A comprehensive review of Intelligent Question-Answering Systems in Traditional Chinese Medicine Based on LLMs

Qilan Xu, Tong Wu, Yiwen Wang, Xingyu Li, Heshui Yu, Shixin Cen, Zheng Li

PII: S2095-1779(25)00223-0

DOI: https://doi.org/10.1016/j.jpha.2025.101406

Reference: JPHA 101406

To appear in: *Journal of Pharmaceutical Analysis*

Received Date: 25 March 2025

Revised Date: 2 July 2025

Accepted Date: 19 July 2025

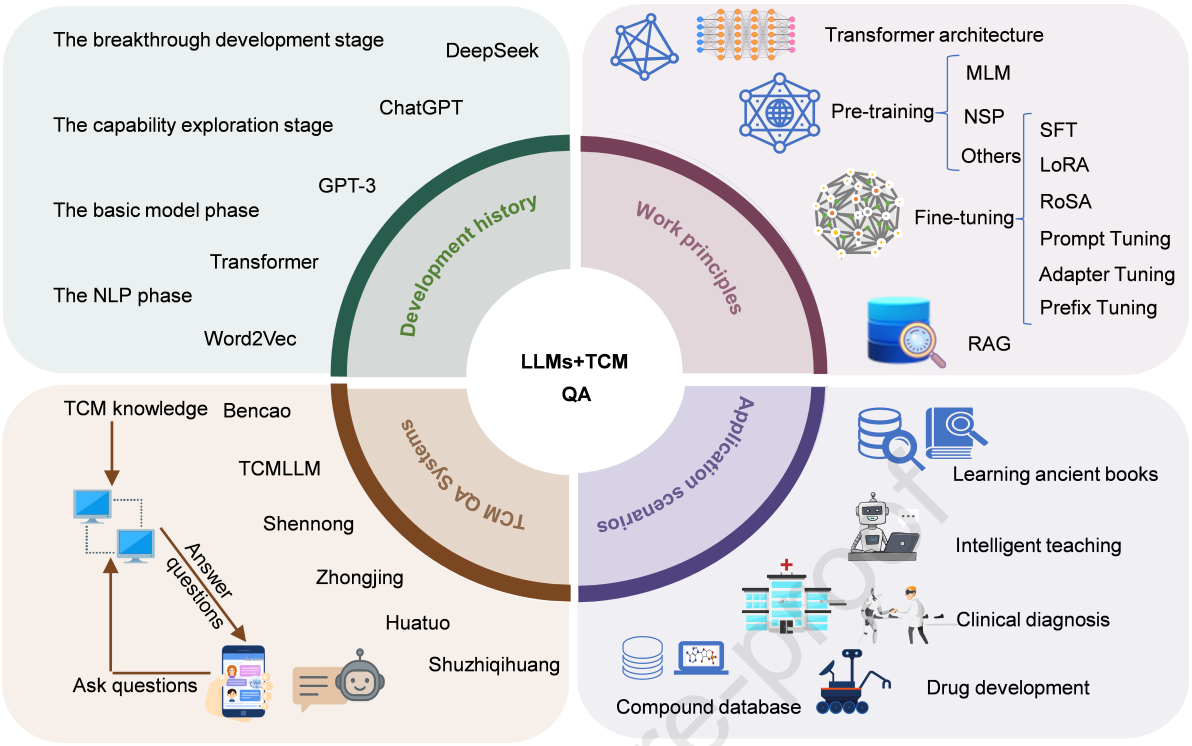
Please cite this article as: Q. Xu, T. Wu, Y. Wang, X. Li, H. Yu, S. Cen, Z. Li, A comprehensive review of Intelligent Question-Answering Systems in Traditional Chinese Medicine Based on LLMs, *Journal of Pharmaceutical Analysis*, https://doi.org/10.1016/j.jpha.2025.101406.

A Comprehensive Review of Intelligent Question-Answering Systems in Traditional Chinese Medicine Based on LLMs

Qilan Xu [a, 1], Tong Wu [a, 1], Yiwen Wang [a], Xingyu Li [a], Heshui Yu [a, b, c, d], Shixin Cen [a, b, c, d, **], Zheng Li [a, b, c, d, e, *]

[a] *College of Pharmaceutical Engineering of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin, 301617, China*
[b] *State Key Laboratory of Component-Based Chinese Medicine, Tianjin, 301617, China*
[c] *State Key Laboratory of Chinese Medicine Modernization, Tianjin, 301617, China*
[d] *Tianjin Key Laboratory of Intelligent and Green Pharmaceuticals for Traditional Chinese Medicine*
[e] *Haihe Laboratory of Modern Chinese Medicine, Tianjin, 301617, China*

[1] Both authors contributed equally to this work.
[*] Corresponding author.
College of Pharmaceutical Engineering of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin, 301617, China.
[**] Corresponding author.
College of Pharmaceutical Engineering of Traditional Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin, 301617, China.
E-mail addresses: lizheng@tjutcm.edu.cn (Z. Li), censhixin@tjutcm.edu.cn (S. Cen).



Review paper

A comprehensive review of Intelligent Question-Answering Systems in Traditional Chinese Medicine Based on LLMs

Abstract：

Large language models (LLMs) are advanced deep learning models with billions or even trillions of parameters, enabling powerful natural language processing and knowledge reasoning capabilities. Their applications in the medical domain have been rapidly expanding, spanning medical research, clinical diagnosis, drug development, and patient management. As a cornerstone of China's healthcare system, traditional Chinese medicine (TCM) faces significant challenges, including difficulties in knowledge extraction, and lack of standardization. The emergence of TCM-focused LLMs presents a transformative opportunity, offering a novel technological framework to process vast amounts of TCM data, uncover hidden theoretical insights, and enhance both research and clinical applications. Despite the growing interest in AI-driven medical solutions, systematic research on LLMs in the TCM domain remains limited. This article provides a comprehensive review of LLM development, detailing their underlying mechanisms, training methodologies, and key technological advancements. It further explores the unique characteristics and diverse application scenarios of existing TCM-LLMs. Additionally, this study also conducts a horizontal comparison of the differences between intelligent question-answering (QA) systems on general LLMs and QA systems on TCM-LLMs, discusses challenges and potential risks, and offers strategic recommendations for future development. By synthesizing current advancements and addressing critical gaps, this work aims to support the continued modernization and intelligent evolution of TCM, fostering its integration into contemporary healthcare systems.

Keywords: Large language models, Traditional Chinese medicine-large language models, Traditional Chinese medicine question-answering systems

1. **Introduction**

Large language models (LLMs) are deep neural networks with a vast number of parameters, designed to develop advanced language comprehension and generation capabilities through large-scale pre-training (PT) tasks. These models learn intricate linguistic structures and patterns, enhancing their ability to process and generate human-like text [1]. Through the processing of massive text data, LLMs perform excellently in various natural language processing (NLP) tasks. For example, in machine translation, text generation, question answering (QA) systems, and sentiment analysis, the capabilities of LLMs significantly surpass those of traditional models, demonstrating strong generalization performance [2,3]. Their deep understanding of complex language features enables them to efficiently handle semantic reasoning, knowledge integration, and task adaptation. Consequently, with their powerful language comprehension and information processing capabilities, LLMs have rapidly expanded their applications in the medical field, including new drug design [4], development of personalized medical treatment plans [5], clinical diagnosis of diseases [6], and medical education [7]. LLMs have demonstrated significant value and become an important technical tool for promoting the intelligent development of healthcare.

Nowadays, the field of traditional Chinese medicine (TCM) faces numerous challenges, such as insufficient data standardization, and difficulty in data integration. TCM data sources are diverse, covering multiple aspects such as ancient books, literatures, and clinical records. The data types include different fields such as biology, chemistry, and pharmacology. In terms of data format, it presents multi-modal characteristics, including various forms such as text, charts, and images. Therefore, we have systematically sorted out and summarized the data sources in TCM (Fig. 1). The integration of TCM information is confronted with complex and highly heterogeneous issues [8,9]. Traditional manual analysis methods are time-consuming，labor-intensive, and difficult to systematically and efficiently uncover the deep-seated patterns behind the data [10].

< **Fig. 1.** Summary of the data in the field of TCM. >

Based on previous reports, some studies have investigated the application of LLMs to overcome the challenges of data integration in TCM analysis [11]. The powerful NLP capabilities of LLMs enable them to quickly process and interpret complex TCM data, especially demonstrating unique advantages in handling TCM's multimodal and heterogeneous data. These models can automatically organize classical Chinese medical literature and extract key information. Additionally, they uncover deep correlations between different data sources through comprehensive analysis of clinical cases and TCM knowledge bases. For instance, by integrating classical TCM literature, modern clinical data, and Chinese medicinal materials knowledge, LLMs can identify diagnostic and treatment patterns, optimize treatment plans, and provide personalized treatment recommendations [12]. LLMs also offer critical technical support for the standardization, digitalization, and intelligent development of TCM [13]. Through their automated data processing and pattern recognition capabilities, LLMs can significantly enhance TCM research and clinical diagnosis. Meanwhile, the model's high-performance computing and data mining potential offer strong support for the transition of TCM from traditional experience-based medicine to modern medicine.

By using the Web of Sciences database, we have retrieved literature related to LLM research on a global scale. The paper search involves keywords such as LLMs, and intelligent QA systems (Fig. 2). After removing irrelevant and duplicate articles, a total of 212 papers have been published since 2020, with 187 published in 2023. Subsequently, we conducted keyword analysis and found that in recent years, the combination of TCM and LLMs has become one of the research hotspots. As a result, a variety of representative TCM-LLMs have emerged, including the Shuzhiqihuang 2.0, Huatuo, Huangdi, and so on. These models have demonstrated extensive application prospects and significant achievements in the practice of TCM.

< **Fig. 2.** Overview analysis diagram of intelligent question-answering (QA) systems based on large language models (LLMs). >

As an important research direction of TCM-LLMs, compared with traditional search engines, intelligent QA systems can quickly and efficiently obtain knowledge and information conveniently. With their powerful automated information retrieval and knowledge reasoning capabilities, they can provide accurate consulting services for TCM and further support clinical decision-making and patient management [14]. Previous reviews of LLMs have focused on their model architectures, key technologies, and training methods [15]. Particularly, in general fields such as NLP and dialogue systems, a relatively systematic understanding has been established [16]. However, with the continuous acceleration of the digital and intelligent transformation of TCM disciplines, TCM-LLMs have gradually become a new research hotspot. Although existing literature have reported some application explorations of TCM-LLMs, such as prescription recommendation, case analysis and QA on TCM knowledge [17], there is still a lack of systematic review and evaluation of their overall development status and potential challenges.

Based on the above background, this work aims to provide a systematic review of relevant research on TCM-LLMs. First, it reviews the evolutionary progression of LLMs as well as the key technologies and core methods, outlines the model architectures and training processes commonly adopted in LLMs. On this basis, this paper further analyzes the unique characteristics and technical advantages of the TCM large language model, deeply explores its diverse application scenarios in practice, and examines the profound impact it may have on the field of TCM. At the same time, this articles also analyzes the core challenges faced by TCM-LLMs in aspects, such as data quality, health management, and ethical norms. These challenges not only determine the technical barriers in this field but also directly affect the sustainable development and wide application of TCM-LLMs. Through comprehensive analysis of relevant research results and limitations, this work summarizes the research status and development trends of TCM-LLMs, puts forward several thoughts and suggestions, and

provides a scientific basis and reference significance for subsequent research and industrial practice.

2. The evolutionary progression of LLMs

The development of LLMs can be roughly divided into four stages: the NLP phase, the basic model phase, the capability exploration stage, and the breakthrough development stage [18,19]. To more systematically show the development process of LLMs, we further sorted and drew its development process diagram (Fig. 3).

< **Fig. 3.** Background of large language models (LLMs) development. >

2.1. The NLP phase (2013-2014)

During this period, word embedding (WE) techniques such as Word2Vec and GloVe rapidly advanced and became a significant milestone in NLP. Unlike traditional methods, such as N-grams or one-hot vectors, WE can map words into a low-dimensional, dense vector space, addressing data sparsity and capturing rich semantic and syntactic features. Word2Vec employs the Continuous-Bag-of-Words and Skip-gram models with shallow neural networks to predict local context, while GloVe generates high-quality word vectors by optimizing the global word frequency co-occurrence matrix. The development of WE technology has spurred innovations in language models and provided efficient, versatile representations for tasks such as sentiment analysis and machine translation [20–22].

2.2. The basic model phase (2017-2018)

In 2017, Vaswani et al. [23] introduced the Transformer architecture, revolutionizing machine translation by overcoming the limitations of traditional Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks. The Transformer's innovative attention mechanism enabled efficient parallel computing, improving both translation quality and model performance. This breakthrough laid the foundation for advancements in NLP tasks such as language generation and question answering, establishing the Transformer as the dominant architecture in deep learning (DL). In 2018, OpenAI released GPT-1, the first pre-trained transformer-based model

with 117 million (M) parameters, followed by Google's BERT model with 340 M parameters [24,25]. The success of these models marked the rise of transformer architectures in NLP, advancing the application of NLP technologies and ushering in a new era in DL dominated by pre-trained models.

2.3. The capability exploration stage (2019-2022)

During this stage, the deployment of large-scale language models faced challenges such as high computational demands, long training times, and the risk of overfitting due to task-specific fine-tuning (FT). Researchers began exploring solutions to eliminate the need for task-specific FT, enhancing model capabilities [26]. Compared to GPT-1, GPT-2 saw significant improvements, with its parameter scale increasing from 117 M to 1.5 billion (B), and training data expanding from 5 GB. These advancements greatly improved GPT-2's generalization, boosting performance across tasks [24,27]. OpenAI then released GPT-3, which further increased its parameter scale to 175 B, positioning it as one of the largest models of its time [24,28]. These advancements in language models provided technical support for new NLP research, shifting tasks from specialized to more generalized. At the same time, large models exhibit extraordinary generality and scalability.

2.4. The breakthrough development stage (2022-Present)

The release of ChatGPT marked a significant milestone in the development of large models, highlighting progress in NLP. With a parameter scale in the hundreds of billions and training data reaching hundreds of terabytes, ChatGPT advanced natural language generation technology and enhanced the intelligence and interactivity of dialogue systems [29]. Building on ChatGPT, OpenAI launched the GPT-3.5 and GPT-4 series, with GPT-4 offering notable improvements in model capabilities and application range, particularly in language generation accuracy and context understanding depth. Additionally, GPT-4's multimodal capabilities enable it to process both text and images, significantly improving user experience [24]. These models provide new directions for multimodal intelligence and complex data analysis,

and offer new tools for the development of TCM.

3. The work principles and key technologies of LLMs

3.1. LLMs working principle:

LLMs are deep neural networks with hundreds of billions of parameters, built on various architectures such as Transformer, MAMBA [30], Falcon Mamba 7B [31], and Receptance Weighted Key Value [32], with transformer being the most fundamental and widely adopted. The Transformer architecture (Fig. 4A), distinct from traditional RNNs and CNNs, enables high parallelism, reduces training time. It also excels at processing large-scale data, making it particularly suitable for tasks like machine translation and text generation. It consists of two core components, the Encoder, which reads and understands input text, and the Decoder, which generates output. This encoder-decoder collaboration underpins its outstanding performance in language modeling and generation tasks [23,24]. The core architecture of the transformer model consists of five fundamental components, namely word embeddings, the attention mechanism, the multi-head attention mechanism, feedforward neural networks and positional encodings.

