IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 1 December 2024, accepted 20 December 2024, date of publication 31 December 2024, date of current version 13 January 2025.

Digital Object Identifier 10.1109/ACCESS.2024.3524588

TOPICAL REVIEW

A Review of Applying Large Language Models in Healthcare

QIMING LIU[1], RUIRONG YANG[2,3], QIN GAO[4], TENGXIAO LIANG[2], XIUYUAN WANG[5], SHIJU LI[6], BINGYIN LEI[1], AND KAIYE GAO[4,7], (Member, IEEE)

[1]School of Economics and Management, Beijing Information Science and Technology University, Beijing 100192, China
[2]Dongzhimen Hospital, Beijing University of Chinese Medicine, Beijing 100029, China
[3]School of Psychology, University of Glasgow, G12 8QQ Glasgow, U.K.
[4]School of Economics and Management, Beijing Forestry University, Beijing 100091, China
[5]Changzhou Liu Guojun Vocational Technology College, Changzhou 213025, China
[6]School of Economics and Management, Beijing Jiaotong University, Beijing 100191, China
[7]Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hong Kong

Corresponding authors: Ruirong Yang (yangruirong365@gmail.com), Tengxiao Liang (13601133923@163.com), and Kaiye Gao (kygao@foxmail.com)

This work was supported in part by Beijing Social Science Fund Project under Grant 22GLC055; in part by Hong Kong Scholar Program under Grant XJ2022050; in part by the Pilot Project for Enhancing Clinical Research and Translational Capacity at Dongzhimen Hospital, Beijing University of Chinese Medicine under Grant DZMG-MLZY-25008; and in part by the Cultivation Program for the Fifth Batch of National Excellent Talents in Clinical Chinese Medicine.

ABSTRACT In response to the growing demand for healthcare and the increasing importance people place on medical services, efficiently meeting these needs within the constraints of limited healthcare resources is of great social and economic benefit. Therefore, research into applying Large Language Models (LLMs) in the healthcare sector holds significant importance. This paper provides a review of the research progress on the application of LLMs in the healthcare field. First, the basic framework of LLMs is summarized, and the training process of LLMs in healthcare is systematically reviewed. Next, six specific application areas of LLMs in healthcare are reviewed: disease diagnosis and decision support, dissemination of medical knowledge, medical assistance, medical image analysis, biomedicine, and medical education. Then, several representative healthcare-specific LLMs are discussed, along with their performance analysis. Following this, the challenges faced by LLMs in healthcare are summarized, and relevant suggestions are provided. The future development trends of LLMs in healthcare are also explored. Finally, a bibliometric analysis is performed. Through the literature review, we found: 1) After pretraining, LLMs are widely adaptable to downstream tasks, significantly enhancing processing performance and efficiency; 2) LLMs in healthcare possess multiple capabilities and can handle multimodal data; 3) Bibliometric analysis shows that researchers are paying increasing attention to the application of LLMs in healthcare; 4) Further research is needed in optimizing, improving reliability, and expanding practical applications of large healthcare models.

INDEX TERMS Healthcare, large language models, literature bibliometric, medical.

I. INTRODUCTION

As the population continues to grow and the aging process accelerates, the number of patients in the healthcare sector is also steadily increasing. The high incidence of chronic diseases and frequent outbreaks of influenza have drawn increasing attention to healthcare issues, with people placing greater importance on life and health [1]. With continuous economic progress, people have higher expectations and a wider variety of needs in healthcare. Compared to other disciplines, medicine has inherent uncertainties [2], and the large volume and complexity of information involved in healthcare make its practice challenging, driving the advancement of relevant technological capabilities.

Artificial Intelligence (AI) has emerged as a powerful tool with transformative potential in healthcare [3], [4], [5], [6],

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

[7], [8]. However, AI still faces significant challenges, such as difficulties in using medical data, low data standardization, challenges in handling complex cases, and insufficient intelligent applications [9]. Large Language Models (LLMs), such as ChatGPT, with their natural language processing and knowledge retention capabilities, can effectively address healthcare-related issues.

Throughout 2023, the release of LLMs has been on the rise, with some specifically designed for healthcare applications [10]. For instance, Surovková et al. [11] have noted that the introduction of AI language models, such as ChatGPT-4, is changing patient-office communication and transforming the role of orthodontic nurses. Therefore, this paper holds significant theoretical and practical value by reviewing the applications of LLMs in healthcare, enhancing the efficiency in addressing healthcare challenges, and laying the foundation for future development.

As a result, various studies concerning LLMs in healthcare have surfaced in recent years.

However, there is a lack of a more systematic and comprehensive review of the application of LLMs in healthcare.

Tian et al. [12] explored the transformative potential of Artificial Intelligence in improving healthcare by addressing historical challenges related to manual image interpretation. Their focus was on the profound impact of LLMs in medical image processing, without covering other healthcare areas. Xiong et al. [13] reviewed and discussed the application of machine learning in health big data. Although they addressed multimodal health big data, their work utilized machine learning methods rather than LLMs. Krishnan et al. [14] reviewed self-supervised methods and models used in medicine and healthcare, while Xue et al. [15] provided an overview of Deep Learning applications in healthcare. Neither study reviewed LLMs. Cascella et al. [10] reviewed research with a focus on recent developments, summarizing the LLMs released over the past year and emphasizing their potential uses in the healthcare field.

Therefore, in order to help designers and researchers to address these challenges and to make potential recommendations for the design and practical application of LLMs in healthcare, this paper provides an overview of LLMs applications in healthcare from six perspectives: (1) LLMs framework; (2) applications in healthcare; (3) Some healthcare Language Models; (4) challenges of LLMs in the field of healthcare; (5) future trends; and (6) bibliometric analysis of the literature. There are several contributions of this work.

