs	GPT-4 outperformed in fuzzy matching (e.g., TCM formulas, ingredients)
[10]	A Chinese medical exam	20 LLMs	GPT-4 achieved 61.6% accuracy (vs. 71.6% human)
[18]	A large-scale Chinese prompt-tuning benchmark for medical domain tasks	9 LLMs	GPT-4 led in few-shot settings (overall score 0.518)

3.3. Large Language Models as Integrators of Research Knowledge

Large Language Models play a pivotal role in knowledge integration and theoretical research within the field of Traditional Chinese Medicine by processing vast amounts of TCM literature, ancient texts, and clinical data. Their core functions encompass text mining, semantic embedding, and structured knowledge generation, enabling the extraction of critical information—such as symptoms, pathogenesis, treatment principles, and formulas—from unstructured TCM texts and transforming it into structured, computable representations. This capability not only supports academic research and interdisciplinary exploration but also reveals the intrinsic logic and relationships within TCM knowledge through the construction of knowledge graphs. For instance, LLMs can integrate TCM theories from ancient texts with modern clinical data, providing researchers with systematic analytical tools. This technological application effectively bridges the gap between traditional TCM knowledge and modern computational techniques, propelling the modernization and innovative development of TCM theoretical research.

In recent years, several studies have demonstrated the potential of LLMs in the integration of TCM knowledge and theoretical research, particularly in enhancing knowledge structuring and research efficiency. Duan et al. [33] developed a TCM case-based question-answering system using GLM-3-Turbo and GPT-3.5-Turbo, which automatically extracts entities such as symptoms, pathogenesis, treatment principles, and prescriptions to construct a TCM case knowledge graph and supports natural language query functions. Li et al. [34] utilized ChatGPT 4.0 to extract entities and relationships related to rheumatology from ancient texts, constructing a TCM rheumatology knowledge graph that provides structured data support for theoretical research. Liu et al. [35] enhanced the knowledge retrieval capabilities for TCM ancient texts through GPT-4-turbo, improving the precision of semantic understanding and information extraction. Zhang et al. [36] built a TCM knowledge graph based on the iFLYTEK SPARK large model, achieving precise identification and association of TCM entities and relationships. These studies collectively indicate that LLMs have broad application prospects in the integration of TCM knowledge and theoretical research, providing significant support for the modernization of TCM. Table 4 summarizes the applications of large language models in the integration of TCM knowledge.

Table 4. Large Language Models in the Integration of Traditional Chinese Medicine Knowledge.

Paper	Object	Foundation Model	Result
[33]	Question-answering system for TCM cases by LLMs and knowledge graphs	GLM-3-Turbo, GPT-3.5-Turbo	Faithfulness = 0.9375, Answer Relevance = 0.9686, Context Recall = 0.9500
[34]	A knowledge graph for rheumatology in ancient Chinese medical literature	ChatGPT 4.0	Constructed 42,972 core entities, 149,925 basic expansion entities, and 58,079 relationships in the additional library
[35]	Enhance knowledge retrieval from classical TCM texts	GPT-4-turbo	Answer Relevance = 0.856, Recall = 0.561, and Context Precision = 0.532
[36]	Construct a TCM knowledge graph	iFLYTEK SPARK Large Model	F1-score = 90.36%, Recall = 87.77%, Precision = 93.168%
[37]	Classifying TCM formulas	ChatGLM-6b, ChatGLM2-6b	Accuracy = 71% (ChatGLM2-6b), and 70% (ChatGLM-6b)
[15]	Standardized curation, acquisition, and translation of natural medicinal material knowledge	ShennongChat	Achieved standardized NMM differentiation, accurate knowledge retrieval via context-aware search
[38]	Extracting five types of acupoint location relations from the WHO Standard Acupuncture Point Locations corpus	5 LLMs	Fine-tuned GPT-3.5 outperformed all models with the highest micro-average F1 score (0.92)
[39]	Address knowledge injection and preference alignment challenges in healthcare	Baichuan-7B	Rouge-1 = 27.44, and BLEU-1 = 16.66
[16]	Predict potential interactions between herbal medicines and drugs	ChatGPT (text-embedding-ada-002), GCN	AUC = 98.48%, AUPR = 98.21%, and F1-score = 95.17%