< **Fig. 4.** Working principle of large language models (LLMs) and training methods. >

3.1.1 Word embedding

Input representation refers to transforming original data (such as text, images or audio) into a numerical form that can be processed by the model, usually represented in the form of vectors. The core goal of this process is to map complex and diverse input data to a unified mathematical space to facilitate the model's understanding and learning [23]. The formula is represented as:

$$E = Embedding(X)$$

Where X denotes the input sequence, and E refers to the embedded representation of the input sequence.

3.1.2 Attention mechanism

The attention mechanism, introduced in 2014 [33], originated in machine translation as a method inspired by human selective focus. It enables models to concentrate on the most relevant parts of input data, improving translation accuracy. By allowing the decoder to dynamically attend to different parts of the source sequence, it alleviates the need for the encoder to compress all information into a fixed-length vector, enabling more effective information flow and selective retrieval during decoding.

The attention mechanism calculates relevance scores using three components: query, key, and value. Analogous to an information retrieval process, the query represents the search intent, the key serves as the index, and the value holds the content. By comparing the query with the key, the model assigns weights to each input element, reflecting their relevance. These weights are then used to generate context-aware outputs. The calculation is as follows [34]:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where Q represents query, K represents key, V denotes value, T is sequence length, and d denotes feature dimension.

3.1.3 Multi-head attention

The multi-head attention mechanism extends traditional attention by computing multiple attention functions in parallel, each as an independent "head." This design significantly improves computational efficiency, particularly in scenarios involving large-scale data and deep neural networks, by facilitating parallelized operations. Moreover, it allows the model to capture information from different representational subspaces across positions, thereby avoiding the over-smoothing effect of single-head attention [35,36]. In practice, we compute attention over a set of queries in matrix Q, along with the corresponding keys and values, to derive the output matrix as follows:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots\ldots, head_H)W^0$$

$$where\ head = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

where H denotes the number of heads, which refers to the number of parallel attention mechanisms, W^0 represents a trainable matrix used for the linear transformation of the output, W_i^Q, W_i^K, and W_i^V represents the linear transformation matrix for the i-th head, QW_i^Q, KW_i^K, and VW_i^V are the results after mapping the query, key, and value to different subspaces.

3.1.4 Feed-forward network (FFN)

Each encoder and decoder layer, aside from the attention sub-layer, contains a fully connected FFN applied independently and identically at each position. This FFN comprises two linear transformations separated by a ReLU activation [37]. Despite its simplicity, the feedforward layer is essential for the transformer's strong performance [38]. Dong et al. [39] highlighted that stacking self-attention modules alone can cause rank collapse, leading to token uniformity bias. The feedforward layer plays a key role in mitigating this issue.

3.1.5 Position encoding

The primary function of a positional encoder is to add a vector representing the position of each input element in a sequence. Since the transformer model lacks the inherent ability to process sequence order, position encoding is added to the input embedding in both the encoder and decoder stacks. This approach addresses the transformer's lack of positional information and enhances its efficiency and flexibility when processing sequence data [23]. By incorporating positional information, the model can better capture sequential relationships. Three common methods for position encoding are as follows:

Sinusoidal Positional Encoding, proposed by Bahdanau et al. [23], is a fixed method that doesn't require additional learning. It uses sine and cosine functions to generate positional encodings, enabling the model to capture relative positional relationships. The periodicity of these functions allows the model to handle sequences of varying lengths [40]. The calculation method of position encoding is as follows:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

where *pos* represents the index of the current element (for instance the position of the word in the sentence), *d* represents the dimension of positional encoding, and *i* refers to the dimension index of the positional encoding.

Learned Positional Encoding treats position encodings as trainable parameters, allowing the model to learn the best representation for each position. Using an embedding layer, the model optimizes encodings through training data, updating them during training, similar to WE [23,41]. Shaw et al. [42] proposed adding learnable relative position embeddings to the attention mechanism, enabling adaptive adjustments of position encodings to enhance task-specific performance.

Relative Positional Encoding offers a more flexible alternative to fixed and learned position encodings by focusing on the relative distances between elements rather than their absolute positions [42–44]. This approach has demonstrated superior performance in specific tasks, particularly those involving long sequences. Notably, Alibi [45] and Rotary Position Embedding (RoPE) [46] are two widely adopted relative position encoding methods in LLMs. The technique has been integrated into several prominent architectures, including Transformer-XL [47], DeBERTa [48], T5 [49], and Roformer [46]. As large-scale models continue to evolve, relative positional encoding is expected to play an increasingly crucial role in managing complex tasks and extensive datasets.

3.2. Training methods

The training of a complete LLM typically involves three stages, i.e., PT from scratch, FT of an existing LLM, and alignment with specific application scenarios via prompt-based methods [50]. In this section, we elaborate on these three stages, PT, FT, and RAG (Fig. 4B).

3.2.1 PT

PT is the first stage in LLMs training, establishing a foundation for model capabilities [51]. During PT, LLMs are trained on large-scale text corpora using

unsupervised or self-supervised learning (SSL) methods to align modalities and acquire multimodal world knowledge [52]. This process enables LLMs to develop rich language understanding and generation skills, including vocabulary relationships, grammar, and context dependencies [53–55].

Common PT strategies include Masked Language Model (MLM), Autoregressive Language Model, and Next Sentence Prediction (NSP). These methods enable the model to learn language patterns from large text corpora and improve performance in various NLP tasks. The MLM captures context by masking part of the input and predicting missing words [56]. The Autoregressive Language Model learns language sequences from left to right, generating text gradually [57]. The NSP analyzes the logical relationship between sentences, capturing long-term dependencies and generating text autoregressively [58].

In addition to these common methods, large-scale PT models also employ advanced techniques such as contrastive learning and multi-task learning, enhancing the model's expressiveness and transferability [59]. Furthermore, emerging methods like Generative Adversarial Networks and meta-learning provide new avenues for improving model learning [60,61]. To explore the practical effects of these PT strategies, we provide a detailed overview of current large language model PT methods, including their underlying structures, model names, parameters, and PT data scales (Table 1) [27,48,49,52, 62-83].

< **Table 1.** Comparison of the model structure and Pre-Training (PT) parameters of general large language models (LLMs). >

3.2.2 FT

After PT, LLMs have obtained general capabilities to solve various tasks. However, research shows that the performance of large models can be further optimized according to specific goals. The goal of FT is to utilize a pre-trained model that has already learned rich language knowledge and make it better adapt to specific tasks or domains by further training on specific datasets [84]. In this process, the PT language

model is initialized with the learned parameters first, and then trained on a specific dataset. In this way, the parameters of the model will be updated according to the data of specific tasks, enabling it to better adapt to the target task [19]. However, the cost of deploying LLM in practical work is very high. Therefore, how to reduce operating costs while maintaining performance has become a new research field. In this section, we summarize several common methods to improve the efficiency of LLM.

Supervised Fine-Tuning (SFT), also known as Instruction Fine-Tuning (IFT) [85], refers to the process of enhancing a pre-trained model's performance through labeled data. Initially, the model learns general knowledge via large-scale unsupervised learning. It is then fine-tuned on annotated data from specific domains or tasks, enabling it to generate more accurate and contextually appropriate outputs for novel inputs [86]. This approach has been widely adopted in leading LLMs such as ChatGPT, FLAN [75], and OPT-IML [87].

Prompt Tuning [56]: Unlike traditional supervised learning, it utilizes a LLM that has been trained on a large-scale text corpus. By defining a new prompt function, the model can achieve better performance on specific tasks.

Prefix Tuning [88]: A lightweight FT method for natural language generation tasks. In prefix tuning, a set of prefix vectors trained for specific tasks are appended to the frozen transformer layers. The prefix vectors are virtual tokens and are attended to by the context tokens on the right. In addition, this method can keep the language model parameters unchanged, thus achieving the purpose of adapting to different tasks.

Adapter Tuning: Adapters are small neural modules inserted between or within transformer layers, enabling efficient task-specific FT without altering the original model parameters [89]. Comprising a dimensionality reduction layer, a nonlinear layer, and a dimensionality expansion layer, adapters introduce minimal trainable parameters while effectively adapting LLMs to downstream tasks. For instance, T5 model employs adapters for FT after PT [90].

Low-Rank Adaptation (LoRA) [91]: LoRA model by inserting trainable low-rank

matrices into key layers, avoiding changes to the original model architecture. Most pre-trained parameters are frozen, with only the low-rank components updated. This approach enables efficient adaptation with minimal parameter updates while preserving the original model's knowledge.

Robust Adaptation (RoSA): Nikdan et al. [92] proposed a parameter-efficient FT method inspired by robust principal component analysis. RoSA jointly trains low-rank and high-sparsity components to enhance model performance. Experimental results show that RoSA outperforms traditional low-rank and sparse FT approaches under limited computational and memory resources.

3.2.3 Distributed training

Due to the extremely large scale of LLMs, training a high-performance LLM poses tremendous challenges. To effectively learn the network parameters of LLMs, distributed training algorithms are usually required, combined with multiple parallel strategies to improve training efficiency. At present, several optimization frameworks for distributed training have been released to promote the implementation and deployment of parallel algorithms, such as DeepSpeed and Megatron-LM.

DeepSpeed [93]: A library for scalable distributed training and inference of DL models not only optimizes memory management but also significantly improves training efficiency.

Megatron-LM [94–96]: An NVIDIA-developed DL library offers optimized support for distributed training through model parallelism, data parallelism, and mixed-precision techniques, significantly enhancing training efficiency across GPUs.

3.3. Retrieval-augmented generation (RAG)

RAG is a technique that enhances LLM performance by integrating external knowledge into the generation process [56]. RAG consists of three key components: retrieval, augmentation, and generation. Specifically, it leverages retrievers to provide relevant context for LLMs, which then utilize their reasoning capabilities to decompose tasks, select appropriate tools, and generate responses [97–99]. Research indicates that

RAG substantially mitigates catastrophic forgetting caused by model weight updates, making it well-suited for domains requiring low tolerance for errors and rapidly evolving information. Compared to traditional FT, RAG allows timely incorporation of new medical knowledge without compromising previously learned information, thereby maintaining output accuracy in dynamic medical settings [50]. Notable implementations of RAG include QA-RAG [100], Almanac [101], Oncology-GPT-4 [102], Impression GPT [103], and Retrieval-Augmented Lay Language Generation [104].

Finally, the three methods described above are all approaches to training LLMs. Compared to PT, FT significantly reduces computational and time costs. However, FT still requires additional model training with high-quality datasets, incurring considerable computational resources and manual effort. In contrast, RAG does not involve updating model parameters, making it a more efficient and convenient approach. Therefore, the choice of training method should be based on the specific requirements of different TCM tasks.

4. TCM QA systems based on LLMs

Nowadays, numerous LLMs have emerged successively, such as Qihuangwendao, Shuzhibencao, Huangdi, and Zhongjing. The development of these TCM-LLMs usually includes the following steps. First, the model is pre-trained through a large-scale general corpus to learn basic language knowledge and patterns. Then, on the basis of PT, professional data in the field of TCM is further used for FT. These professional data cover classical literature of TCM, case data, prescription compatibility, disease diagnosis and treatment, etc., aiming to endow the model with domain knowledge and context understanding ability of TCM. Finally, the fine-tuned LLM can be widely applied to TCM-related tasks, such as TCM diagnosis support, personalized treatment suggestions, drug recommendations, and medical QA [9,10,105]. Through this development process, the model can deeply master the terminology, knowledge systems, and diagnosis and treatment rules of TCM, thereby significantly improving its

performance in TCM-related tasks. This section will focus on introducing the currently developed TCM-LLMs and classifying them according to different task requirements and application scenarios. It will also focus on elaborating on specific aspects such as technical architecture, training dataset size, FT methods, and performance indicators.

4.1. Huatuo

HuaTuo is a TCM-LLM developed based on tuning LLaMA-7B architecture. It integrates both structured and unstructured knowledge from the Chinese Medical Knowledge Graph (CMeKG). Rather than a simple aggregation of resources, HuaTuo incorporates targeted architectural refinements informed by a comprehensive understanding of medical task requirements. As a result, it performs effectively in complex tasks such as medical question answering and the interpretation of domain-specific terminology. To facilitate SFT, over 8,000 high-quality, domain-specific instruction samples were curated by extracting knowledge instances from CMeKG and enhancing them using the OpenAI API. Although relatively limited in size, this dataset is highly specialized and contextually relevant, enabling the model to acquire professional knowledge efficiently, mitigate interference from general-purpose data, and improve its domain-specific reasoning and application capabilities.

During training, general-purpose instructions were excluded due to the specificity of the medical domain, and only the input components were retained. This design encourages the model to concentrate on learning knowledge and generating accurate, professional responses, thereby improving its practical utility. In the evaluation phase, HuaTuo introduced an evaluation metric called Safety, Usability, and Smoothness (SUS), with scores ranging from 1 (unacceptable) to 3 (good). Compared with models such as LLaMA, Alpaca, and ChatGLM, HuaTuo achieved superior performance, obtaining a safety score of 2.88, a knowledge usability score of 2.12, and a smoothness score of 2.47. These results indicate that HuaTuo performs well across multiple dimensions and holds significant potential for intelligent QA systems [106].

4.2. Huangdi

Huangdi is developed based on Ziya-LLaMA-13B-V1. Building on this foundation, it integrates corpora from TCM textbooks, various TCM websites, and other related data sources to construct a pre-trained language model with domain-specific knowledge in TCM. Utilizing extensive dialogue and instruction datasets, the model undergoes SFT, enabling it to effectively respond to questions related to classical TCM texts.

During the PT stage, data from "13th Five-Year Plan" TCM textbooks and folk medicine websites were used to establish a foundational understanding of TCM theory and clinical knowledge. The FT process consists of two parts: general IFT and instruction-based dialogue FT using classical TCM texts. The former uses 52k Chinese data from Alpaca-GPT4 to improve the generality of the model. The latter, based on TCM ancient books, constructs a professional dataset containing more than 500,000 dialogue data, covering fields such as basic TCM theory, disease diagnosis, and the application of prescriptions and herbs.

In terms of data scale, the PT dataset is approximately 0.5 GB, and the processed ancient book dataset reaches 338 MB. These datasets together form diverse dialogue types, supporting the model in comprehensively mastering knowledge of TCM. FT adopts SFT and Direct Preference Optimization (DPO) optimization. The learning rate is set to 3×10^{-4}, lora rank is set to 16, and the number of training rounds is 6, enhancing the model's generation ability and user-friendliness. At the level of model performance evaluation, the train loss drops from 1.456 in the PT stage to 0.00276 in the DPO stage, and the eval loss is 1.258, indicating good generalization ability. Compared with Qwen, the Huangdi performs better in aspects such as TCM theory, diagnosis, and the application of prescriptions, demonstrating its application value and promotion potential in the TCM field.

4.3. Zhongjing

Zhongjing is developed by a research group at Zhengzhou University based on the Baichuan2-13B-Chat and Qwen1.5-1.8B-Chat models, with the goal of enhancing the

application capabilities of LLMs in the field of TCM. Its technical framework includes continuous PT, SFT, and reinforcement learning from human feedback (RLHF). During the continuous PT phase, Zhongjing was trained on diverse real-world medical text data from multiple sources, including electronic health records, medical consultation transcripts, textbooks, and other medical literature. In the next SFT stage, the model was trained using four types of instruction datasets. Among them, the Chinese multi-turn medical dialogue dataset, which includes 70,000 QA pairs across 14 medical departments and more than 10 real-world scenarios, significantly improved the model's conversational ability. During the RLHF stage, the researchers established a comprehensive set of annotation guidelines and enlisted six medical experts to rank 20,000 model-generated sentences. A reward model was then trained using the Proximal Policy Optimization algorithm to better align the model's outputs with expert preferences.