Unlike previous reviews, this paper provides a more comprehensive systematic literature review (SLR) of the application of large models in the healthcare field, offering a complete overview of existing knowledge, identifying research gaps, and guiding future research directions. First, this paper presents a more exhaustive review of the various scenarios where LLMs are applied in healthcare, addressing issues only partially resolved by prior reviews. Second, the paper systematically discusses the application of LLMs in healthcare from multiple perspectives and analyzes several representative LLMs in this domain. Next, it identifies key challenges and provides insights into the future development trends in the field. Finally, a bibliometric analysis of relevant literature is conducted, enabling a quantitative evaluation of the scientific literature. This analysis reveals the characteristics, trends, and impact of research across different disciplines, helping to identify new research directions and offering valuable references for the future development of LLMs in healthcare.

The remainder of this paper is arranged as follows. Section II provides an overview of the structure of LLMs. Section III introduces the application of LLMs in the field of healthcare. Section IV lists and analyzes some healthcare language models. Section V addresses the challenges of LLMs in the field of healthcare. Section VI looks forward to future development trends. Section VII includes the bibliometric analysis on the optimization of LLMs. Section VII gives a summary of the article.

II. CONSTRUCTION OF LARGE LANGUAGE MODELS

A. THE MECHANISMS OF LARGE LANGUAGE MODELS

Language Models (LM) aim to model the probability of generating word sequences [16]. They provide a probability distribution for the next character in a given sentence, based on the preceding n characters. The widely used language models currently typically employ a left-to-right approach for predicting words sequentially, that is:

$$P(w_1, w_2, \cdots, w_T) = \prod_{t=1}^{T} P_\theta(w_t | w_0, w_1, \cdots, w_{t-1}) \quad (1)$$

where w_0 represents the start token and w_T represents the end token. After training is completed, the language model can auto-regressively generate text from left to right.

Generally, a Large Language Model (LLM) refers to a language model with over ten billion parameters. Since the introduction of the Transformer, which offers greater computational efficiency and performance compared to Recurrent Neural Networks (RNNs), it has become the foundational architecture for LLMs due to its ability to efficiently handle longer text sequences [17]. Most current mainstream language models are trained based on the Transformer architecture [18].

In this model, Self-Attention (SA) is a core component of the Transformer model, enables the model to weigh the importance of each word in a sequence when generating predictions. SA projects the input sequence into a set of queries Q, keys K, and values V with dimensions C using three learnable linear mapping matrices WQ, WK, and WV. The self-attention weights are then computed using the formula given by (2):

$$\mathrm{SA}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right) V \quad (2)$$

Through the linear transformation of the input sequence, SA is capable of capturing the semantic characteristics and long-range dependencies of the input sequence.

Additionally, the Transformer model utilizes a mechanism called Multi-head Self-Attention (MSA), which is an extension of the self-attention approach and is presented in equation (3). It is composed of n self-attention heads, meaning that the input sequence undergoes multiple linear transformations to yield distinct projection matrices.

$$\mathrm{MSA}\,(Q, K, V) = \mathrm{Concat}\,(\mathrm{SA}_1, \mathrm{SA}_2, \cdots, \mathrm{SA}_n) \times W^0 \quad (3)$$

where W^0 represents the linear transformation matrix, and SA_n denotes the output of the n-th self-attention head.

The Transformer architecture comprises multiple encoders and decoders, where the encoder is composed of a MSA module and a Feed-Forward Neural Network (FFN) module. The MSA module leverages the self-attention mechanism to capture the interrelationships among various elements within the input sequence. The FFN module, on the other hand, consists of two linear layers and an activation function, which is used for further non-linear transformations. Both the MSA and FFN modules employ residual connections and Layer Normalization (LN) techniques to facilitate the training of deep networks [19] and is presented in equation (4).

$$\begin{cases} x_0 = x_0 + x_{pos} \\ y_k = x_{k-1} + \mathrm{MSA}\,(\mathrm{LN}\,(x_{k-1})) \\ x_k = y_k + \mathrm{FFN}\,(\mathrm{LN}\,(y_k)) \end{cases} \quad (4)$$

where x_{pos} represents the positional embedding, and x_k represents the output of the k-th encoder.

The structure of the decoder differs slightly from that of the encoder. The decoder includes an additional layer, and the multi-head self-attention mechanism in the first layer incorporates masking. The second layer features a multi-head cross-attention mechanism [20].

Based on the Transformer's encoder-decoder architecture, unsupervised pre-training using large datasets can be performed to develop foundational models with universal language capabilities. The existing frameworks are broadly categorized into three types: transformer decoders only, transformer encoders only, and transformer encoder-decoder hybrids. As shown in Figure 1, x represents the original input sequence, $x_t (t = 1, 2, \ldots, T)$ denotes the t-th label, T is the sequence length, $M(x)$ represents the mask label of x, and S represents the initial label of the sequence embedding. $p1, p2, p3$ and $p4$ represent the position embedding tags from first to fourth, respectively, while P indicates the conditional probability. i and j denote the start and end indices of the encoder input, respectively [21].

B. TRAINING LARGE LANGUAGE MODELS FOR MEDICAL APPLICATIONS

In general, training LLMs involves five main steps. First, data collection is performed. Second, the data is preprocessed. Third, the model undergoes pre-training using a large corpus

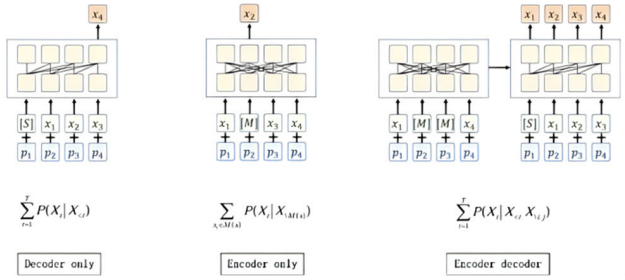

FIGURE 1. Existing prevalent model training frameworks [22].

to acquire broad knowledge. Next, the model is fine-tuned to enable it to handle a variety of tasks effectively. Finally, the model is evaluated and optimized, for example, by using Reinforcement Learning from Human Feedback (RLHF) to align the model with human values [23].