4. Challenges and Research Directions

Despite the widely recognized potential of large language models in the field of traditional Chinese medicine, their practical deployment still faces multiple technical challenges. Data scarcity is the primary bottleneck for the application of large language models in traditional Chinese medicine. The knowledge system of TCM originates from numerous textual contents, which are difficult to use directly for model training due to their unique language style, format heterogeneity, and privacy sensitivity. For instance, ancient texts are mostly written in classical Chinese and contain a large number of metaphorical expressions and specialized terms, such as "pi-xu-shi-kun (spleen deficiency with dampness)" or "yin-xu-huo-wang (yin deficiency with exuberant fire)". These require not only cross-linguistic and cross-temporal semantic analysis but also in-depth annotation by TCM experts to convert them into structured data. This process involves complex NLP tasks [40], including named entity recognition, relationship extraction, and semantic disambiguation [41,42]. However, the accuracy of current automated tools in processing ancient texts remains insufficient.

The adaptability of models in the field of TCM further exacerbates the technical challenges. The theoretical framework of TCM is grounded in concepts such as Yin-Yang and the Five Elements, the Zang-Fu organs and meridians, and syndrome differentiation and treatment [43]. These principles diverge significantly from the biomedical paradigm of modern medicine, imposing unique demands on the semantic modeling and reasoning capabilities of LLMs. General-purpose LLMs, such as those in the GPT series, excel in broad language tasks but encounter a domain gap in semantic understanding when applied to TCM contexts. For instance, the term "Qi stagnation and blood stasis" encompasses not only symptom descriptions but also a holistic assessment of etiology, pathogenesis, and treatment principles. General models, however, often remain limited to superficial text matching, failing to grasp their deeper semantic implications. Moreover, TCM diagnosis relies on multidimensional contextual reasoning, necessitating the integration of diverse factors such as the patient's chief complaints, physical signs, environmental influences, and the physician's clinical experience [7]. Yet, the attention mechanisms of current LLMs prove inadequate in modeling long-sequence dependencies and cross-modal reasoning. Taking the Transformer architecture as an example, while its self-attention mechanism effectively captures local semantic relationships, its performance markedly declines when tasked with inference chains spanning hundreds of steps or navigating nonlinear causal relationships.

Evaluation and standardization issues similarly constrain the technological advancement of LLMs in the field of TCM. Due to the complexity of TCM tasks, general NLP evaluation metrics such as BLEU, ROUGE, and F1 fail to comprehensively measure model performance. For instance, in prescription generation tasks, high textual similarity does not necessarily imply clinical efficacy; validation requires medical indicators such as therapeutic consistency or safety. However, current research suffers from highly fragmented datasets and evaluation methods, with significant variations in task definitions, annotation standards, and data scales across datasets, thereby limiting fair comparisons between models. Moreover, the gold standard for TCM diagnosis is difficult to standardize due to variations in physicians' experience, often necessitating the inclusion of multi-expert consensus as a reference in evaluations, which increases assessment costs.

The application of LLMs in TCM differs from modern medicine due to distinct theoretical foundations, data characteristics, and clinical paradigms: First, TCM relies on holistic concepts like Yin-Yang and syndrome differentiation, whereas modern medicine is rooted in evidence-based, reductionist approaches (e.g., pathophysiology). LLMs in TCM must interpret abstract terms (e.g., "Qi deficiency"), requiring deeper semantic modeling. Second, TCM data are often unstructured, historical, and in classical Chinese, contrasting with modern medicine's structured, standardized records (e.g., electronic health records). This demands more preprocessing and domain adaptation for TCM LLMs. Also, TCM emphasizes individualized, experience-based treatments (e.g., herbal formulas), while modern medicine follows protocol-driven interventions. LLMs in TCM thus need to handle complex, multi-factorial reasoning. These differences highlight the need for TCM-specific LLM strategies, enhancing their ability to bridge traditional and modern medical practices.