To further optimize performance, the group avoided exclusive reliance on distilled data and carefully balanced the proportion of single-turn and multi-turn medical dialogues during the FT process. In terms of evaluation metrics, Zhongjing demonstrates strong performance across both single-turn and multi-turn interactions in three key dimensions: safety, professionalism, and fluency. In most scenarios, Zhongjing outperforms baseline models, and in multi-turn dialogue specifically, it surpasses all compared models except ChatGPT [9,13,107,108].

4.4. BianQue

Research has found that LLMs such as ChatGLM, ChatGPT, DoctorGLM, and ChatDoctor, perform well in generating general and widely applicable health suggestions in single-turn conversations. However, they exhibit a significant limitation: the lack of Chain-of-Thought (CoT) capability. Unlike real TCM experts, these models struggle to gather comprehensive patient information through continuous and interactive questioning. To alleviate this issue, BianQue has been developed. It has been developed based on the open-source ChatGLM-6B architecture, which features strong

Chinese language understanding and generation abilities. To support the training of the model, the group developed the BianQueCorpus, a health-related big dataset with tens of millions of samples. It integrated some multi-turn conversation datasets such as MedDialog-CN, IMCS-V2, and MedDG, and collected multi-turn health conversations (There are 243,7190 samples, of which 46.2% are the doctors' answers, and the rest are suggestions) in the real world through data outsourcing services. To ensure the quality of the dataset, researchers have developed an automated data cleaning mechanism based on regular expressions. In addition, the model uses ChatGPT to polish the doctors' suggestions for multi-turn conversations. This is because doctors often provide very brief responses through internet platforms, lacking detailed analysis and suggestions.

In terms of performance evaluation, the BianQue model performs outstandingly in multiple Chinese multi-turn medical dialogue datasets. It is evaluated using metrics such as Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and the self-defined Proactive Questioning Ability (PQA). The results show that it outperforms baseline models like ChatGLM-6B, ChatGPT, and DoctorGLM in all indicators, demonstrating excellent generation quality and interactive questioning capabilities. Especially on the MedDG dataset, the BianQue has a remarkably high PQA value of 0.81, which proves that it has a strong proactive questioning ability and excellent medical dialogue generation performance [106,109].

4.5. TCMLLM-PR

In recent years, some open-source LLMs such as ChatGLM, ChatGPT, and LLaMA have all been trained on general-domain data, acquired good general-task processing capabilities, but their proficiency in specific medical fields is limited. To address this, researchers have successively developed specialized TCM-LLM. Although these LLMs have achieved certain progress, they still have not fully addressed the core issues in the TCM field, especially in the area of TCM prescription recommendation. To address the above challenges, researches proposed the TCMLLM-PR model, which is a LLM tailored to TCM prescription recommendation tasks using

the ChatGLM-6B architecture. The dataset for training TCMLLM-PR integrates multi-source heterogeneous information from eight channels, covering four TCM textbooks, *Pharmacopoeia of the People's Republic of China* (2020 Edition) [110], Chinese Medicine Clinical Cases, splenic-stomach disease, and hospital clinical records covering lung disease, diabetes, liver disease, stroke. TCMLLM-PR was then trained using the ChatGLM-6B architecture with P-Tuning v2 technology. On this basis, an instruction-tuning dataset containing 68,654 samples was constructed, with a total scale of approximately 10 M tokens. Ultimately, the model is able to accurately focus on the scenario of recommending TCM prescriptions.

In terms of evaluation metric, the model mainly uses the following indicators Precision@K, Recall@K, and F1 score@K. The experiment results demonstrate that TCMLLM-PR significantly outperforms baseline models on pharmacopoeia datasets and TCM textbooks, achieving F1@10 improvements of 59.48% and 31.80%, respectively. In the cross-dataset transfer task, it performed best when transferring from textbook data to the liver disease dataset, with F1@10 reaching 0.1551. The analysis of real-world cases further confirmed that this model performs outstandingly in the prescription recommendation task. The output results are highly consistent with the real doctors' prescriptions, demonstrating great potential for clinical applications [111].

4.6. TCMChat/TCMGPT

While LLMs have shown excellent performance in medical tasks like QA and diagnosis, TCM-LLMs (such as Ben Cao, Bian Que, HuaTuoGPT) still struggle with limited corpora, data inaccuracies, subjective evaluations, and inadequate tools. These problems have restricted their popularization and application. In this context, TCMChat emerged. The development of TCMChat begins with the utilization of the Baichuan2-7B chat base model. Moreover, it also adopts the typical transformer decoder architecture and has carried out a number of optimizations in the model structure. For example, it uses root mean square normalization (RMSNorm) instead of the normalization layer to achieve a more stable normalization process. It introduces rotary

positional encoding to enhance the sequence modeling ability, and replaces the activation function with SwiGLU which has a stronger expressive ability. It is committed to building a high-performance conversational large-scale model for TCM, enhancing the application efficiency of artificial intelligence (AI) in the modernization process of TCM, and providing technical support for the standardization and popularization of TCM knowledge. To support model training, the research team has constructed a large-scale and high-quality PT and SFT dataset. The PT data covers multiple sources such as books (20 M), web crawlers (30 M), open-source data (352 M), and literature (715 M), and a total of approximately 1G of unsupervised corpus has been constructed. The SFT dataset covers seven task scenarios, including TCM knowledge QA, multiple-choice questions, reading comprehension, entity extraction, medical case diagnosis, Chinese herbal medicines, formula recommendation, and ADMET prediction, with a total of approximately 600,000 QA pairs.

TCMChat has parameters for two stages, PT and SFT. For the PT process, the learning rate is 2×10^{-4}, the batch size is 32 per GPU, and the maximum context length is 1024 tokens. For the SFT process, full-parameter FT is used. The learning rate has been adjusted to 2×10^{-5}, the batch size per GPU is 16, and the maximum context length is limited to 1,024 tokens. In addition, the model also uses the AdamW optimizer and sets the weight decay to 1×10^{-4} to prevent overfitting. In terms of performance evaluation, TCMChat demonstrates strong results across a range of tasks. For multiple-choice questions related to herbs and formulas, the model achieves accuracy rates of 71.6% and 76.8%, respectively. In the reading comprehension task, it reaches a BLEU score of 0.584 and a BertScore as high as 0.886. The entity extraction task yields an impressive F1 score of 0.907. In medical case diagnosis, the model achieves an accuracy of 0.847. For herb and formula recommendation, it records a Mean Reciprocal Rank of 0.536 and a Normalized Discounted Cumulative Gain of 0.439. Lastly, in the ADMET prediction task, TCMChat attains a classification accuracy of 0.818 and a ROC-AUC of 0.830. Overall, TCMChat demonstrates powerful understanding and

generation capabilities in multiple sub-tasks, providing a strong support for AI systems in TCM [16].

4.7. Qibo

Qibo is a model based on LLaMA. During training, it first acquired the basic knowledge and the theoretical framework of TCM through continuous PT, developing capabilities in comprehension, dialectical analysis, and entity recognition of TCM. Subsequently, SFT with diverse datasets was employed to improve its dialogue and instruction following abilities. To enhance TCM inquiry and syndrome differentiation, Qibo incorporates a retrieval-enhanced approach using an external knowledge base. Furthermore, a CoT mechanism is implemented to simulate the real-world TCM consultation process, enabling multi-turn information integration and informed decision-making in prescription generation.

The PT data includes modern medical textbooks, TCM reading comprehension materials, TCM textbooks, TCM prescriptions, and so on. With a total size of approximately 2 GB, this dataset provides the model with a rich and comprehensive knowledge base of TCM. The SFT dataset contains seven types of tasks such as TCM QA, reading comprehension, and prescription recommendation, totaling approximately 600,000 QA pairs. During the training, four types of data are used and converted into the Alpaca format, including single-round conversations, multi-turn conversations in TCM departments, NLP instruction tasks, and general medical conversations, comprehensively improving the generalization and robustness of the model in the fields of TCM.

In terms of model evaluation, Qibo demonstrates significant advantages in multiple dimensions. In subjective evaluation, the model performs outstandingly in terms of professionalism, safety, and fluency compared to the baseline models. Qibo-7B has an average subjective win rate of 63% on 150 TCM questions, and it particularly has a clear advantage in the safety dimension. The objective evaluation takes the form of multiple-choice questions. Based on the test of 3,175 questions related to TCM

professional practice examinations, Qibo's accuracy has increased by 23%-58% compared to the baseline models. In addition, in TCM NLP tasks, such as Entity Recognition Ability, TCM Reading Comprehension, and TCM Syndrome Differentiation. Qibo has achieved ROUGE-Longest (ROUGE-L) scores of 0.72, 0.61, and 0.55 respectively. Although it still does not reach the optimal performance of task-specific models. overall, it outperforms existing medical large language models, fully demonstrating its potential in TCM language understanding and application scenarios [112].

4.8. Lingdan

Based on the Baichuan2-13B-Base model, continuous PT is carried out to obtain the Lingdan model. Furthermore, the Lingdan Traditional Chinese Patent Medicine Chat (Lingdan-TCPM-Chat) model and the Lingdan-prescription recommendation (Lingdan-PR) model are developed. Experimental results show that these two models perform excellently in the tasks of TCM clinical knowledge answering and herbal prescription recommendation. Among them, the Lingdan-PR model has improved by 18.39% in the Top@20 F1 metric compared with the best baseline model, providing strong support for promoting the integrated development of TCM and AI. To support the construction of the above models, the research team has created a large-scale TCM Pre-trained dataset that covers multi-source content such as ancient TCM books and textbooks. They have also used the Baichuan2-13B-Base model to translate ancient TCM texts and have linguistically processed the information about TCM herbs.

During the training stage, the Baichuan2-13B-Base model was used as the base model, and training was conducted on 6 NVIDIA A100-80G GPUs using Quantized-LoRA (QLoRA) and Zero Redundancy Optimizer. Training balance was achieved by configuring LoRA parameters and applying diverse data sampling strategies. Regarding the Lingdan-TCPM-Chat model, 200,000 single-turn dialogue data have been generated through knowledge QA transformation, and 1,599 multi-turn consultation data have been constructed based on the TCM Interactive Diagnostic Dialogue Framework. As

for the Lingdan-PR, a spleen and stomach herbal prescription recommendation dataset has been constructed based on the data from the Department of Spleen and Stomach Diseases in Guang'anmen Hospital. After data augmentation, FT has been carried out on two pre-trained models respectively, and finally, it has comprehensively outperformed existing baseline models in multiple Top@K indicators, verifying the effectiveness and robustness of the method.

The Lingdan outperforms the baseline models in multiple evaluation indicators. In the F1@5 indicator, it has increased by up to 5.82% compared to the highest score of the baseline models. In the F1@10 indicator, it has increased by 11.89%; and in the F1@20 indicator, it has increased by 18.39%. This shows that the Lingdan-PR model has better prediction accuracy, recall rate, and comprehensive performance in the task of TCM prescription recommendation [113].

4.9. Biancang

Biancang is a TCM-specific LLM. It uses a two-stage training process. First, it injected domain-specific knowledge, and then aligns it through targeted stimulation. During the FT phase, four major methods are employed. The first method is SFT, which responds to TCM task instructions through structured QA for training models. The second method is RLHF, which optimizes the model output based on preference scoring. The third method is knowledge enhancement, which integrating the structural information of TCM knowledge graph into the process of contextual understanding. The fourth method is multi-task learning, which integrates tasks such as QA, summarization, translation, and dialogue to enhance the model's generalization ability. In multiple evaluation tasks, Biancang demonstrates outstanding performance. In the Chinese medical exam, its accuracy rate reaches 94.97%, outperforming GPT-4 (82.69%) and Qwen2-7B-Instruct (83.35%). Biancang-Qwen2.5-7B-Instruction has achieved further improvements, reaching an accuracy rate of 82.10% on the TCM syndrome differentiation test set. At the same time, the manual evaluation by TCM experts has also verified Biancang's leading performance in terms of professionalism,

fluency, and security, demonstrating its strong potential in the application of TCM-LLM [17].

4.10. Haiheqibo

Haiheqibo is a knowledge graph model developed for the TCM domain. The construction process of it can be divided into three main stages. Firstly, in the data processing stage, data is collected from medical books, open-source datasets, and crawled data (Wikipedia and Baidu encyclopedia). The data undergoes through processes such as unified formatting (format preprocessing), noise removal (data cleaning), sample deduplication, and quality assessment to ensure the accuracy and diversity of the training data. At each stage, sampled data is manually evaluated to enhance overall data quality. Subsequently, in the PT stage, the above high-quality data is used to construct a PT dataset. Continuous PT is then carried out based on the LLaMA model to generate the TCM-Base Model, which has basic knowledge of TCM but lacks dialogue ability. Then, in the FT stage, data including multi-turn dialogues, single-turn QA, and various NLP tasks are converted into instruction-style formats to construct an instruction dataset. Afterward, SFT is performed on the TCM-Base Model using this dataset to enhance its ability to follow task-specific instructions. Finally, Haiheqibo with both TCM knowledge and dialogue ability is obtained.

4.11. Congbaosuwen

On November 2, 2024, the 5.0 version of the Congbaosuwen was officially released. First, it can accurately understand users' needs. Whether it is professional Chinese medicine consultation or the public's inquiry about health preservation knowledge, it can achieve efficient interaction and enhance the user experience. Second, it features high flexibility. It supports specific vertical-field scenarios and multi-modal data. It can be optimized for scenarios such as clinical practice and teaching. For example, in tongue diagnosis, it can combine with images for assisted diagnosis. Third, its security has been significantly improved. The self-reflection mechanism ensures the accuracy and compliance of the content. Finally, it is compatible API service by

OpenAI, which simplifies development and deployment and accelerates the launch of TCM products.

4.12. Pangu

In 2024, Zhejiang Jiuwei Health Technology Co., Ltd. and Huawei Cloud Computing Technology Co., Ltd. jointly launched the PanGu. This model is a large-scale pre-trained model based on DL technology, specifically designed and optimized for the field of TCM. This model is trained using massive amounts of TCM data, enabling it to deeply understand the language and culture of TCM, providing strong support for the research, development, and application of TCM. At the level of data quality, the PanGu integrates various types of data such as classic TCM literature, TCM prescriptions, medicinal material information, and clinical cases, forming a vast and comprehensive TCM knowledge base. These data not only cover all aspects of TCM but also have been carefully cleaned and annotated to ensure data quality and accuracy. In terms of technology, the PanGu adopts the transformer architecture in DL, which is a neural network structure with powerful feature extraction and context-understanding capabilities. Through large-scale PT, the model can automatically learn the complex knowledge and patterns in the field of TCM, providing a solid foundation for subsequent applications.

In application, it shows broad prospects and potential. First, in the recommendation of TCM prescriptions, the model can intelligently recommend personalized TCM prescriptions based on the patient's symptoms and constitution, improving the accuracy and effectiveness of TCM treatment. Second, in the quality control of medicinal materials, the model can assist in identifying the authenticity and quality of medicinal materials by analyzing information such as the characteristics, origin, and harvesting time of the medicinal materials, ensuring the quality and safety of the medicinal materials. In addition, this model can also play an important role in auxiliary disease diagnosis, new drug research and development, and health management.