1) DATA COLLECTION

Publicly accessible natural language data comes in a variety of forms [24], [25]. After careful processing, such data can be transformed into large-scale, high-quality, and diverse language datasets, providing the foundation for training LLMs. General-purpose datasets offer rich training data and linguistic knowledge, enhancing the performance and applicability of LLMs. For instance, new models such as GPT-4 and LLaMA are typically pre-trained on datasets containing over a trillion words.

Additionally, synthetic data generated by advanced language models like ChatGPT has been used for training, resulting in language models with impressive performance. This synthetic data is widely used due to its low cost and high quality [26].

As a result, medical large models often utilize sources such as online Q&A, physician-patient dialogues, knowledge bases, publicly available medical datasets, and ChatGPT-based self-instruction methods to construct various instruction datasets [23]. Table 1 provides an overview of popular biomedical datasets commonly used in training medical LLMs, illustrating their diverse content and applications.

2) DATA PREPROCESSING

During the training of LLMs for medical applications, data preprocessing directly impacts the quality and effectiveness of model training. Data cleaning is essential, including the removal of irrelevant, duplicate, and low-quality data. Next, tokenization and part-of-speech tagging are performed, along with constructing a specialized vocabulary tailored for the medical domain. Medical data often include sensitive information, such as patient identifiers and medical history, which require strict anonymization to comply with privacy laws like HIPAA and GDPR. Additionally, unstructured formats, such as free-text clinical notes, demand specialized tokenization and vocabulary building tailored for the medical domain.

TABLE 1. Popular biomedical benchmark datasets.

Name	Main content
MedMCQA [27]	Over 194,000 high quality AIIMS and NEET PG Entrance Exam MCQs, covering 2400 medical topics and 21 medical subjects.
MultiMedQA [28]	It combines six existing open-ended medical Q&A datasets, covering professional medical exams, scientific research, etc.
MIMIC-III [29]	Includes demographic information, laboratory test information, patient medication information, nursing-related information, patient imaging reports, and information on each patient's hospital entry and exit.
Chest X-ray 14 [30]	Contains 112,120 X-ray images with 14 disease labels and medical diagnostic reports from 30,805 patients.
MedicationQA [31]	Questions and answers about medication cover a variety of medication-related topics, including information on medication guidelines, drug dosages, and more.

3) PRE-TRAINING

Most Large LLMs are pre-trained on large, general-purpose datasets. The purpose of pre-training is to allow the model to learn advanced features that can be transferred to specific tasks during fine-tuning. Without pre-training, the model tends to memorize questions and answers mechanically without grasping deeper knowledge.

The pre-training process for medical LLMs involves several steps. First, medical text data—including medical records, literature, and pathology reports—is preprocessed and transformed into numerical representations suitable for model input. The model parameters are then randomly initialized, and the numerical representations of the text are fed into the model. A loss function is used to measure the difference between the model's output and the actual next word in the sentence. This loss function can be adjusted based on specific medical standards and metrics. The model's parameters are then optimized to minimize the loss, repeating the process until the model's output reaches an acceptable accuracy level. Throughout this process, expert knowledge and domain-specific data can be incorporated to enhance the model's performance and applicability in the medical field.

Generally, the higher the quality and the larger the amount of pre-training data used, the better the model's performance. Additionally, to update or expand the model's knowledge base for new datasets or domains, continual pre-training is performed. Continual pre-training is a critical phase for the model's knowledge accumulation, during which most of its understanding of medical concepts is developed [32].

4) FINE-TUNING

General-purpose LLMs, such as ChatGPT, have demonstrated significant potential across various applications and industries. However, despite their excellent performance in general domains, they often underperform in specialized fields due to a lack of domain-specific data, potentially leading to unreliable decisions [33]. For instance, in healthcare, these models may struggle due to insufficient specialized training datasets, which may limit their effectiveness. Simply memorizing medical knowledge to pass medical tests is not equivalent to providing safe and effective clinical services [34].

When performing downstream tasks, pre-trained models can be used to initialize weights for domain-specific fine-tuning. Fine-tuning requires domain-annotated data, enabling pre-trained models to adapt effectively to specialized tasks. By introducing a small number of domain-specific examples, fine-tuning retrieves and activates the knowledge stored in the pre-trained model, aligning it with the requirements of specialized fields [35].

Fine-tuning is a crucial step in transforming LLMs from general-purpose to specialized models by leveraging domain-specific expertise. Medical LLMs differ from general LLMs by focusing specifically on healthcare applications. Fine-tuning in the medical domain involves additional training on diverse datasets, such as medical literature, clinical notes, and diagnostic images. These datasets must address unique challenges, such as handling unstructured clinical texts, ensuring data privacy, and incorporating domain-specific terminology. Through this process, the model learns medical language rules, contextual understanding, and specialized knowledge, enabling it to meet the complex requirements of healthcare tasks. The detailed process is illustrated in Figure 2. By integrating healthcare knowledge, fine-tuned and specialized LLMs can ensure responses meet established medical standards, enhancing their quality and usability in real-world healthcare scenarios. For example, in diabetes care, fine-tuning a model with datasets containing clinical guidelines, patient case studies, and treatment protocols enabled the development of an insulin regimen guidance model. Leveraging pre-trained models like Clinical BERT and NLP techniques led to an insulin regimen guidance model [36], showing promising results in improving response accuracy and relevance, addressing the dynamic needs of diabetes management [37].