Future technological research should focus on the following directions: First, to address the issue of data scarcity, the development of a specialized NLP toolchain for ancient texts can be considered. This includes deep learning-based classical Chinese word segments [44], dependency parsers [45], and semantic parsing modules integrated with TCM knowledge graphs to improve the efficiency and accuracy of text structuring. Second, regarding the model's adaptability issue, existing hybrid architecture models can be utilized to combine the generative capabilities of LLMs with the structured reasoning capabilities of knowledge graphs [46]. This would enhance the model's multi-hop reasoning ability or incorporate reinforcement learning mechanisms [47] to optimize the model's decision-making pathways. Finally, to tackle the issue of evaluation standardization, a TCM-specific evaluation framework should be designed. This involves constructing open-source TCM benchmark datasets, defining unified annotation standards and task protocols, and establishing a dynamic evaluation mechanism that updates assessment criteria in real-time based on clinical feedback.

5. Conclusions

This paper systematically explores the current application status, challenges, and future development directions of large language models in the field of Traditional Chinese Medicine. Through an in-depth analysis of multiple dimensions, including dataset construction, clinical decision support, educational assessment and assistance, and knowledge integration, we reveal the immense potential of LLMs in advancing the modernization of TCM research and practice. However, the widespread application of LLMs in TCM still faces numerous challenges, such as the scarcity of high-quality TCM data, insufficient model adaptability to the unique context of TCM, and the lack of standardized evaluation systems. Future research should focus on these critical areas, enhancing the applicability and reliability of LLMs in TCM through interdisciplinary collaboration and technological breakthroughs. With continuous technological advancements, LLMs are expected to serve as a bridge connecting traditional wisdom with modern technology, injecting new vitality into the inheritance and development of TCM while contributing unique Eastern wisdom to global health initiatives.