4.13. Tianhelingshu

Tianhelingshu is a professional LLM designed for the field of TCM acupuncture and moxibustion. It is built on professional data, including classic Chinese medicine works, acupuncture and moxibustion clinical practice evidence-based database, and TCM evidence-based knowledge map. This model has systematically studied hundreds of classic TCM works and been trained on tens of thousands of pieces of evidence-based data. It has profound knowledge of TCM theory and can serve as an intelligent assistant for TCM to provide users with accurate and professional answers. Whether it is an in-depth discussion of TCM theory or a detailed analysis of health problems, the model can quickly give detailed responses. When users seek advice on acupuncture and moxibustion treatment, it can rapidly analyze users' conditions and put forward personalized suggestions, including various acupuncture and moxibustion treatment methods such as acupuncture, moxibustion, and acupressure.

4.14. Hengqin

Hengqin aggregates a vast amount of TCM data, including 10 B characters of TCM knowledge texts and digital cases from TCM hospitals. Relying on a highly reliable TCM diagnosis and treatment knowledge base, it assists doctors in accurate diagnosis and treatment and provides personalized treatment plans. The Intelligent and Automated Integrated Innovation Platform for New TCM Drugs, through engineering development, realizes a one-stop solution for the entire experimental process of TCM ingredient acquisition, structural characterization, and bioactivity determination based on robotics and automation technologies.

Recently, Hengqin Rheumatoid Arthritis v1 (Hengqin-RA-v1) has been developed. It is specifically designed for the diagnosis and treatment of RA. This model adopts a progressive training workflow, optimizing RA-specific datasets while preserving existing knowledge. Hengqin-RA-v1 integrates domain-specific knowledge through segmented structured data and enhances model performance using instance-oriented and entity-relationship-oriented retrieval enhancements. A sliding window strategy is

723 also employed during training to refine contextual logic and improve the model's
724 understanding of the complex diagnostic context in TCM. Hengqin-RA-v1 outperforms
725 other LLMs in the medical domain, achieving an accuracy rate of 54% in TCM
726 examinations. This result significant outperformed better than both Chinese and non-
727 Chinese models. It excels in generating diagnostic recommendations and treatment
728 plans for rheumatoid arthritis, surpassing traditional approaches in certain scenarios and
729 even outperforming human expert evaluations in some diagnostic cases [114].

730 4.15. Shennong

731 Shennong is jointly completed by the Intelligent Knowledge Management and
732 Service Team of the School of Computer Science and Technology at East China Normal
733 University. It aims to promote the development and implementation of large models in
734 the field of TCM and enhance the knowledge of large models in TCM and their ability
735 to answer medical consultations. Shennong is obtained by using FT with LoRA
736 (rank=16). It is based on an open-source knowledge graph of TCM, with LLaMA
737 serving as the base model. Through an entity-centered self-instruction method, more
738 than 110,000 TCM instruction data is obtained by calling ChatGPT, promoting the
739 inheritance of TCM empowered by large models. However, there are also some
740 shortcomings at the same time. For example, the data relies on an open-source
741 knowledge graph of TCM.

742 Compared with the Chinese LLaMA-7B model, the Shennong demonstrates
743 superior overall performance in TCM QA tasks. First, Shennong exhibits more natural
744 and human-centered language expression; its responses not only address the patient's
745 condition but also convey empathy and attention to emotional needs, thereby enhancing
746 the user experience. Second, owing to large-scale training on TCM-specific
747 instructional data, the model possesses a strong foundation in domain knowledge. It can
748 deliver detailed, actionable treatment suggestions tailored to specific symptoms,
749 including common herbal formulas, methods of administration, and relevant
750 precautions. In contrast, the Chinese LLaMA-7B model, lacking domain-specific

optimization, tends to produce relatively brief and generic responses, with limited professionalism, which makes it insufficient for practical applications in TCM diagnostic and therapeutic contexts. Consequently, Shennong is better suited for intelligent diagnosis and decision-making tasks in TCM.

4.16. Shuzhibencao

Shuzhibencao was jointly developed by Huawei Cloud and Tasly Pharmaceutical Group Co., Ltd. This model integrates an extensive database, including over 1,000 ancient texts and their translations, more than 90,000 traditional prescriptions, upwards of 40,000 Chinese patent medicines, over 40 M literature abstracts, more than 3 M natural products, data on over 20,000 target gene pathways, more than 100,000 clinical treatment protocols, over 160,000 Chinese medicine patents, and a wide range of pharmacopoeia and policy guidelines. This model has 38 B parameters and has been pre-trained on a vast corpus of TCM texts. By integrating and combining the reinforcement of vector library retrieval and FT in multiple scenarios of Chinese medicine research and development, it can better assist researchers in mining and summarizing the theoretical evidence of TCM.

Shuzhibencao was pre-trained based on billions of molecular structures and further fine-tuned using 3.5 M unique natural product molecules. This enables it to more accurately perform computational tasks, such as characterizing natural product structures and improving the prediction of their downstream properties. By integrating with appropriate algorithms, the model can also accelerate the screening and optimization of medicinal materials and compound prescriptions [115].

4.17. Bencaozhiku

BenCaozhiku was released on 2024, during the second "Thousand Herbal Genomes Project" symposium. This model integrates core foundational data for TCM research, including 15 M genomic records of source species for medicinal herbs, over 30 M records on interactions between TCM compounds and their targets, and more than 4 M compounds. It forms a knowledge graph comprising over 20 M entities and more

than 2 B relationships, covering the entire TCM industry chain. Powered by the Wenxinyiyan LLM with hundreds of billions of parameters, and enhanced through instruction tuning and RAG techniques, the model supports three key functions: extraction and generation of TCM knowledge, delivery of domain-specific TCM solutions, and comprehensive digital services for the TCM industry. It achieves seamless integration of foundational research data with critical stages across the TCM value chain.

4.18. Qihuangwendao

In July 2023, Baidu Health and GuShengTang Incorporated have released the Qihuangwendao. Based on the training of TCM knowledge, this model takes more than 1,000 ancient Chinese medical books and TCM documents such as *Huangdi Neijing* and *Treatise on Cold Damage and Miscellaneous Diseases* as its core data foundation. It covers 11 M pieces of data in the knowledge graph of TCM, 2 M pieces of real clinical diagnosis and treatment data of TCM, 100 thousand pieces of real medical case data of TCM experts, and 100 thousand pieces of data on pulse conditions, tongue manifestations, meridians, and acupoints. It has efficient operation capabilities and accurate syndrome differentiation capabilities, and has achieved a high level of specialization.

Qihuangwendao consists of two application functionalities: the medical LLM and the health preservation LLM. The medical LLM is further divided into two sub-models: providing prescriptions for confirmed diagnoses and conducting diagnostic assessments based on symptoms. Through the construction technology of the knowledge graph of the experience of famous veteran TCM doctors, their experiences are sorted out to provide knowledge support for the model. The natural language recognition technology of TCM is applied to improve the ability to interpret text related to diseases and symptoms. With the help of the technology for constructing standardized symptoms and signs in TCM, expressions are standardized for accurate analysis. The big data mining technology of TCM diagnosis and treatment is adopted to extract key

information from massive data and conduct training. Ultimately, the model can accurately diagnose complex diseases, conduct syndrome differentiation of symptoms, and make professional judgments. It can customize personalized treatment plans for different users, such as recommending TCM prescriptions, guiding meridian massage, and providing dietary therapy suggestions. The health preservation LLM, creating personalized multidimensional wellness plans to help maintain health and prevent diseases [9,13].

4.19. Shuzhiqihuang 2.0

On November 25 in 2024, at the fourth "Big Data and AI in TCM" Shanghai Forum, Northeast Normal University, in collaboration with multiple institutions, released the Shuzhiqihuang 2.0 multi-modal large model in the field of TCM. This model contains 32 B parameters and covers two main modules of TCM and western medicine: it has more than 200,000 and 100,000 instruction data respectively. In addition, the model covers over 80,000 TCM prescriptions, more than 40,000 TCM ingredients, encompasses 9,000+ kinds of TCM materials, incorporates 2,000+ TCM syndromes, and contains 1,000+ ancient books; including 18,000+ targets, 2,000+ diseases, 2.4 M compounds (of which about 410,000 are natural products), and more than 2 M documents. With its advanced multi-modal capabilities and extensive knowledge base coverage, Shuzhiqihuang 2.0 shows a new breakthrough in the field of intelligent diagnosis and treatment.

4.20. MedChatZH

MedChatZH is a dialogue model specifically designed and optimized for TCM consultations, demonstrating significant advantages in aspects such as technical architecture, training data, and FT methods. This model is constructed based on the Baichuan-7B (similar to LLaMa). RMSNorm is adopted to normalize the input of each sublayer, which improves the stability of the training process and the effect of layer normalization. By incorporating the Rotational Position Embedding mechanism, it integrates relative and absolute positional information, thus significantly enhancing the

model's generalization and understanding capabilities for different text lengths and structures. At the data level, the PT dataset of MedChatZH encompasses more than 1,000 TCM ancient books and modern books, including *Treatise on Febrile Diseases*, *the Yellow Emperor's Canon of Internal Medicine*, and *A Barefoot Doctor's Manual*. It covers a wide range of TCM theories and clinical experiences. In addition, the medical IFT dataset (med-mix-2M) is introduced, which contains 763,629 medical instructions and 1,305,194 general instructions. The data sources integrate multiple projects such as belle-3.5M, medical, medical-dialogue, and so on, laying a solid foundation for the application of the model in professional and general dialogue scenarios.

In terms of the FT, MedChatZH addresses the differences between the language style of TCM books and the requirements of modern conversations. First, use Baidu Wenxinyiyan to convert ancient Chinese texts into modern Chinese, and then leverage the ChatGPT API to optimize the translation quality, constructing a high-quality PT corpus. The model undergoes complex data processing steps, including heuristic and model-based filtering to remove irrelevant or sensitive content. Medical instruction data further undergoes a three-fold processing: First, personal privacy information is removed through regular expressions. Second, the trained Ziya-LLaMA-7B-Reward model is used to score the data, and low-quality samples with a score lower than 0.5 are eliminated. Third, the format of numerical symbols is unified to enhance data standardization. In the FT stage, reinforcement learning (RL) techniques is adopted to convert the instruction data into a QA template in the "Human-Assistant" format. Only the answer part generated by the model is used to calculate the loss and update. In the webMedQA dataset test, MedChatZH performs outstandingly on automatic evaluation metrics such as BLEU, GLEU, and ROUGE. Under the BLEU-1 metric, MedChatZH scores 56.14, while ChatGLM-Med scores 32.18 and BenTsao scores 32.02. In terms of the ROUGE-L metric, MedChatZH reaches 35.99, ChatGLM-Med is 26.14, and BenTsao is only 17.72. This indicates that the answers generated by MedChatZH are superior to models such as ChatGLM-Med and BenTsao in terms of similarity to

reference sentences, fluency, and quality evaluation based on word matching [3].

4.21. Chinese patent medicine instructions-ChatGLM (CPMI-ChatGLM)

CPMI-ChatGLM is based on the ChatGLM-6B model and adopts a prefix decoder-only transformer framework. It integrates the bidirectional and unidirectional attention mechanisms, which are used to process input and output information respectively. The model ensures the stability of the training process through the gradient scaling embedding layer and the post-LN layer normalization method. At the same time, it introduces RoPE to replace the traditional absolute position encoding, and adopts the GeLU activation function in the FFNs to enhance the expressive ability and generalization performance. In terms of training data, the core dataset of CPMI-ChatGLM is derived from the Entity Recognition of TCM, *Standard Therapeutic Guidelines for National Essential Drugs*, and instructions. After data cleaning, denoising, and the removal of drugs with unknown attributes, data augmentation is achieved in combination with ChatGLM. Eventually, a dataset containing 3,906 high-quality data records is constructed.

In terms of FT methods, CPMI-ChatGLM applies the parameter-efficient FT technology, focusing on comparing two methods, LoRA and P-Tuning v2. LoRA optimizes model parameters by introducing low-rank matrices to assist in the update process, while P-Tuning v2 optimizes performance by introducing continuous prompts at each layer of the model. Experiments show that FT based on P-Tuning v2 performs better in multiple evaluation metrics. For example, ROUGE-1 F1 score is 33.81% higher than that of LoRA. In addition, the model is also fine-tuned with instruction data containing TCM knowledge. For domain-specific tasks, a small set of instruction data is often sufficient to guide generation, in contrast to general-domain models. This strategy improves the performance of the model in tasks such as the recommendation. For performance evaluation, the model comprehensively assesses the similarity and quality of the generated text using automatic evaluation metrics such as BLEU, ROUGE, and BARTScore metrics. It achieves a score of 0.7641 on the BLEU-4 metric,

demonstrating excellent performance. Regarding manual evaluation, the SUS standard is introduced to evaluate the reliability and practicality of the content generated by the model from three aspects: safety, usability, and fluency. The results show that CPMI-ChatGLM scores highly in all three indicators, indicating that it has good application prospects in the field of Chinese herbal medicine instruction manuals [116].

4.22. MING-mixture-of-expert (MING-MOE)

MING-MOE, a medical LLM based on MOE, is designed to manage diverse and complex medical tasks. It does not require task-specific annotations, thus improving its usability across extensive datasets. MING-MOE adopts a Mixture of Low-Rank Adaptation technique. This technique allows for efficient parameter utilization by keeping the basic model parameters static while enabling adaptation through a minimal set of trainable parameters. Literature demonstrate that MING-MOE achieves state-of-the-art performance in more than 20 medical tasks, indicating a significant improvement over existing models. This approach not only expands the capabilities of medical language models but also enhances the inference efficiency. MING-MOE is a bilingual (Chinese and English) medical LLM built upon the foundation of the Qwen1.5-Chat. The FT process mainly focuses on enhancing the model's stability in handling specific tasks in the medical field. At the same time, a certain proportion of medical QA and interaction data are retained in the FT dataset reserve for the model's general capabilities. Based on the size of the base model, four different sizes were proposed, including MING-MOE (1.8 B), MING-MOE (4 B), MING-MOE (7 B), and MING-MOE (14 B).

The results of the automatic evaluation metrics showed that two studies reported accuracy in medical tasks. For TCM diagnosis, TCM-GPT achieved an accuracy of 0.264, while MING-MOE achieved 41.58 (1.8 B), 50.31 (4 B), 57.03 (7 B), and 63.2 (14 B). In TCM examinations, the reported accuracy was 0.29 for TCM-GPT and 33.96 (1.8 B), 45.00 (4 B), 49.58 (7 B), and 59.79 (14 B) for MING-MOE. In summary, MING-MOE provides accurate answers and reasonable interpretations for a variety of

tasks, demonstrating strong capabilities in knowledge application and problem-solving, and has high application value in the field of TCM [117].

4.23. IvyGPT

IvyGPT is built on the LLM architecture and adopts a two-stage training strategy to form its technical system. In the first stage SFT process, it is optimized based on LLaMA-33B. By introducing the LoRA method, the weights of the original model are frozen, and trainable low-rank matrices are injected into each layer of the transformer. This significantly reduces the number of trainable parameters required in downstream tasks. On this basis, the QLoRA technology is further adopted to perform 4-bit quantization on the base model. This enables the efficient FT of large-scale models even in low memory resource environments, significantly improving the training efficiency and adaptability. In the second stage, RLHF is introduced to optimize the model. During the training process, the model generates responses based on real instructions and multi-turn conversations. Then, human evaluators score the responses from four dimensions: considering informativeness, coherence, adherence to human preferences, and factual accuracy, which is used to guide the learning of the reward model. At the same time, to prevent the model trained through RL from deviating too far from the model in the SFT stage. Constraints are introduced during the training process to ensure that the model achieves better results without deviating from the track. In addition, the top k responses generated by the model are scored, and by defining a reward function, the model is guided to generate answers that are more in line with human preferences, thus achieving the unity of performance and stability.