Clinical knowledge-encoded and fine-tuned LLMs, such as Clinical BERT, Med-PaLM2, GatorTron, MedGPT, and Huatuo-GPT, have demonstrated significant potential in medical NLP and clinical applications.

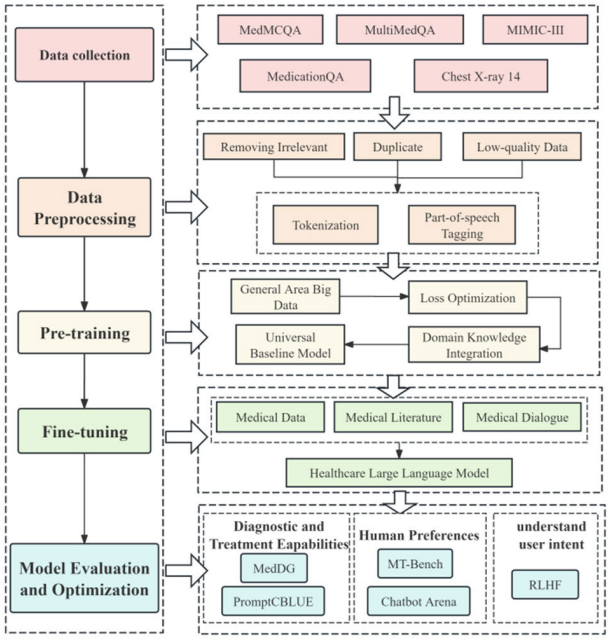

FIGURE 2. The framework for the healthcare large language model.

5) MODEL EVALUATION AND OPTIMIZATION

With the increasing use of medical LLMs, objectively and comprehensively evaluating their medical capabilities has become increasingly important. The core evaluation of LLMs focuses on assessing their demonstrated abilities and requires a multi-faceted examination to align with real medical scenarios.

The capabilities of general LLMs can be categorized into basic and advanced abilities [38]. Basic abilities include language generation, knowledge application, and complex reasoning. Advanced abilities encompass alignment with human values, external environment interaction, and tool usage.

Medical LLMs are applications of general LLMs in healthcare, and their evaluation standards should be similar to those of general models. In evaluating diagnostic and treatment capabilities, the MedDG dataset from PromptCBLUE can assess a model's ability in lightweight diagnostic scenarios based on multi-turn conversations. PromptCBLUE also offers a comprehensive evaluation benchmark for medical LLMs focused on traditional Chinese medicine.

To evaluate alignment with human preferences, MT-Bench and Chatbot Arena [39] are two commonly used benchmarks. MT-Bench assesses LLMs' performance in multi-turn conversations and instruction adherence, while Chatbot Arena enhances the breadth and diversity of evaluation results. The Safety-Prompts dataset can evaluate whether medical LLMs provide positive and fair responses to psychological and mental health issues.

To better understand user intent and ensure the generated content meets user expectations, RLHF is employed. The primary goal is to capture the entirety of human preferences, driving the AI-generated responses to align with human values. This approach further enhances the LLM's performance in real user-facing scenarios, improving healthcare quality, safety, and efficiency while adhering to ethical and legal standards [40].

C. PARADIGMS OF LARGE MEDICAL LANGUAGE MODELS

LLMs undergo an initial training phase in which they learn a comprehensive understanding of language by analyzing diverse texts. This enables them to broadly grasp general language patterns and subtle contextual nuances. Following this, fine-tuning is performed to adapt the pre-trained LLM for specific domains, refining its capabilities for specialized tasks. The model builds foundational language knowledge through pre-training, which is then applied effectively to target tasks after fine-tuning with domain-specific knowledge.

In recent years, to better leverage the features of LLMs while avoiding training all model parameters, a new paradigm focused on prompt learning has gained attention [41]. Prompt learning reduces the gap between pre-training and fine-tuning by minimizing the difference in data formats, enabling more efficient application of LLMs to downstream tasks. Prompts can be categorized as either discrete prompts or continuous prompts. Discrete prompts are typically natural language text, which helps the pre-trained model understand downstream tasks. Continuous prompts, on the other hand, are continuous vectors that are not restricted by word embeddings. These prompts often include a small number of trainable parameters that can be updated during training for downstream tasks [42].

III. APPLYING LARGE LANGUAGE MODELS IN HEALTHCARE

A. DISEASE DIAGNOSIS AND DECISION SUPPORT

The vast and continually expanding volume of medical literature makes it challenging for healthcare professionals to promptly understand and acquire knowledge across various medical fields. The increasing specialization within medical disciplines has also led to cognitive barriers between different areas of expertise [43]. The complexity of today's medical system poses significant challenges for healthcare providers; however, LLMs have the capability to learn contextual information from large datasets, understand complex language structures, and capture subtle nuances, making them well-suited for applications in medical text comprehension and language generation.

Medical LLMs exhibit strong predictive and classification abilities, integrating personalized patient information with extensive medical data to provide individualized diagnostic and treatment plans. This customized clinical decision support can help healthcare professionals improve clinical outcomes. For example, the HuaTuo model, developed by the Chinese University of Hong Kong [44], leverages a fine-tuned Chinese Medical Knowledge Graph (CMeKG) to enhance its performance in medical question answering tasks. Through

instruction fine-tuning, HuaTuo integrates real-world physician feedback and high-quality domain-specific datasets, enabling it to provide contextually accurate and professional responses. For example, in a pilot study evaluating HuaTuo's diagnostic accuracy, the model achieved an improvement of 20% in multi-turn consultations compared to general-purpose LLMs like ChatGPT. This fine-tuning approach addresses the limitations of general LLMs, such as insufficient understanding of medical terminology and inability to handle nuanced medical contexts, thereby demonstrating superior diagnostic support in Chinese medical settings.