Author Contributions

All authors contributed to the study's conception and design. R.Z.: methodology, data collection, writing, visualization, investigation; D.T.: data collection, writing; Y.W.: methodology, writing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Yang, J.; Jin, H.; Tang, R.; et al. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Trans. Knowl. Discov. Data* **2024**, *18*, 1–32.
2. Kumar, P. Large language models (LLMs): Survey, technical frameworks, and future challenges. *Artif. Intell. Rev.*. **2024**, *57*, 260.
3. Zhou, X.; Peng, Y.; Liu, B. Text mining for traditional Chinese medical knowledge discovery: A survey. *J. Biomed. Inform.* **2010**, *43*, 650–660.
4. Scheid, V. *Chinese Medicine in Contemporary China: Plurality and Synthesis*; Duke University Press: Durham, NC, USA, 2002.
5. Zhao, W.X.; Zhou, K.; Li, J.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
6. Cheng, C. The yi-jing and yin-yang way of thinking. In *The Routledge History of Chinese Philosophy*; Routledge: London, UK, 2008; pp. 83–118.
7. Zhang, Y.-H.; Lv, J.; Gao, W.; et al. Practitioners' perspectives on evaluating treatment outcomes in traditional Chinese medicine. *BMC Complement. Altern. Med.* **2017**, *17*, 269.
8. Chang, Y.; Wang, X.; Wang, J.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45.
9. Hua, R.; Dong, X.; Wei, Y.; et al. Lingdan: Enhancing encoding of traditional Chinese medicine knowledge for clinical reasoning tasks with large language models. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2019–2029.
10. Liu, J.; Zhou, P.; Hua, Y.; et al. Benchmarking large language models on CMExam—A comprehensive Chinese medical exam dataset. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 52430–52452.
11. Tan, Y.; Zhang, Z.; Li, M.; et al. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. *Comput. Biol. Med.* **2024**, *172*, 108290.
12. Li, Y.; Huang, S.; Qi, J.; et al. Exploring the comprehension of ChatGPT in traditional Chinese medicine knowledge. *arXiv* **2024**, arXiv:2403.09164.
13. Luo, Z.; Yuan, C.; Xie, Q.; et al. Are Large Language Models True Healthcare Jacks-of-All-Trades? Benchmarking Across Health Professions Beyond Physician Exams. *arXiv* **2024**, arXiv:2406.11328.
14. Wang, X.; Chen, G.H.; Song, D.; et al. CMB: A comprehensive medical benchmark in Chinese. *arXiv* **2023**, arXiv:2308.08833.
15. Yang, Z.; Yin, Y.; Kong, C.; et al. ShennongAlpha: An AI-driven sharing and collaboration platform for intelligent curation, acquisition, and translation of natural medicinal material knowledge. *Cell Discov.* **2025**, *11*, 32.
16. Yuan, S.; Zhou, Z.; Jin, X.; et al. Enhancing Herbal Medicine-Drug Interaction Prediction Using Large Language Models. *IEEE J. Biomed. Health Inform.* **2025**. https://doi.org/10.1109/JBHI.2025.3558667.
17. Zhang, B.; Yi, Z.; Wang, J.; et al. HBot: A Chatbot for Healthcare Applications in Traditional Chinese Medicine Based on Human Body 3D Visualization. *arXiv* **2024**, arXiv:2408.00481.