Compared with models such as HuaTuo, Shennong, ChatMed, MedicalGPT, and ChatGPT, IvyGPT achieved the highest semantic similarity score (93.58) in 100 QA tasks, demonstrating stronger semantic understanding and expression capabilities. Moreover, the average word count of responses generated by IvyGPT is 271.05, which is higher than that of other current models, indicating that its answers are more informative. Overall, IvyGPT performs outstandingly in terms of training efficiency,

output quality, and semantic consistency, showing great potential for medical applications [118].

Furthermore, we have also conducted a systematic compare of TCM-LLMs (Table S1). In practical applications, the selection of TCM-LLMs should be approached from multiple dimensions. First, it is essential to define the specific application scenario. For diagnostic reasoning tasks, models such as Zhongjing and Qibo demonstrate strong performance. For prescription recommendation, TCMLLM-PR and Lingdan are more suitable. In QA and consultation contexts, Huatuo, MedChatZH, and Biancang are recommended. Open-source models like GLM-130B, Huatuo, and TCMChat should be prioritized, as they facilitate local deployment and personalized FT, particularly important in handling sensitive data. Model adaptability should also be considered in relation to training methods. For instance, Huatuo and TCMLLM-PR employ IFT, while IvyGPT and Biancang incorporate RLHF. Regarding evaluation, Biancang achieves an accuracy of 94.97% in chronic disease management, and the BLEU score of MedChatZH also serves as a useful performance indicator. For tasks involving multimodal input and knowledge graph integration, models such as CPMI-ChatGLM and Shuzhiqihuang 2.0 are optimal. Ensuring the stable operation of deployed models through continuous optimization is also recommended.

< **Table S1.** Summary of existing intelligence traditional Chinese medicine (TCM) intelligent question-answering (QA) systems based on large language models (LLMs) in terms of open-source availability, model development, parameter size, training methods, features, and data sources**.** >

5. Prospects for the application of TCM-LLMs

According to the information, as of November 30, 2022, more than 100 M users have retrieved LLMs and medicine in PubMed. The number of included articles increased from 182 in 2020 to 867 in 2023. Thus, it can be seen that large models have become a new direction to help the development of TCM [9]. As a medical auxiliary tool, large models in the field of TCM have shown great application potential in medical

education, new drug research and development, and medical clinical practice. The application prospects in the field of TCM are very broad. In addition, we have drawn a figure illustrating the application prospects of LLMs in the field of TCM (Fig. 5).

< **Fig. 5.** Prospects for the application of traditional Chinese medicine (TCM)-large language models (LLMs).>

5.1 Medical education

5.1.1 Learning ancient books

Ancient TCM books not only record the theoretical innovations, clinical experiences and medical techniques of ancient practitioners, but also provide valuable references and inspirations for modern TCM diagnosis and treatment of various diseases. However, due to the complexity of content, systematically absorbing the vast information in these ancient books has always been a challenge. For this reason, many researchers are focusing on using advanced technologies to conduct in-depth studies on ancient texts. LLMs can establish a knowledge base of TCM through learning and analyzing TCM literature, providing researchers with more comprehensive and accurate knowledge of TCM. Since TCM research requires a large number literature as support, and these literature materials are often scattered in various fields and disciplines, LLMs can integrate these scattered literature materials to establish a comprehensive and accurate knowledge base of TCM [13, 119, 120].

LLMs achieve comprehensive learning and in-depth exploration of TCM knowledge by integrating a large amount of TCM literature, ancient books, medical records, and modern TCM research results. At the same time, they cover core TCM theories such as the theory of yin and yang, the five elements theory, syndrome differentiation and treatment, and meridian theory, as well as treatment plans for different diseases. The Huangdi constructed by Zhang and co-workers [121] and others provides users with knowledge services in multiple aspects such as answering questions about ancient books, TCM consultations, treatment suggestions, and preventive health preservation through natural language conversations. This model not only deeply

excavates the knowledge value contained in ancient books but also explores a new paradigm for the research and application of TCM ancient books.

5.1.2 Intelligent teaching

Compared with traditional teaching methods, LLMs can automatically generate the required teaching materials and even provide virtual patient cases for teachers and students to conduct discussion, analysis, and simulated diagnosis. At the same time, teachers can use large models to formulate targeted teaching plans and realize the personalization of TCM teaching. By inputting individual information such as students' strengths, weaknesses, learning goals, and preferences, the model can generate personalized learning plans that meet the specific needs of students, thus providing a basis for precise tutoring and significantly improving teaching efficiency [122]. In language learning, LLMs can simulate conversations in multiple languages, correct grammar, increase vocabulary, and help improve pronunciation to meet needs [123].

5.2. Drug development

New drug research and development is crucial for promoting the development of China's healthcare industry. It not only fills the gaps of existing drugs but also meets unmet treatment needs. Drug research and development can be roughly divided into four stages: (1) Target Selection and Validation. Drug research and development usually starts from identifying targets related to specific diseases. At this stage, multiple technical means such as cell and gene target assessment, genome and proteome analysis, and bioinformatics prediction need to be combined. (2) Compound Screening and Lead Optimization. By identifying active compounds for screening, methods such as combinatorial chemistry, high-throughput screening, and virtual screening are usually employed to find candidate compounds from public databases (such as PubChem, ChemBL Database, etc.) (3) Preclinical Research. Through structure-activity relationship research and computer simulation technology, combined with cell function testing, repeated iterative optimization is carried out to improve the functional performance of newly synthesized candidate drugs. (4) Clinical Trials. Before entering

the preclinical research stage, candidate compounds need to undergo in vivo testing in animal models, including pharmacokinetic studies and toxicity tests, to ensure their basic safety and effectiveness [124–127].

According to relevant literature, new drug research and development has the largest share in the global medical AI market, reaching 35% [128]. However, new drugs also face many difficulties in the research and development stage, including long research and development cycles, high costs, and low success rates [129]. Therefore, integrating LLMs into the field of drug discovery and development marks a major paradigm shift and provides new methods for understanding disease mechanisms, promoting drug discovery, and optimizing the clinical trial process [4]. Traditional drug screening methods include molecular docking, pharmacophore matching, and similarity search [128]. Different from the traditional drug research and development process, AI has been widely used in all aspects of new drug research and development through key AI technologies such as NLP, ML, DL, and knowledge graphs. including drug target identification, active compound screening, compound property prediction, molecular generation, and protein structure and protein-ligand interaction prediction [130,131]. The TCM-LLMs can quickly discovers the connection between drugs and diseases, and between diseases and genes through DL technology, thereby shortening the drug research and development cycle, reducing drug research and development costs, and reducing the risk of new drug research and development [132,133]. At the same time, LLMs can rapidly screen out required TCM articles through automated literature retrieval and analysis. They can extract key information and potential drug candidates, synthesize innovative new drug prescriptions. Additionally, they can extract available data from experimental reports and generate initial report texts to assist in designing experimental schemes and enhance research efficiency [11].

5.3. Clinical diagnosis and treatment

5.3.1 Clinical data processing

These big models possess highly advanced computational capabilities. They can rapidly and accurately search through a large amount of medical data, collect the latest information on diseases, and provide reliable scientific bases for researchers. They can conduct precise case analyses and thus more quickly raise the level of medical research. LLMs can automatically extract and deeply understand massive amounts of medical data from medical literature, electronic medical records, clinical trials, and more. This helps construct a more comprehensive and accurate medical knowledge graph and enables researchers to quickly sift out useful information from a large amount of complex data [134]. In addition, these LLMs also have strong information integration capabilities and can achieve cross-modal fusion between different data sources. Patel et al. [135] found through research that big models can quickly generate standardized discharge summaries, reducing the work burden on doctors and saving more time for doctors to take care of patients.

5.3.2 Clinical decision support

LLM-based intelligent QA systems can automatically extract patients' symptom information, analyze and diagnose it according to TCM theory, and provide reference opinions for doctors. In terms of case analysis, through the analysis of text data such as medical records and symptom descriptions, large models can help doctors quickly understand patients' conditions and provide preliminary diagnostic suggestions. In terms of drug recommendation, LLMs can automatically match suitable drugs according to patients' conditions and symptoms and put forward reasonable medication suggestions. In terms of disease prediction, through the analysis of a large number of case data, LLMs can predict the occurrence of certain diseases [136].

5.3.3 Providing diagnostic and treatment recommendations

In the field of TCM, LLMs have been widely applied in fields such as diagnosis and treatment, medicine, and health preservation due to their excellent performance [137]. These models have unique advantages. For example, they can be perfectly combined with "the four diagnostic methods", achieve a perfect match between TCM

natural language and SSL, and be adaptable to the characteristics of TCM compound prescriptions. In addition, these models can help TCM professionals in diagnosis and treatment research and assist in TCM diagnosis and treatment [9]. Treating diseases according to their characteristics and providing precise treatment requires doctors to provide personalized diagnosis and treatment plans according to the conditions of different patients. LLMs can analyze patients' personal health information and other relevant data, striving for "one prescription for one person", "one prescription for one symptom", and "one strategy for one person", thereby improving the patient experience.

We selected a question (Fig. 6), and the differences between intelligence TCM QA systems based on LLMs and general LLMs were compared. Six TCM-LLMs exhibit distinct characteristics. Shuzhiqihuang lacks detailed guidance on drug dosage. TCMChat adopts syndrome differentiation based on classical theories, but demonstrates limited applicability to contemporary diseases. Shuzhibencao offers flexibility and diversity in treatment approaches. Congbaosuwen primarily emphasizes dietary and lifestyle regulation, but lacks complete treatment protocols. Huatuo provides an integrated treatment plan, highlighting external therapies and requiring active patient participation. CMLM-Zhongjing provides a user-friendly and efficient solution without the necessity of local deployment. In the horizontal comparison with general LLMs, TCM-LLMs excels in providing a comprehensive and traditional diagnosis by emphasizing patterns of disharmony in the body and offering well-established herbal remedies rooted in ancient wisdom. Its response is more in-depth and patient-centric, deeply tied to the TCM approach. The general LLM focuses more on immediate symptom relief and general medical advice, which can be very useful for users seeking quick, practical solutions. However, it lacks the deeper, more nuanced diagnosis that a TCM-specific model would offer. It also offers more conventional treatments that align with modern medical practices. Therefore, based on a comprehensive analysis of the characteristics and applicability of these TCM-LLMs, it is recommended that patients flexibly apply these models during the treatment process,

combining their personal symptoms and physical constitutions in a comprehensive and complementary manner.

< **Fig. 6.** Comparison of intelligence traditional Chinese medicine (TCM) question-answering (QA) systems with general large language models (LLMs). >

5.3.4 Equitable distribution of medical resources

At present, China faces severe challenges in uneven distribution of medical resources, shortage of primary care doctors, and prevention and treatment of chronic diseases. Especially in remote areas, the problems of scarce medical resources and limited diagnosis and treatment levels are more prominent, directly affecting the health security and quality of life of the people [138]. The development of LLMs in the field of TCM provides new ideas and technical support for solving these problems. By integrating abundant knowledge of TCM and modern medical research results, LLMs can realize the sharing of medical resources nationwide, provide real-time diagnostic assistance, treatment suggestions, and chronic disease management plans for primary care doctors. The development of LLMs in the field of TCM can promote the sharing of medical resources nationwide, improve the medical diagnosis level in areas with limited medical resources, and provide innovative solutions for China's primary health care services [139].

6. Discussion

Currently, TCM intelligent QA systems based on LLMs have been widely applied. However, these systems continue to face numerous challenges in handling upstream tasks. Firstly, in actual diagnostic consultations, the massive clinical data and patient health information recorded by hospitals are sensitive and must be collected, stored, and used in strict compliance with regulations. To prevent patient information leaks, advanced technical measures must be adopted alongside strict adherence to privacy regulations. Federated Learning, an emerging privacy-preserving framework where "data remains local and only model parameters are uploaded", has garnered widespread attention [140]. This mechanism allows medical institutions to participate in

collaborative model training without moving their local data from its original storage environment, effectively alleviating conflicts between data sharing limitations and privacy risks. Drawing on the principles of Federated Learning, its application to the training and optimization of TCM-LLMs may offer a viable solution to the current challenges of high data sensitivity and difficult cross-institutional collaboration.

Simultaneously, due to the large scale of current model parameters, it is quite difficult for patients to deploy the model on terminal devices. To enhance the accessibility and practicality of the TCM-LLMs in practical applications, model lightweighting has become a crucial direction. Through technologies such as model compression, knowledge distillation, and parameter pruning, the computational resources required during the inference phase can be significantly reduced, making it feasible to deploy the model on mobile devices and facilitating the provision of convenient medical services. In addition, MoE architecture achieves an optimized balance between performance and computational efficiency by selectively activating some expert sub-modules, providing a promising research direction for the efficient deployment and application of TCM-LLM in the future [141].

TCM diagnosis involves inspection, auscultation and olfaction, inquiry, and palpation. However, most of the currently constructed TCM-LLMs still mainly rely on text data and lack the ability to deeply integrate unstructured and multimodal data such as images (such as tongue images, facial complexion), voice, odor descriptions, and pulse conditions [142,143]. This holistic approach to data integration can lead to more comprehensive and accurate medical insights, enabling TCM-LLMs to provide a more complete picture of a patient’s health status. Comprehensive data integration not only helps to enhance the model's comprehensive perception of patients' health conditions, but also significantly improves the application performance of TCM-LLMs in clinical auxiliary diagnosis and intelligent QA systems. This enables them to respond more accurately to complex medical queries and provide personalized diagnosis and treatment recommendations. However, there are still many technical challenges at

present. For example, cross-modal learning involves how to transfer information and knowledge between different modalities, especially how to design effective multi-modal alignment methods. Data from different modalities often contain features at different levels, how to integrate this information through algorithms while maintaining its independence and diversity is a major difficulty in the development of multi-modal TCM-LLMs.

Additionally, in the process of drug development, compound-related data exhibit highly complex and heterogeneous characteristics [144]. Common data types include SMILES representations of molecules, two-dimensional molecular structure diagrams, three-dimensional conformational information, medicinal material source literature, and formula context [145]. These data originate from diverse sources, ranging from modern laboratory measurements to historical documents or pharmacopoeia extractions, leading to inconsistent formats, semantic discrepancies, and uneven data quality. For example, while SMILES data are concise and easy to process, they lack spatial positional information and struggle to accurately reflect stereochemical properties; image or structural data, though information-rich, still lack standardized expression and annotation methods. Effectively integrating multimodal data remains a key technical bottleneck in constructing unified and high-quality molecular representation models. Additionally, TCM drug development emphasizes the holistic synergistic effects of compound formulas, with their mechanisms of action generally following a systemic regulatory pattern of multi-component-multi-target-multi-pathway. Traditional modeling paradigms centered on single-drug-single-target-single-action-pathway are ill-suited to this context, failing to effectively capture synergistic or antagonistic interactions between components [146]. Therefore, there is an urgent need to develop large models capable of integrating structured chemical data with TCM compound formula knowledge to support key processes such as efficacy prediction, mechanism-of-action modeling, and candidate drug screening.