B. POPULARIZING MEDICAL KNOWLEDGE

Even though ChatGPT has not been specifically fine-tuned for the healthcare domain, it still demonstrates a substantial ability to handle health-related questions. MedGPT, a healthcare-specific Large Language Model, has shown superior performance in patient education and medical knowledge dissemination. By leveraging multi-turn dialogue training datasets, MedGPT can provide detailed explanations of medical conditions, educate patients about treatment protocols, and address common health concerns in a conversational format. In a deployment within outpatient clinics, MedGPT successfully answered over 85% of patient queries with accuracy similar to that of human medical assistants. Additionally, its ability to simulate real-world patient-doctor interactions has proven valuable in improving health literacy and empowering patients to make informed decisions. Notably, ChatGPT has shown proficiency in the United States Medical Licensing Exam (USMLE), generally achieving passing scores and providing reasonable explanations [45].

Medical cognitive intelligence, based on LLMs, complements human cognitive abilities and facilitates human-machine collaborative cognition, which makes it easier to navigate the increasingly complex medical landscape. Consequently, patients can use these medical language models to access desired medical knowledge, enhancing their health literacy and contributing to the broader dissemination of medical information.

C. MEDICAL ASSISTANT

The time and energy of physicians are limited, and there are challenges in healthcare due to the need for more medical resources and expertise. LLMs can rapidly assimilate, summarize, and restate information, and they can provide primary care diagnostic advice to patients [46]. They are frequently used in chatbots and virtual assistants and possess the capability to serve as medical question-answering assistants, being able to handle specific tasks traditionally performed by general practitioners. For example, the medical variant Med-PaLM, based on PaLM2, can retrieve medical knowledge, answer questions, and decode medical terminology.

The application of large models in the healthcare domain allows for the use of question-and-answer interaction to address common patient inquiries, perform routine disease diagnoses, explain medical conditions, provide medical advice, or search for and recommend similar cases, all in support of assisting physicians in formulating diagnosis and treatment plans.

For example, in applying the DoctorGLM large model in medical dialogue, PromptDesigner was designed to address the issue of long-context problems by extracting relevant keywords from the input, thereby enhancing the precision of the large model's responses [47]. The MedGPT large model, after acquiring a substantial amount of accurate and structured medical knowledge, was trained on real-world medical dialogues between doctors and patients, laboratory test results, and medical records. Subsequently, it underwent fine-tuning using high-quality structured clinical diagnosis and treatment data, and finally, it underwent reinforcement learning through honest physician feedback. This approach has led to the pioneering achievement of continuous free-form dialogue capability between large AI models and actual patients [48]. This model supports multimodal input and output in medical consultations, recommends appropriate medical examination items for patients after the consultation, and devises treatment plans based on the consultation and examination results. As a large model for chronic diseases, ClouD GPT can assist physicians and pharmacists in controlling the quality of prescriptions, thereby enhancing the efficiency and accuracy of physician diagnosis and treatment plans. It has successfully developed the first publicly published clinical study in cardiovascular disease treatment using digital therapy interventions to regulate blood lipids. Based on the Large Language Model ChatGLM2, after fine-tuning with medical knowledge question-and-answer data, it can automatically generate internship physical examination conclusions [49].

LLMs can also alleviate the administrative burden on clinical physicians. Discharge summaries are a more repetitive task involving interpreting and compressing information with little problem-solving [50]. In addition, some technology companies are conducting pilot programs using multimodal large models to read patient information and electronic medical records, and draft physician responses. This approach aims to reduce the time required for medical staff to respond to patients [34]. Emerging multimodal models will further expand capabilities and incorporate more data sources, even enabling the automatic and accurate interpretation of physicians' handwritten notes [51].

D. MEDICAL IMAGE ANALYSIS

Medical imaging refers to non-invasive techniques used to obtain internal images of the human body [52], serving as a critical basis for diagnosing patients. The applications of LLMs in medical image analysis are extensive. These models can assist physicians and researchers by interpreting handwritten notes and prescriptions [53], thereby supporting disease diagnosis, evaluating patient conditions, and guiding the development of treatment plans.

E. BIOMEDICAL ADVANCES IN MEDICINE AND BIOLOGY

LLMs also demonstrate promising prospects in various biomedical applications. They might analyze genomic data to identify patterns that indicate a predisposition to certain diseases or optimize personalized treatment regimens [53] to enhance the level of precision medicine. LLMs can identify potential drug-drug interactions in treatment, thus preventing potential adverse effects [53] and contributing to safer patient care.

LLMs can assist in drug development, such as developing vaccines for infectious diseases [54]. In addition, AlphaFold deduces protein structure from amino acid sequences in protein design and engineering. The ProGen large model can predict the function of protein sequences [55]. Furthermore, the transfer learning capability of LLMs may aid in identifying and analyzing promoter regions in bacterial DNA from newly isolated strains with similar lineages. TSSNote-CyaPromBERT identifies promoter regions in bacterial DNA [56]. Large multimodal models are also used in drug development, particularly for new drug design, to develop new compounds with specific properties [57].

F. MEDICAL EDUCATION

The application of LLMs in medical education is extensive. These models can generate exercises and quiz scenarios for classroom use, helping medical students practice and improve their diagnostic and treatment planning abilities. They provide timely access to cutting-edge and authoritative medical knowledge, encompassing diseases, treatment options, and treatment procedures. Furthermore, LLMs present significant opportunities for medical education, including realistic simulations, digital patients, personalized feedback, assessment methods, and the elimination of language barriers [58]. These advanced technologies create an immersive learning environment, enhancing the efficiency of medical students' education.