18. Zhu, W.; Wang, X.; Zheng, H.; et al. PromptCBLUE: A Chinese prompt tuning benchmark for the medical domain. *arXiv* **2023**, arXiv:2310.14151.
19. 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. Update* **2024**, *6*, 100158.
20. Zhu, J.; Gong, Q.; Zhou, C.; et al. ZhongJing: A Locally Deployed Large Language Model for Traditional Chinese Medicine and Corresponding Evaluation Methodology: A Large Language Model for data fine-tuning in the field of Traditional Chinese Medicine, and a new evaluation method called TCMEval is proposed. In Proceedings of the Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science, Chengdu, China, 20–22 October 2023; pp. 1036–1042.
21. Qi, J.; Wang, X.; Yang, T. Traditional Chinese medicine prescription recommendation model based on large language models and graph neural networks. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 5–8 December 2023; pp. 4623–4627.
22. Qi, K.; Chen, L.; Yao, J; et al. Feasibility and challenges of interactive AI for traditional Chinese medicine: An example of ChatGPT. *Chin. Med. Cult.* **2024**, *7*, 174–181.
23. Liu, Y.; Luo, S.; Zhong, Z.; et al. Hengqin-RA-v1: Advanced Large Language Model for Diagnosis and Treatment of Rheumatoid Arthritis with Dataset based Traditional Chinese Medicine. *arXiv* **2025**, arXiv:2501.02471.
24. Su, X.; Gu, Y. Implementing retrieval-augmented generation (rag) for large language models to build confidence in traditional Chinese medicine. 2024. https://doi.org/10.31219/osf.io/ns2v3
25. Zhou, X.; Dong, X.; Li, C.; et al. TCM-FTP: Fine-Tuning Large Language Models for Herbal Prescription Prediction. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; pp. 4092–4097.
26. Zhou, Z.; Yang, T.; Hu, K. Traditional chinese medicine epidemic prevention and treatment question-answering model based on llms. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 5–8 December 2023; pp. 4755–4760.
27. Zhang, H.; Wang, X.; Meng, Z.; et al. Qibo: A large language model for traditional Chinese medicine. *arXiv* **2024**, arXiv:2403.16056.
28. Yue, W.; Wang, X.; Zhu, W.; et al. Tcmbench: A comprehensive benchmark for evaluating large language models in traditional Chinese medicine. *arXiv* **2024**, arXiv:2406.01126.
29. Peng, S.; Zhu, Y.; Wang, B.; et al. Performance of GPT-4 and mainstream Chinese Large Language Models on the Chinese Postgraduate Examination Dataset: Potential for AI-assisted Traditional Chinese Medicine. *Preprints.* **2024**. https://doi.org/10.21203/rs.3.rs-4392855/v1.
30. Li, K.-C.; Bu, Z.-J.; Shahjalal, M.; et al. Performance of ChatGPT on the Chinese master's degree entrance examination in Clinical Medicine. *PLoS ONE* **2024**, *19*, e0301702.
31. Wu, Y.-C.; Wu, Y.-C.; Wu, C.-L.; et al. GPT-3.5, GPT-4, and GPT-4o Performance in the Traditional Chinese Medicine Licensing Examination in Taiwan: The Need for Improved Reliability in AI Solutions. *Preprints.* **2025.** https://doi.org/10.20944/preprints202501.1787.v1.
32. Tong, X.; Smirnova, N.; Upadhyaya, S.; et al. Utilizing Large Language Models for Named Entity Recognition in Traditional Chinese Medicine against COVID-19 Literature: Comparative Study. *arXiv* **2024**, arXiv:2408.13501.
33. Duan, Y.; Zhou, Q.; Li, Y.; et al. Research on a traditional Chinese medicine case-based question-answering system integrating large language models and knowledge graphs. *Front. Med.* **2025**, *11*, 1512329.
34. Li, H.; Xia, C.; Hou, Y.; et al. TCMRD-KG: Innovative design and development of rheumatology knowledge graph in ancient Chinese literature assisted by large language models. *Front. Pharmacol.* **2025**, *16*, 1535596.
35. Liu, C.; Sun, K.; Zhou, Q.; et al. CPMI-ChatGLM: Parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions. *Sci. Rep.* **2024**, *14*, 6403.
36. Zhang, Y.; Hao, Y. Traditional Chinese medicine knowledge graph construction based on large language models. *Electronics* **2024**, *13*, 1395.
37. Wang, Z.; Li, K.; Ren, Q.; et al. Traditional Chinese Medicine Formula Classification Using Large Language Models. In Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 5–8 December 2023; pp. 4647–4654.
38. Li, Y.; Peng, X.; Li, J.; et al. Relation extraction using large language models: A case study on acupuncture point locations. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 2622–2631.
39. Ye, Q.; Liu, J.; Chong, D.; et al. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv* **2023**, arXiv:2310.09089.
40. Chowdhary, K.; Chowdhary, K. Natural language processing. In *Fundamentals of Artificial Intelligence*; Springer: New Delhi, India, 2020; pp. 603–649.
41. Li, J.; Sun, A.; Han, J.; et al. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*

2020, *34*, 50–70.
42. Nasar, Z.; Jaffry, S.W.; Malik, M.K. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.* **2021**, *54*, 1–39.
43. Jiang, M.; Lu, C.; Zhang, C.; et al. Syndrome differentiation in modern research of traditional Chinese medicine. *J. Ethnopharmacol.* **2012**, *140*, 634–642.
44. Ma, J.; Ganchev, K.; Weiss, D. State-of-the-art Chinese word segmentation with Bi-LSTMs. *arXiv* **2018**, arXiv:1808.06511.
45. Zhang, S.; Wang, L.; Sun, K.; et al. A practical Chinese dependency parser based on a large-scale dataset. *arXiv* **2020**, arXiv:2009.00901.
46. Sun, J.; Xu, C.; Tang, L.; et al. Think-on-graph: Deep and responsible reasoning of *a* large language model on *a* knowledge graph. *arXiv* **2023**, arXiv:2307.07697.
47. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285.