In future research, QA must exhibit traceability and interpretability in clinical applications to gain the trust of both healthcare professionals and patients. Future research should incorporate techniques such as causal reasoning, symbolic logic, and visual path analysis to make the model's diagnostic and recommendation processes transparent and controllable, reducing the risk of "black-box" phenomena.

7. Conclusion

This work provides a comprehensive review of the development of LLMs and offers an in-depth analysis of their key components, including model architecture, training methods, and core development technologies. Building on this foundation, we systematically analysis existing TCM intelligent QA systems based on LLM, highlighting their promising application prospects in areas such as medical education, drug development, and clinical diagnosis and treatment. By reviewing these systems, the study aims to offer valuable references for future research and development of LLMs in the TCM domain.

References：

[1] Y. Yao, J. Duan, K. Xu, et al., A survey on large language model (llm) security and privacy: The good, the bad, and the ugly, High-Confid. Comput. 4 (2024), 100211.

[2] J. Yang, H. Jin, R. Tang, et al., Harnessing the power of llms in practice: A survey on chatgpt and beyond, ACM. Trans. Knowl. Discov. Data. 18 (2024) 1–32.

[3] Y. Tan, Z. Zhang, M. Li, et al., MedChatZH: A tuning LLM for traditional Chinese medicine consultations, Comput. Biol. Med. 172 (2024), 108290.

[4] Y. Zheng, H.Y. Koh, M. Yang, et al., Large language models in drug discovery and development: From disease mechanisms to clinical trials, arXiv. 2024. https://arXiv.org/abs/2409.04481.

[5] J. Cosentino, A. Belyaeva, X. Liu, et al., Towards a Personal Health Large Language Model, arXiv. 2024. https://arXiv.org/abs/2406.06474.

[6] C. Wu, Z. Lin, W.L. Fang, et al., A medical diagnostic assistant based on LLM, In China Health Information Processing Conference, Springer, Berlin, 2024, pp. 135–147.

[7] X. Meng, X. Yan, K. Zhang, et al., The application of large language models in medicine: A scoping review, Iscience 27 (2024), 109713.

[8] K. Liu, H. Zhang, H. Liu, et al., Research and Practice on The Establishment of Intelligent TCM Dialectical Treatment Platform, Chinese Journal of Health Informatics and Management 20 (2023) 333–338.

[9] Z. Chen, W. Peng, D. Zhang, et al., Application, Challenges, and Prospects of Large Language Model in the Field of Traditional Chinese Medicine, Med. J. PUMCH. 16 (2025) 83–89.

[10] W. Xiao, C. Song, S. Chen, et al., Key technologies and construction strategies of large language models for traditional Chinese medicine, CHM. 55 (2024) 5747–5756.

[11] L. Yang, Z. Wang, K. Yao, et al., Prospective Reflections on Application of Large Language Models in Field of Traditional Chinese Medicine, Chin. Arch. Tradit. Chin. Med. 43 (2025) 16–24.

[12] X. Wang, T. Yang, K. Hu, Research on Personalized Prescription Recommendation of Traditional Chinese Medicine Based on Large Language Pre-Training Model, Chin. Arch. Tradit. Chin. Med. 42 (2024) 15–18+264.

[13] C. Bai, J. Wang, The Application of the Artificial Intelligence Large Language Model in the Field of Traditional Chinese Medicine, Journal of Xichang University 38 (2024) 62–69.

[14] Y. Yin, L. Zhang, Y. Wang, et al., Question answering system based on knowledge graph in traditional Chinese medicine diagnosis and treatment of viral hepatitis B, Biomed. Res. Int. (2022), 7139904.

[15] M.U. Hadi, A.-T. Qasem, R. Qureshi, et al., Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects, Authorea Preprints 1 (2023) 1–26.

[16] Y. Dai, X. Shao, J. Zhang, et al., TCMChat: A generative large language model for traditional Chinese medicine, Pharmacol. Res. 210 (2024), 107530.

[17] S. Wei, X. Peng, Y.-F. Wang, et al., BianCang: A Traditional Chinese Medicine Large Language Model, arXiv. 2024. https://arXiv.org/abs/2411.11027.

[18] Q. Zhang, T. Gui, R. Zheng, et al., Large-scale language models: from theory to practice, 2023, pp. 5–6.

[19] W.X. Zhao, K. Zhou, J. Li, et al., A survey of large language models, arXiv. 2023. https://arXiv.org/abs/2303.18223.

[20] B. Wang, A. Wang, F.X. Chen, et al., Evaluating word embedding models: Methods and experimental results, APSIPA trans. signal. 8 (2019), e19.

[21] T. Mikolov, K. Chen, G. Corradio, et al., Efficient estimation of word representations in vector space, arXiv. 2013. https:// arXiv.org/abs/1301.3781.

[22] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[23] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need, arXiv. 2017,

https://arXiv.org/abs/1706.03762.

[24] Y. Wang, Q. Li, Z. Dai, et al., Current status and trends in large language modeling research, Chin. J. Eng. 46 (2024) 1411–1425.

[25] J. Devlin, M.-W. Chang, K. Lee,et al., Bert: Pre-training of deep bidirectional transformers for language understanding, Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (long and short papers), 2019, pp. 4171–4186.

[26] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, arXiv. 2018. https://arXiv.org/abs/1801.06146.

[27] A. Radford, J. Wu, R. Child, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019), 9.

[28] T.B. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, arXiv. 2020. https://arXiv.org/abs/2005.14165.

[29] Y.Q. Shen, L. Heacock, J. Elias, et al., ChatGPT and other large language models are double-edged swords, Radiology. 307 (2023), e230163.

[30] J.T. Halloran, M.S.Gulati, P.F. Roysdon, Mamba state-space models can be strong downstream learners, arXiv. 2024. https://arXiv.org/abs/2406.00209.

[31] J.W. Zuo, M. Velikanov, D.E. Rhaiem, et al., Falcon mamba: The first competitive attention-free 7b language model, arXiv. 2024. https://arXiv.org/abs/2410.05355.

[32] B. Peng, E. Alcaide, Q. Anthony, et al., Rwkv: Reinventing rnns for the transformer era, arXiv. 2023. https://arXiv.org/abs/2305.13048.

[33] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv. 2014. https://arXiv.org/abs/1409.0473.

[34] J. Liu, K.H. Lim, R.K.-W. Lee, et al., Towards Objective and Unbiased Decision Assessments with LLM-Enhanced Hierarchical Attention Networks, arXiv. 2024. https://arXiv.org/abs/2411.08504.

[35] L. Yi, X. Zhou, W. He, et al., LongHeads: Multi-Head Attention is Secretly a Long Context Processor, arXiv. 2024. https://arXiv.org/abs/2402.10685.

[36] G. Xiao., J. Tang, J. Zuo, et al., DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads, arXiv. 2024. https://arXiv.org/abs/2410.10819.

[37] Z. Liu, Q. Song, Q.C. Xiao, et al., FFSplit: Split Feed-Forward Network For Optimizing Accuracy-Efficiency Trade-off in Language Model Inference, arXiv. 2024. https://arXiv.org/abs/2401.04044.

[38] T. Lin, Y. Wang, X. Liu, et al., A survey of transformers, AI open 3 (2022) 111–132.

[39] Y. Dong, J.-B. Cordonnier, A. Loukas, Attention is not all you need: Pure attention loses rank doubly exponentially with depth, International Conference on Machine Learning, PMLR. (2021) 2793–2803.

[40] A. Haviv, O. Ram, O. Press, et al., Transformer language models without positional encodings still learn positional information, arXiv. 2022. https://arXiv.org/abs/2203.16634.

[41] S. Sukhbaatar, A. Szlam, J. Weston, et al., End-to-end memory networks, arXiv. 2015. https://arXiv. org/abs/1503.08895.

[42] P. Shaw, J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, arXiv. 2018. https://arXiv.org/abs/1803.02155.

[43] N.-Q. Pham, T.-L. Ha, T.-N. Nguyen, et al., Relative Positional Encoding for Speech Recognition and Direct Translation, arXiv. 2020. https://arXiv.org/abs/1803.02155.

[44] K. Wu, H. Peng, M. Chen, et al., Rethinking and Improving Relative Position Encoding for Vision Transformer, Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10033–10041.

[45] O. Press, N.A. Smith, M. Lewis, et al., Train Short, Test Long: Attention with linear biases enables input length extrapolation, arXiv. 2021. https://arXiv.org/abs/2108.12409.

[46] J. Su, Y. Lu, S. Pan, et al., Roformer: Enhanced transformer with rotary position embedding, Neurocomputing 568 (2024), 127063.

[47] Z. Dai, Z. Yang, Y. Yang, et al., Transformer-xl: Attentive language models beyond a fixed-length context, arXiv. 2019. https://arXiv.org/abs/1901.02860.

[48] P. He, X. Liu, J. Gao, et al., Deberta: Decoding-enhanced bert with disentangled attention, arXiv. 2020. https://arXiv.org/abs/2006.03654.

[49] C. Raffel, N. Shazeer, A. Roberts, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 1–67.

[50] H. Zhou, F. Liu, B. Gu, et al., A survey of large language models in medicine: Progress, application, and challenge, arXiv. 2023. https://arXiv.org/abs/2311.05112.

[51] T.B. Brown, B. Mann, N. Ryder, et al., Language Models are Few-Shot Learners, NeurIPS. 33 (2020) 1877–1901.

[52] A. Chowdhery, S. Narang, J. Devlin, et al., Palm: Scaling language modeling with pathways, J. Mach. Learn. Res. 24 (2023) 1–113.

[53] C. Zhang, S. Bengio, M. Hardt, et al., Understanding deep learning (still) requires rethinking generalization, Commun. Acm. 64 (2021) 107–115.

[54] N. Mehrabi, F. Morstatter, N. Saxena, et al., A survey on bias and fairness in machine learning, Acm. Comput. Surv. 54 (2021) 1–35.

[55] T. Li, S. Shetty, A. Kamath, et al., CancerGPT for few shot drug pair synergy prediction using large pretrained language models, NPJ. Digit. Med. 7 (2024), 40.

[56] A. Wettig, T. Gao, Z. Zhong, et al., Should you mask 15% in masked language modeling? arXiv. 2022. https://arXiv.org/abs/2202.08005.

[57] Z. Du, Y. Qian, X. Liu, et al., Glm: General language model pretraining with autoregressive blank infilling, arXiv. 2021. https://arXiv.org/abs/2103.10360.

[58] H. An, Y. Chen, Z. Sun, et al., SentenceVAE: Enable next-sentence prediction for large language models with faster speed, higher accuracy and longer context, arXiv. 2024. https://arXiv.org/abs/2408.00655.

[59] A. Bindal, S. Ramanujam, D. Golland, et al., Improved Content Understanding With Effective Use of Multi-task Contrastive Learning, arXiv. 2024. https://arXiv.org/abs/2405.11344.

[60] Y. Ling, X. Jiang, Y. Kim, MALLM-GAN: Multi-Agent Large Language Model as Generative Adversarial Network for Synthesizing Tabular Data, arXiv. 2024. https://arXiv.org/abs/2406.10521.

[61] S.A. Sahoo, Meta-Learning for Large Language Models: Teaching LLMs to Learn New Tasks with Minimal Data, Available at SSRN 4977093. 2024.

[62] H. Naveed, A.U. Khan, S. Qiu, et al., A comprehensive overview of large language models, arXiv. 2023. https://arXiv.org/abs/2307.06435.

[63] J. Devlin, M.-W. Chang, K. Lee, et al., Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv. 2019. https://doi.org/10.18653/v1/N19-1423.

[64] Y. Liu, M. Ott, N. Goyal, et al., Roberta: A robustly optimized bert pretraining approach, arXiv. 2019. https://arXiv.org/abs/1907.11692.

[65] K. Clark, M.-T. Luong, Q.V. Le, et al., ELECTRA: Pre training text encoders as discriminators rather than generators, in Proc. 8th ICLR. Addis Ababa, Ethiopia, 2020, pp. 118.

[66] H. Wang, J. Li, H. Wu, et al., Pre-trained language models and their applications, Engineering 25 (2023), 51–65.

[67] A. Liu, B. Feng, B. Xue, et al., Deepseek-v3 technical report, arXiv. 2024. https://arXiv.org/abs/2412.19437.

[68] Vicuna, An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys. org/blog/2023-03-30-vicuna. (Accessed 20 December 2024).

[69] R. Taori, I. Gulrajani, T. Zhang, et al., Stanford alpaca: An instruction-following llama model, 2023. https://github.com/tatsu-lab/stanford_alpaca.

[70] A.Q. Jiang, A. Sablayrolles, A. Mensch, et al., Mistral 7B, arXiv. 2023. https://arXiv.org/abs/2310.06825.

[71] H. Touvron, T. Lavril, G. Izacard, et al., Llama: Open and efficient foundation language models, arXiv.2023. https://arXiv.org/abs/2302.13971.

[72] H. Touvron, L. Martin, K. Stone, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv. 2023. https:// arXiv.org/abs/2307.09288.

[73] B. Cohen-Wang, H. Shah, K. Georgiev, et al., Contextcite: Attributing model generation to context, NeurIPS. 37 (2024) 95764–95807.

[74] J. Bai, S. Bai, Y. Chu, et al., Qwen technical report, arXiv. 2023. https://arXiv.org/abs/2309.16609.

[75] H.W. Chung, L. Hou, S. Longpre, et al., Scaling instruction-finetuned language models, J. Mach. Leaen. Res. 25 (2024) 1–53.

[76] A. Mihalache, J. Grad, N.S. Patil, et al., Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment, Eye. 38 (2024) 2530–2535.

[77] L. Ouyang, J. Wu, X. Jiang, et al., Training language models to follow instructions with human feedback, NeurIPS. 35 (2022) 27730–27744.

[78] J. Achiam, S. Adler, S. Agarwal, et al., Gpt-4 technical report, arXiv. 2023. https:// arXiv.org/abs/2303.08774.

[79] R. Kurokawa, Y. Ohizumi, J. Kanzawa, et al., Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases, Jpn. J. Radiol. 42 (2024) 1399–1402.

[80] M. Lewis, Y. Liu, N. Goyal, et al., Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv. 2019. https://arXiv.org/abs/1910.13461.

[81] A. Zeng, X. Liu, Z. Du, et al., Glm-130b: An open bilingual pre-trained model, arXiv. 2022. https://arXiv.org/abs/2210.02414.

[82] P.M. Abhinav, R.M. SujayKumar, O. Christopher, Machine Translation with Large Language Models: Decoder Only vs. Encoder-Decoder. arXiv. 2024. https://arXiv.org/abs/2409.14747.

[83] Y. Tay, M. Dehghani, V.Q. Tran, et al., Ul2: Unifying language learning paradigms, arXiv. 2022. https://arXiv.org/abs/2205.05131.

[84] L. Wu, Z. Zheng, Z. Qiu, et al., A survey on large language models for recommendation, World Wide Web. 27 (2024), 60.

[85] J. Wei, M. Bosma, V.Y. Zhao, et al., Finetuned language models are zero-shot learners, arXiv. 2021. https://arXiv.org/abs/2109.01652.

[86] V. Sanh, A. Webson, C. Raffel, et al., Multitask prompted training enables zero-shot task generalization, arXiv. 2021. https://arXiv.org/abs/2110.08207.

[87] S. Iyer, X.V. Lin, R. Pasunuru, et al., Opt-iml: Scaling language model instruction meta learning through the lens of generalization, arXiv. 2022. https://arXiv.org/abs/2212.12017.