The strong performance of GPT-4 and Med-PaLM 2 in medical examinations suggests that LLMs could serve as valuable teaching tools for medical students [51]. The use of LLMs to construct a Traditional Chinese Medicine (TCM) knowledge graph can also provide a solid foundation for the study, research, and application of TCM [59]. Additionally, Galactica (GAL), trained on an extensive and highly curated scientific knowledge base—including over 48 million research papers, textbooks, lecture notes, millions of compounds and proteins, scientific websites, and encyclopedias [60]—is also an effective tool for learning medical knowledge.

IV. LARGE LANGUAGE MODELS IN HEALTHCARE

LLMs have made significant contributions to the healthcare field. This article provides an overview of various medical language models, introducing their features, discussing use cases in healthcare, evaluating their performance, and highlighting areas for improvement.

A. ChatGPT

Although ChatGPT has not been specifically fine-tuned for the healthcare domain, its strong language capabilities enable it to perform effectively in various medical assessments. As such, it is also included in this discussion. ChatGPT [61] is based on a decoder-only autoregressive model that uses an autoregressive approach to generate responses. It predicts the next word based on previously derived words, learning in sequence from the most frequent words in the training data rather than deducing correct answers. Regardless of the factual accuracy of the generated content, it presents fluent and coherent contexts [62].

The updated version of ChatGPT (i.e., ChatGPT-4.0) achieved a 90% accuracy rate on questions from the United States Medical Licensing Exam (USMLE) [63]. In a study by Nori et al. [64], ChatGPT-4.0 was evaluated on four public datasets—MedQA, PubMedQA, MedMCQA, and MMLU—that included content based on medical literature, clinical cases, and user-generated data. The results showed that its accuracy was comparable to that of human experts [65].

B. BioBERT

BioBERT is based on the BERT model and is specifically designed for biomedical text mining tasks such as named entity recognition (NER), relation extraction, and question answering. BioBERT uses an encoder-based BERT word encoding model, where words in the input sequence are randomly masked, and the model predicts the generation probability of the masked words using the unmasked ones. It is pre-trained to learn deep bidirectional representations from unlabeled text by jointly conditioning on both the left and right contexts in all layers. By adding an extra output layer, the pre-trained BERT model can be fine-tuned to create advanced models for tasks like question answering and language inference without substantial modifications to task-specific architectures [66].

BioBERT [67] outperforms BERT and previous state-of-the-art models on three representative biomedical text mining tasks: biomedical named entity recognition (F1 score improvement of 0.62%), biomedical relation extraction (F1 score improvement of 2.80%), and biomedical question answering (mean reciprocal rank (MRR) improvement of 12.24%).

C. PUBMEDGPT

PubMedGPT is a Large Language Model trained to interpret biomedical language, utilizing portions of the Pile dataset, specifically PubMed Abstracts and PubMed Central content. It achieves an accuracy of 50.3% on the MedQA dataset, 74.4% on the PubMedQA dataset, and 95.7% on the BioASQ dataset.

D. HUANG-DI

The Huang-Di [33] medical Large Language Model integrates and utilizes traditional Chinese medicine (TCM)

classical texts, providing diverse knowledge services such as answering questions from classical texts, conducting TCM consultations, and offering health and wellness advice. Built on the Ziya-LLaMA-13B-V1 open-source model, it underwent further pre-training, supervised fine-tuning, and DPO (Direct Policy Optimization) optimization to create a generative dialogue model for TCM literature. During the model's automated evaluation, the training loss function was found to have successfully converged. However, the BLEU and ROUGE scores across different dialogue categories were relatively low, suggesting the model possesses a strong capacity for domain-specific creativity.

The model has several limitations. Future work should focus on harmless fine-tuning of the corpus, removing potentially discriminatory, harmful, or unethical content. In addition, as the Huang-Di model primarily relies on textual information, it currently lacks the capability to process multimodal information, such as classical text images.

E. BIOMEDLM

BioMedLM [68] is an open, cost-effective autoregressive GPT-style model with 2.7 billion parameters, specifically trained on PubMed abstracts and full articles. After fine-tuning, BioMedLM demonstrates strong performance in multiple-choice biomedical question answering. In relation extraction tasks, it has a lower false positive rate and higher accuracy compared to other models, largely due to the absence of irrelevant out-of-domain data. BioMedLM achieved a score of 57.3% on the MedMCQA (development set) and 69.0% in the MMLU medical genetics examination.

F. WINGPT

WiNGPT is a Large Language Model developed specifically for the healthcare sector, aimed at integrating professional medical knowledge, health information, and data to provide intelligent question answering, diagnostic support, and medical information services to improve diagnostic efficiency and healthcare quality [69]. Developed by the Weining Health AI Laboratory, WiNGPT was fully self-developed, from pre-training to fine-tuning. To address the lack of Chinese healthcare evaluation resources, the research team created the WiNEval-MCKQuiz, a medical evaluation dataset consisting of 13,060 multiple-choice questions covering 17 medical subjects. By comparing the model's predicted answers with correct answers, the WiNGPT2-34B-0317-DPO model achieved an accuracy rate of 88.1%.

V. CHALLENGES OF LARGE LANGUAGE MODELS IN THE HEALTHCARE FIELD

A. THE INTERPRETABILITY OF MODELS

The interpretability of healthcare LLMs is often challenging due to their black-box nature, making it difficult to explain the internal decision-making process. Decision-makers such as doctors need to understand the decision process or reasoning behind the LLMs to trust the conclusions and recommendations they provide. Strategies to address this issue include conducting research on the interpretability of LLMs, developing methods and tools to provide explanations and understanding of the decision-making process of LLMs, and developing interpretable techniques and methods suitable for the medical field to ensure the interpretability, understandability, and controllability of model decisions.

B. COMPLEX AND DIVERSE NATURE OF LINGUISTIC DATA

LLMs are trained on text streams and have yet to be designed to handle specific complex data structures [23]. However, electronic health records comprise various documents encompassing structured and unstructured data. Employing LLMs to generate electronic health records necessitates integrating a substantial amount of structured electronic health record data and the relationships between them. Strategies to address this issue include establishing standardized data cleaning and preprocessing procedures; targeted fine-tuning of LLMs based on the specific data structures and relational characteristics of electronic health records; and the construction of a dedicated electronic health record large-scale model using transfer learning, thereby facilitating the integration, sharing, and empowerment of electronic health record data [9].

C. PRIVACY AND DATA SECURITY

Ensuring data privacy and compliance with regulations such as GDPR and HIPAA remains a critical concern. Healthcare LLMs often require large amounts of personal health information data for training and validation. If not properly secured, there is a risk of personal information and other data leakage and misuse, especially when it involves sensitive information such as patient diagnoses, medical records, and imaging data. Strategies to address this issue include using encryption technology to transmit and store private data, anonymizing sensitive data, strictly controlling relevant data access permissions, and implementing security audits and monitoring to identify and address information security vulnerabilities promptly.

D. ETHICAL AND LEGAL REGULATIONS

When using large healthcare language models, there may be discriminatory or unfair outcomes, necessitating the assurance of fairness and equity in model decisions that align with medical ethical requirements. Strategies to address this issue include establishing an ethics committee or relevant expert team to review the training data and algorithms of large models; ethical review and regulation are required to ensure compliance with medical ethical norms and legal requirements while safeguarding patient rights and safety; the development of international standards and industry regulations related to large medical models is essential; and the formulation of technical manuals related to the development, application, and promotion of large medical models.

E. THE MECHANISMS OF LARGE LANGUAGE MODELS

During the training of LLMs, the samples may come from specific sources or types of data, or there may be issues with a lack of diversity in the data, leading to inaccuracies and the generation of deceptive, false, or low-quality information [70], as well as the potential for bias and discrimination. Instances have been observed where LLMs have produced unreliable results, such as erroneous diagnostic outcomes and treatment plans in medical scenarios [71]. Strategies to address this issue include the use of accurate, comprehensive, and unbiased training data, as well as the incorporation of diverse, high-quality data sources. Additionally, human experts can verify and validate the information generated by LLMs to mitigate inaccuracies and biases.

VI. FUTURE TRENDS

A. MULTIMODAL DIAGNOSTICS

In June 2023, Google Research and DeepMind released a multimodal Large Language Model named Med-PaLM [72]. Multimodal refers to the ability of a model to process and utilize different types of information, such as text, images, and audio. Multimodal models can accept multiple types of inputs and produce outputs that are not limited to the format of the input data. In the healthcare domain, multimodal data can include electronic medical records, patient case files, and medical images [73]. Healthcare data often consists of text, image, and numerical data [13], and large multimodal models are highly capable of managing these complex medical data types, solving related issues, and becoming increasingly widespread in healthcare applications.

To more effectively integrate different diagnostic modes, Niu et al. [74] introduced EHR-KnowGen, a multimodal learning model enhanced with external domain knowledge. This model fully integrates the multimodal information from electronic health records (EHRs) into a unified feature space and utilizes LLMs to generate disease diagnoses. Luo et al. [75] suggested that diagnosis systems based on large-scale pre-trained multimodal models could assist dermatologists in developing effective diagnostic and treatment strategies, indicating a transformative era in healthcare.

Multimodal LLMs have also shown significant advancements in analyzing medical images. The research team from the University of Cambridge developed an open-source multimodal medical model named Visual Med-Alpaca [76]. This model is capable of detecting nodules, analyzing lesions, and providing diagnostic recommendations for CT scans. In the radiology domain, an internal radiological image captioning model called Med-GIT was trained, which uses a classifier to determine which medical visual expert is responsible for interpreting an input image. Subsequently, the designated expert converts the image into text prompts, and a prompt manager merges the converted visual information with text queries, enabling Med-Alpaca to generate appropriate responses. Additionally, OpenMEDLab's XrayPULSE [77] and XrayGPT from the Mohamed bin Zayed University of Artificial Intelligence [78] can analyze and interpret chest X-rays and perform question-and-answer functions based on these analyses.

Currently, multimodal models are rarely used in medical research. Although Visual Med-Alpaca has made some significant progress in the field of medical multimodal models, its datasets consist of English diagnostic reports, which limits its applicability for advancing multimodal medical models in the Chinese context. XrayGLM, developed by Macao Polytechnic University, addresses this issue and has shown remarkable potential in medical image diagnostics and multi-round interactive dialogue [79]. It is evident that multimodal LLMs are continuously improving in the healthcare sector, providing more precise medical services to patients.

B. VIRTUAL REALITY IN SURGERY

The integration of LLMs with virtual reality (VR) technology opens new possibilities for both treatment and medical education. LLMs can understand and generate human language and perform complex reasoning tasks, while VR technology provides immersive interactive experiences. By combining these two technologies, it is possible to create more intuitive and interactive medical education and training systems, improve surgical planning and execution, and ultimately enhance the quality of medical treatment.

VII. BIBLIOMETRIC ANALYSIS OF LITERATURE

The application of LLMs in the field of healthcare has been a hot topic of research in recent years. Conducting a bibliometric analysis of relevant studies can aid researchers in understanding the research trends in a specific field, identifying current research hotspots and frontiers, and facilitating the establishment of collaborations and academic exchanges. This review summarizes the literature on applying LLMs in healthcare published in SCI-indexed journals from 1 January 2018 to 22 March 2024, utilizing Cite Space. The search terms employed were 'LLMs' and 'Healthcare.' The paper's representative institution and country were determined based on the first author's affiliation and location.