[88] X.L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, arXiv. 2021. https://arXiv.org/abs/2101.00190.

[89] D. Yin, L. Hu, B. Li, et al., Adapter is all you need for tuning visual tasks. arXiv. 2023. https://arXiv.org/abs/2311.15010.

[90] N. Houlsby, A. Giurgiu, S. Jastrzebski, et al., Parameter-efficient transfer learning for NLP, Proceedings of the 36 th International Conference on Machine Learning, Long Beach, California, 2019, pp. 2790–2799.

[91] E. Hu, Y. Shen, P. Wallis, et al., Lora: Low-rank adaptation of large language models, arXiv. 2021. https://arXiv.org/abs/2106.09685.

[92] M. Nikdan, S. Tabesh, E. Crnčević, et al., Rosa: Accurate parameter-efficient fine-tuning via robust adaptation, arXiv. 2024. https://arXiv.org/abs/2401.04679.

[93] J. Rasley, S. Rajbhandari, O. Ruwase,et al., Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters, Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, 2020, pp. 3505–3506.

[94] M. Shoeybi, M. Patwary, R. Puri, et al., Megatron-lm: Training multi-billion parameter language models using model parallelism, arXiv. 2019. https://arXiv.org/abs/1909.08053.

[95] D. Narayanan, M. Shoeybi, J. Casper, et al., Efficient large-scale language model training on gpu clusters using megatron-lm, arXiv. 2021. https://arXiv.org/abs/zenodo.5181820.

[96] V. Korthikanti, J. Casper, S. Lym, et al., Reducing activation recomputation in large transformer models, Proceedings of Machine Learning and Systems, 5 (2023) 341–353.

[97] D. Gao, L. Ji, L. Zhou, et al., Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn, arXiv. 2023. https://arXiv.org/abs/2306.08640.

[98] P. Lu, B. Peng, H. Cheng, et al., Chameleon: Plug-and-play compositional reasoning with large language models, arXiv. 2023. https://arXiv.org/abs/2304.09842.

[99] B. Paranjape, S. Lundberg, S. Singh, et al., Art: Automatic multi-step reasoning and tool-use for large language models, arXiv. 2023. https://arXiv.org/abs/2303.09014.

[100] J. Kim, M. Min, From rag to qa-rag: Integrating generative ai for pharmaceutical regulatory compliance process, arXiv. 2024. https://arXiv.org/abs/2402.01717.

[101] C. Zakka, A. Chaurasia, R. Shad, et al., Almanac-retrieval-augmented language models for clinical medicine, NEJM. AI. 1 (2024), AIoa2300068.

[102] D. Ferber, I.C. Wiest, G. Wölflein, et al., GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines, NEJM. AI. 1 (2024), AIcs2300235.

[103] C. Ma, Z. Wu, J. Wang, et al., An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT, IEEE Trans. Artif. Intell. 8 (2024) 4163–4175.

[104] Y. Guo, W. Qiu, G. Leroy, et al., Retrieval augmentation of large language models for lay language generation, J. Biomed. Inform. 149 (2024), 104580.

[105] C. Wang, M. Li, J. He, et al., A survey for large language models in biomedicine, arXiv. 2024. https://arXiv.org/abs/2409.00133.

[106] H. Wang, C. Liu, N. Xi, et al., Huatuo: Tuning llama model with chinese medical knowledge, arXiv. 2023. https://arXiv.org/abs/2304.06975.

[107] S. Yang, H. Zhao, S. Zhu, et al., Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue, Proceedings of the AAAI conference on artificial intelligence 38 (2024) 19368–19376.

[108] Y. Kang, Y. Chang, J. Fu, et al., CMLM-ZhongJing: Large language model is good story listener, GitHub Repository 2023. https://github.com/pariskang/CMLM-ZhongJing. (Accessed 20 May 2025).

[109] Y. Chen, Z. Wang, H. Zheng, et al., Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt, arXiv. 2023. https://arXiv.org/abs/2310.15896.

[110] Chinese Pharmacopoeia Commission. People's Republic of China, 2020 edition. Beijing: The Medicine Science and Tech nology Press of China, 2020.

[111] H. Tian, K. Yang, X. Dong, et al. TCMLLM-PR: evaluation of large language models for prescription recommendation in traditional Chinese medicine, Digital Chinese Medicine 7 (2024) 343–355

[112] H. Zhang, X. Wang, Z. Meng, et al., Qibo: A Large Language Model for Traditional Chinese Medicine, arXiv. 2024. https://arXiv.org/abs/2403.16056.

[113] R. Hua, X. Dong, Y. Wei, et al., Lingdan: enhancing encoding of traditional Chinese medicine knowledge for clinical reasoning tasks with large language models[J]. J. Am. Med. Inform. Assn. 2024, 31(9) 2019–2029.

[114] Y. Liu, S. Luo, Z. Zhong, et al., Hengqin-RA-v1: Advanced Large Language Model for Diagnosis and Treatment of Rheumatoid Arthritis with Dataset based Traditional Chinese Medicine, arXiv. 2025. https://arXiv.org/abs/2501.02471.

[115] W. Zhu, W. Yue, X. Wang, ShenNong-TCM: A Traditional Chinese Medicine Large Language Model, GitHub 2023. https://github.com/michael-wzhu/ShenNong-TCM-LLM. (Accessed 20 May 2025).

[116] C. Liu, K. Sun, Q. Zhou, et al., CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions, Sci. rep.14 (2024), 6403.

[117] Y. Liao, S. Jiang, Y. Wang, et al. MING-MOE: Enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts, arXiv. 2024. https://arXiv.org/abs/2404. 09027.

[118] R. Wang, Y. Duan, C. Lam, et al. Ivygpt: Interactive chinese pathway language model in medical domain, CAAI International Conference on Artificial Intelligence. ingapore: Springer Nature, Singapore, 2023, pp. 378–382.

[119] L. Gao, C.-H. Jia, W. Wang, Recent advances in the study of ancient books on traditional Chinese medicine, World J. Tradit. Chin. Med. 6 (2020) 61–66.

[120] I. Abdelaziz, A. Fokoue, O. Hassanzadeh, et al., Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions, J. Web. Semant. 44 (2017) 104–117.

[121] J. Zhang, S. Yang, J. Liu, et al., AIGC Empowering the Revitalization of Ancient Books on Traditional Chinese Medicine: Building the Huang-Di Large Language Model, Library Tribune 44 (2024) 103–112.

[122] W. Dai, J. Lin, F. Jin, et al., Can large language models provide feedback to students? A case study on ChatGPT, ICALT. IEEE. (2023) 323–325.

[123] J.C. Young, M. Shishido, Investigating OpenAI's ChatGPT potentials in generating Chatbot's dialogue for English as a foreign language learning, Int. J. Adv. Comput. Sci. Appl. 14 (2023) 65–72.

[124] S. Kim, P.A. Thiessen, E.E. Bolton, et al., PubChem substance and compound databases, Nucleic. Acids. Res. 44 (2016) D1202–D1213.

[125] S.R. Lynch, T. Bothwell, L. Campbell, A comparison of physical properties, screening procedures and a human efficacy trial for predicting the bioavailability of commercial elemental iron powders used for food fortification, Int. J. Vitam. Nutr. Res. 77 (2007) 107–124.

[126] T. Andrysek, Impact of physical properties of formulations on bioavailability of active substance: current and novel drugs with cyclosporine, Mol. Immunol. 39 (2003) 1061–1065.

[127] H.C.S. Chan, H. Shan, T. Dahoun,et al., Advancing drug discovery via artificial intelligence, Trends. Pharmacol. Sci. 40 (2019) 592–604.

[128] F. Huang, H. Yang, X. Zhu, Progress in the Application of Artificial Intelligence in New Drug Discovery, Progress in Pharmaceutical Sciences 45 (2021) 502-511.

[129] Z. Liu, R.A. Roberts, M. Lal-Nag, et al., AI-based language models powering drug discovery and development, Drug Discov. Today 26 (2021) 2593–2607.

[130] L. Liang, C. Deng, Y. Zhang, et al., Application and Challenges of Artificial Intelligence in Drug Discovery, Progress in Pharmaceutical Sciences 44 (2020) 18–27.

[131] L. Patel, T. Shukla, X. Huang, et al., Machine learning methods in drug discovery, Molecules. 25 (2020), 5277.

[132] J.M. Stokes, K. Yang, K. Swanson, et al. A deep learning approach to antibiotic discovery, Cell. 180 (2020) 688–702.

[133] T. Wu, R. Lin, P. Cui, et al. Deep learning-based drug screening for the discovery of potential therapeutic agents for Alzheimer's disease, J. Pharm. Anal. 14 (2024), 101022.

[134] X. Yang, A.K. Chen, N. PourNejatian, et al., A large language model for electronic health records, Npj. Digit. Med. 5 (2022), 194.

[135] S.B. Patel, K. Lam, ChatGPT: the future of discharge summaries? Lancet Digit. Health 5 (2023) e107–e108.

[136] B. Fatani, ChatGPT for future medical and dental research, Cureus. 15 (2023), e37285.

[137] W. Ma, M. Meng, H. Dai, et al., A Comprehensive Review of the Applications of Large Language Models in Clinical Medicine with ChatGPT as a Representative, Journal of Medical Intelligence 44 (2023) 9–17.

[138] R. Khera, A.J. Butte, M. Berkwits, et al., AI in medicine—JAMA's focus on clinical outcomes, patient-centered care, quality, and equity, Jama-J. Am Med. Assoc. 330 (2023) 818–820.

[139] W. Yan, J. Hu, H. Ceng, et al., The Application of Large Language Models in Primary Healthcare Services and the Challenges, Chinese General Practice 28 (2025) 1–6.

[140] L. Li, Y. Fan, M. Tse, et al. A review of applications in federated learning, Compu. Ind. Eng. 149 (2020), 106854.

[141] X. Du, T. Gunter, X. Kong, et al. Revisiting MoE and Dense Speed-Accuracy Comparisons for LLM Training, arXiv. 2024. https://arXiv.org/abs/2405.15052.

[142] Y. Bao, H. Ding, Z. Zhang, et al., Intelligent Acupuncture: Data-driven Revolution of Traditional Chinese Medicine, Acupuncture and Herbal Medicine 3 (2023) 271–284.

[143] X. Liu, T. Gong, Artificial Intelligence and Evidence-Based Research Will Promote the Development of Traditional Medicine, Acupuncture and Herbal Medicine 4 (2024) 134–135.

[144] C. Chakraborty, M. Bhattacharya, S.-S. Lee, Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development, Mol. Ther. Nucleic Acids 33 (2023) 866–868.

[145] S. Liu, H. Wang, W. Liu, et al. Pre-training molecular graph representation with 3d geometry, arXiv. 2021. https://arXiv.org/abs/ 2110.07728

[146] Y. Li, X. Liu, J. Zhou, et al. Artificial intelligence in traditional Chinese medicine: advances in multi-metabolite multi-target interaction modeling, Front. Pharmacol. 16 (2025), 1541509.

Figures captions:

Fig. 1. Summary of the data in the field of traditional Chinese medicine (TCM). TCM data has rich types, mainly covering the following four aspects: experience data, pharmacological data, chemical data, and clinical data. The sources of these data are very extensive, including ancient books, experts, experiments, public databases, and literature. Their types show a high degree of diversity, and the formats also have multimodal characteristics, involving various forms such as text, image, and table.

Fig. 2. Overview analysis diagram of intelligent question-answering (QA) systems based on large language models (LLMs). (A) Co-occurrence of keywords. (B) Emergent map of key words in TCM-LLMs.

Fig. 3. Background of large language models (LLMs) development. (A) Open source LLMs. (B) Closed source LLMs. WE: word embedding; PT: pre-training; FT: fine-tuning; RAG: retrieval-augmented generation; NLP: natural language processing.

Fig. 4. Working principle of large language models (LLMs) and training methods. (A) Diagram of Transformer architecture. (B) Three training methods of LLMs. RAG: retrieval-augmented generation; Norm: normalization.

Fig. 5. Prospects for the application of traditional Chinese medicine (TCM)-large language models (LLMs).

Fig. 6. Comparison of intelligence traditional Chinese medicine (TCM) question-answering (QA) systems with general large language models (LLMs). (A) Intelligence TCM QA systems based on LLMs. (B) General LLMs.

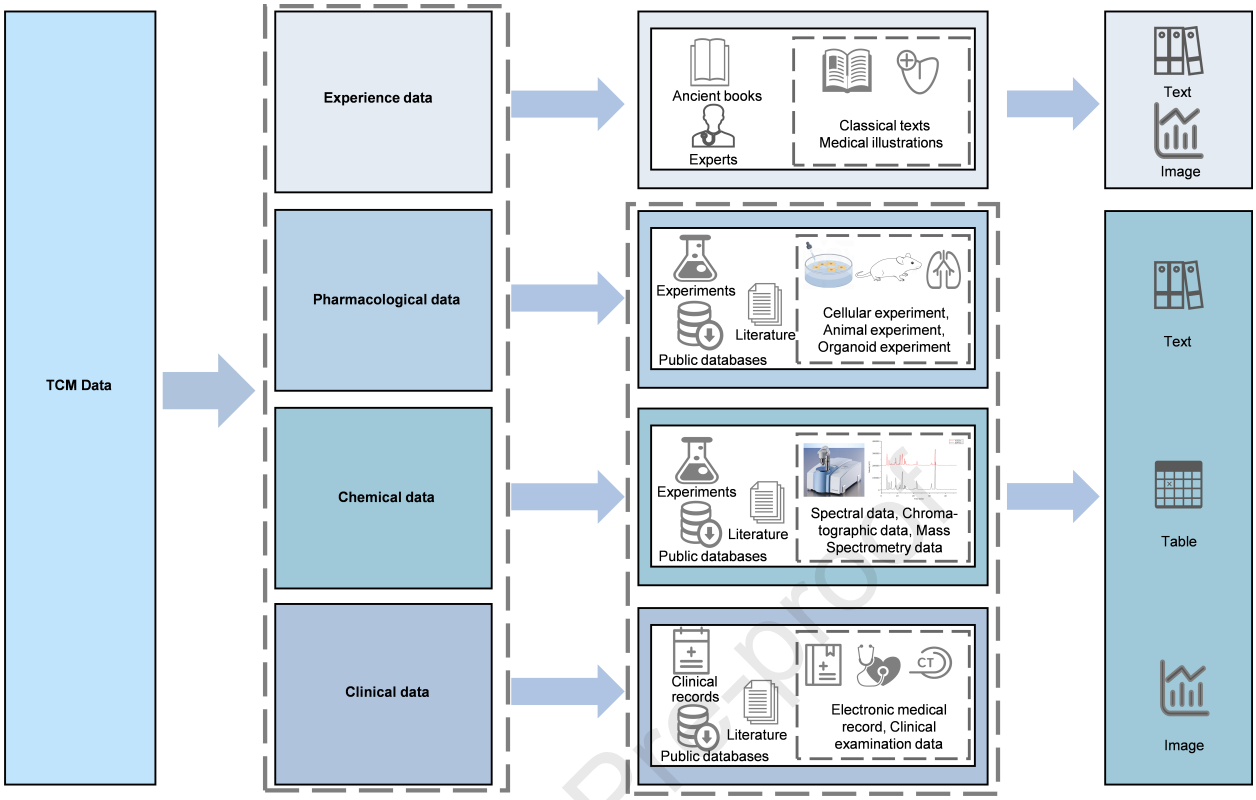
Acknowledgments

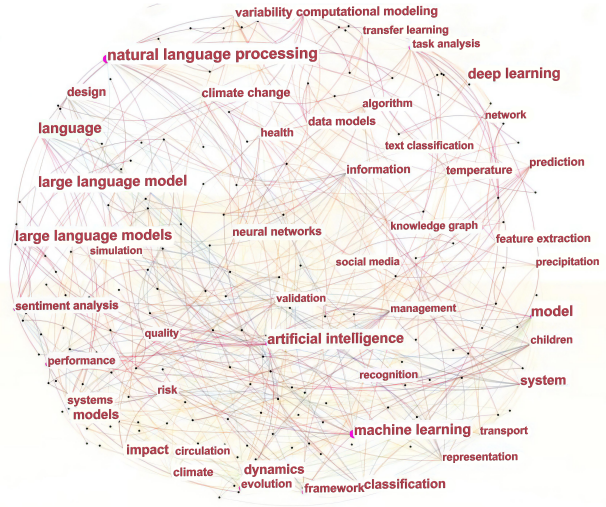
This work was supported by the Special Project for Technological Innovation in New Productive Forces of Modern Chinese Medicines (Grant No.: 24ZXZKSY00010), Science and Technology Program of Tianjin (Grant No.: 24ZXZSSS00460).

Table 1. Comparison of the model structure and Pre-Training (PT) parameters of general large language models (LLMs).

Num	Model name	Model structure	Parameters	Refs.
1	GPT-2	Decoder only	1.5B	[27]
2	DeBERTa	Encoder only	1.5B	[48]
3	T5	Encoder and decoder	11B	[49]
4	PaLM	Decoder only	8B/62B/540B	[52]
5	GPT-3	Decoder only	6.7B/13B/175B	[62]
6	BERT	Encoder only	110M/340M	[63]
7	RoBERTa	Encoder only	355M	[64]
8	ELECTRA	Encoder only	335M	[65]
9	XLNet	Encoder only	360M	[65]
10	CPM	Encoder only	2.6B	[66]
11	CPM-2	Encoder only	11B	[66]
12	GLaM	Encoder only	1.2 trillion	[66]
13	Gopher	Decoder only	280B	[66]
14	DeepSeek-V3	Decoder only	671B	[67]
15	Vicuna	Decoder only	7B/13B	[68]
16	Alpaca	Decoder only	7B/13B	[69]
17	Mistral	Decoder only	7B	[70]
18	LLaMA	Decoder only	7B/13B/33B/65B	[71]
19	LLaMA-2	Decoder only	7B/13B/34B/70B	[72]
20	LLaMA-3	Decoder only	8B/70B	[73]
21	Qwen	Decoder only	1.8B/7B/14B/72B	[74]
22	FLAN-PaLM	Decoder only	540B	[75]
23	Gemini	Decoder only	–	[76]
24	GPT-3.5	Decoder only	–	[77]
25	GPT-4	Decoder only	–	[78]
26	Claude-3	Decoder only	–	[79]
27	BART	Encoder and decoder	140M/400M	[80]
28	GLM	Encoder and decoder	130B	[81]
29	mT5	Encoder and decoder	300M	[82]
30	UL2	Encoder and decoder	19.5B	[83]

–: no data; M: Million; B: Billion.



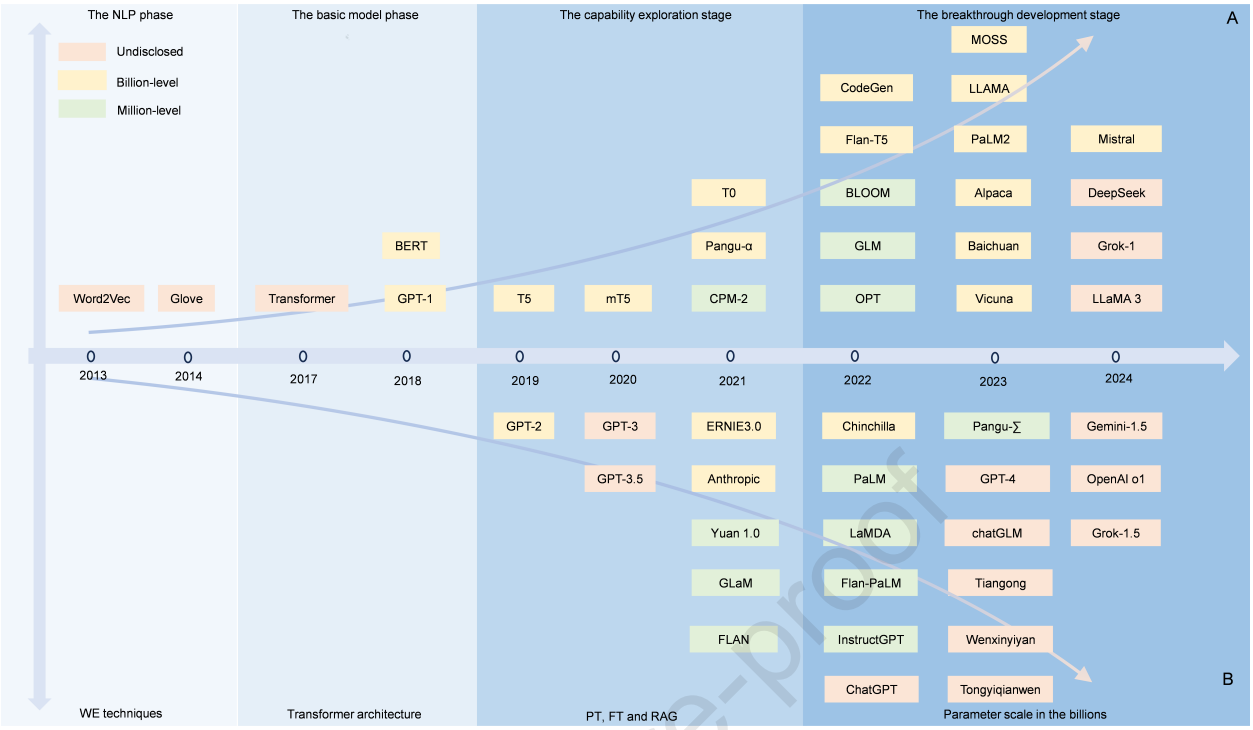


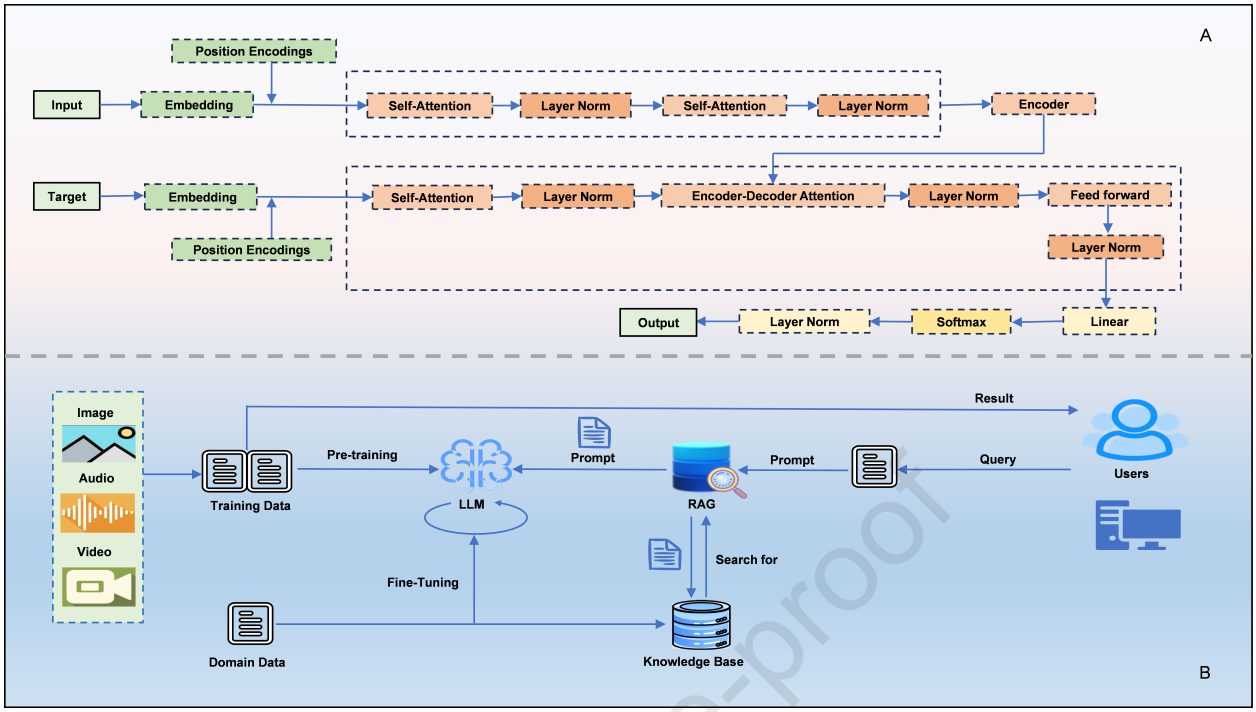
A

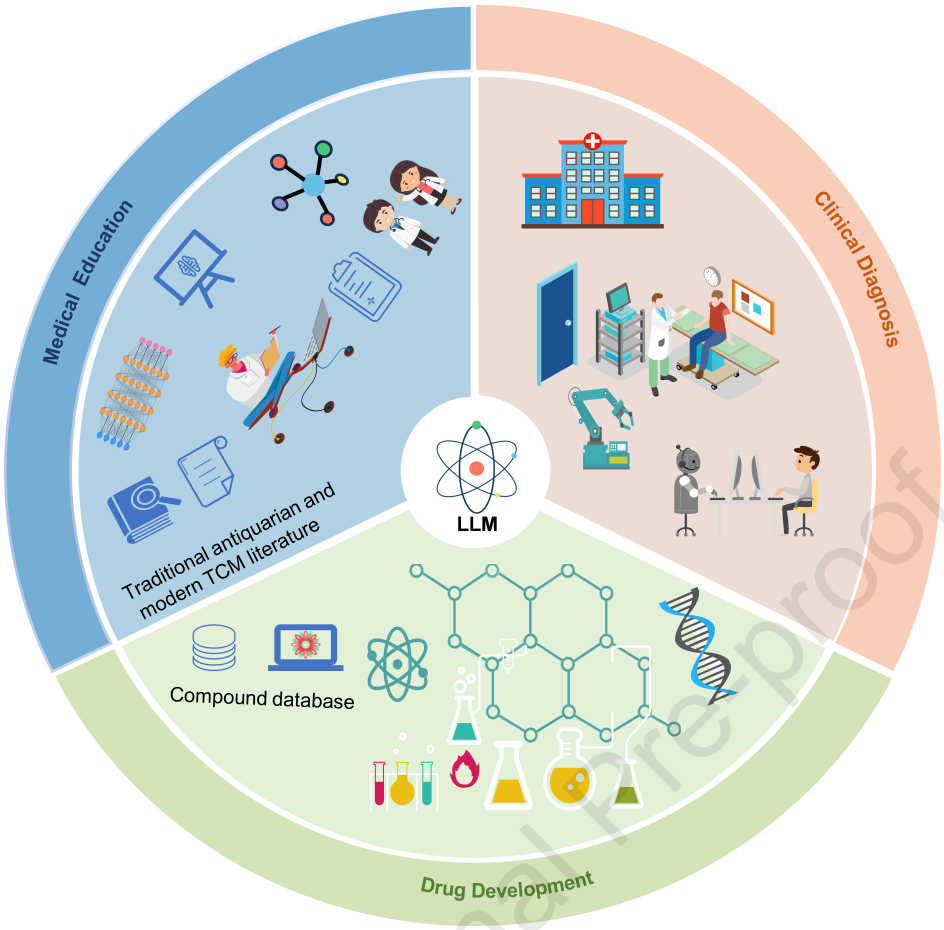
Top 25 Keywords with the Strongest Citation Bursts

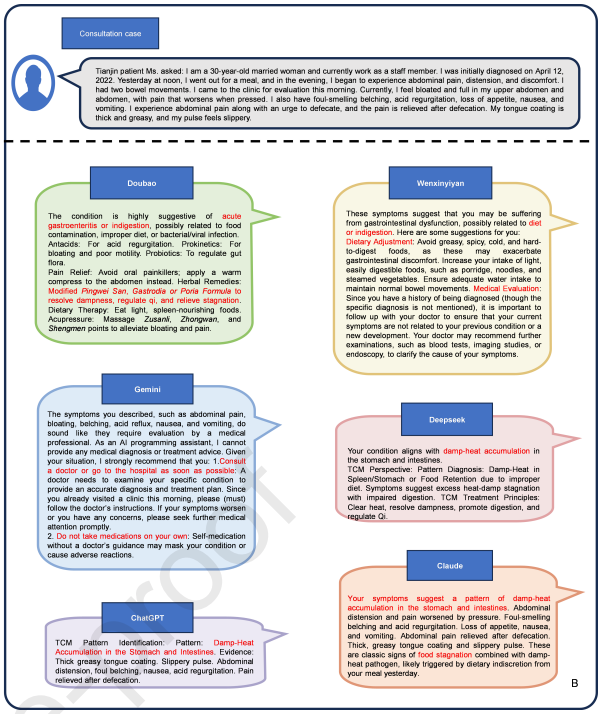
Keywords	Years	Strength	Begin	End	2020-2025
Temporary employment	2020	0.67	2020	2020	
Anxiety disorders	2020	0.67	2020	2020	
individuals	2020	0.67	2020	2020	
Psychosocial work environment	2020	0.67	2020	2020	
Unemployment	2020	0.67	2020	2020	
Marginal structural models	2020	0.67	2020	2020	
Sleep	2020	0.67	2020	2020	
Predictors	2020	0.67	2020	2020	
Psychotropic drugs	2020	0.67	2020	2020	
Metanalysis	2020	0.67	2020	2020	
Job insecurity	2020	0.67	2020	2020	
Depressive symptoms	2020	0.67	2020	2020	
Medical education	2023	0.68	2023	2023	
Question-answering approach	2023	0.62	2023	2023	
Contextual explanations	2023	0.62	2023	2023	
Clinical vignettes	2023	0.62	2023	2023	
Evidence-based dentistry	2023	0.62	2023	2023	
Low kappa	2023	0.62	2023	2023	
Clinical practice	2023	0.62	2023	2023	
Corpus	2023	0.62	2023	2023	
Clinical decision-making	2023	0.62	2023	2023	
Advanced terminology	2023	0.62	2023	2023	
High agreement	2023	0.62	2023	2023	
Interrater reliability	2023	0.62	2023	2023	
Deep learning	2024	0.62	2023	2024	

B









Highlights

- This article is the first systematic review of TCM-LLMs.
- Discussed the characteristics of LLMs in four different stages of development.
- Summarized and compared the working principles and key technologies of LLMs.
- Evaluated the advantages and limitations of open-source and closed-source TCM-LLMs.
- Discussed the prospects and potential impact of TCM-LLM applications.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

CRediT authorship contribution statement

Qilan Xu: Writing – original draft, Methodology, Investigation, Formal analysis. **Tong Wu**: Writing – original draft, Methodology, Investigation, Formal analysis. **Yiwen Wang**: Methodology, Investigation. **Xingyu Li**: Formal analysis. **Heshui Yu**: Conceptualization. **Shixin Cen**: Writing – review & editing, Visualization, Supervision. **Zheng Li**: Conceptualization, Funding acquisition, Writing – review & editing.