A. METHODS AND DATA

1) METHODS

For a more intuitive analysis, the Java program Cite Space, developed in 2006, was used for knowledge graph visualization analysis [80]. Cite Space is a powerful tool for literature analysis and bibliometric visualization. It is a freely available scientific literature database analysis software focusing on bibliometrics and scientometrics. Its significant features include co-citation analysis, which identifies relationships between papers, revealing patterns of interaction and academic exchange among different publications. Visualization tools such as cluster maps and timeline views can be used to explore research topics and potential future directions and discover research hotspots and frontier dynamics in the field [81], [82].

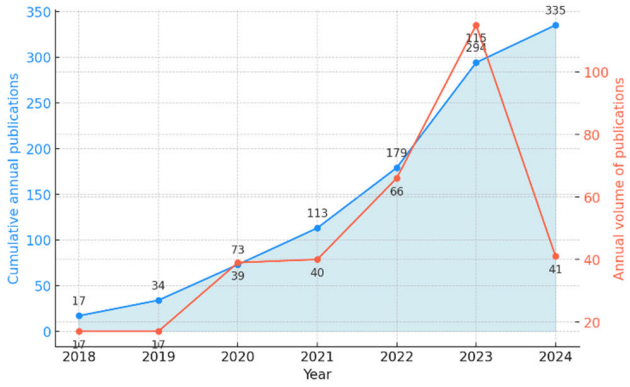

FIGURE 3. Annual distribution of papers and annual cumulative publications (2018–2024).

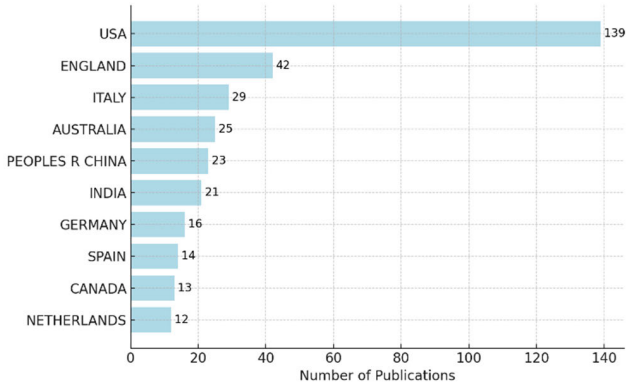

FIGURE 4. Annual bar chart depicting the top ten countries based on the number of publications.

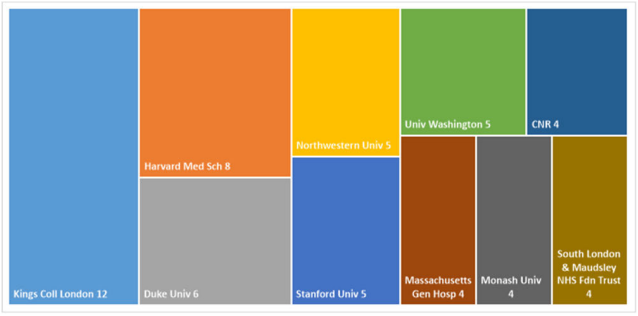

FIGURE 5. Dendrogram of the top ten institutions ranked by the number of publications.

2) DATA

The Web of Science (WOS) collected bibliographic data on applying LLMs in healthcare over seven years (2018–2024). A total of 335 records are present in WOS, excluding reviews articles. WOS is a global academic literature retrieval platform covering many disciplines, including natural sciences, social sciences, and humanities, as well as numerous well-known journals and academic publications. It has multidimensional keyword search capabilities and provides citation analysis tools to gain insights into the connections between literature, facilitating readers' in-depth understanding of literature and scientific research. This study selected the core database of WOS for retrieval to conduct a more scientifically objective bibliometric analysis using higher quality and reliable literature. The focus of this study is the application of large-scale language models in medicine and healthcare. Therefore, the key search terms for literature retrieval are 'Large Language Models' and 'Healthcare.'

B. RESULTS

This study utilizes Cite Space analysis to evaluate contributions from different countries, institutions, journals, and scholars. Cite Space is used to summarize the annual and cumulative publication volumes of relevant literature on applying LLMs in the medical and healthcare field, as shown in Figure 3.

Since 2018, 335 relevant papers on applying LLMs in healthcare have been published. From 2018 to 2023, there has been a significant increase in the number of related documents, rising from 17 to 294. Specifically, from 2021 to 2023, there has been a substantial annual increase in the number of new articles, growing from 40 to 115 per year. The cumulative number of yearly publications demonstrates a noteworthy upward trend, reflecting an increasing scholarly emphasis on research on applying LLMs in the medical and healthcare field.

1) COUNTRY AND PUBLISHER

As shown in Figure 4, the countries with the most published papers on the application of LLMs in the healthcare field are as follows: the United States with 139 papers, the United Kingdom with 42 papers, Italy with 29 papers, Australia with 25 papers, and China with 23 papers. Notably, the United States has the highest number of related documents, which is 3.31 times that of the second-ranked United Kingdom, as presented in figure 4, indicating the highest level of emphasis on applying large models in the medical and healthcare field.

As shown in Figure 5, the institutions with the most published papers on the application of LLMs in the healthcare field are as follows: Kings Coll London (12 papers), Harvard Med Sch (8 papers), Duke Univ (6 papers), Northwestern Univ (5 papers), Stanford Univ (5 papers), and Univ Washington (5 papers). Among the top six institutions, five are in the United States and one in the United Kingdom. This indicates a strong interest in applying LLMs in the medical and healthcare field, particularly in the United States. Furthermore, it suggests that, compared to other countries, the United States has undertaken substantial work and extensive efforts in developing LLMs in the medical and healthcare field.

2) AUTHOR

To further analyze the relevant research on applying LLMs in the healthcare field, this paper utilized Cite Space to statistically summarize the research authors who have already been published and identified several influential authors in this field, as shown in Figure 6.

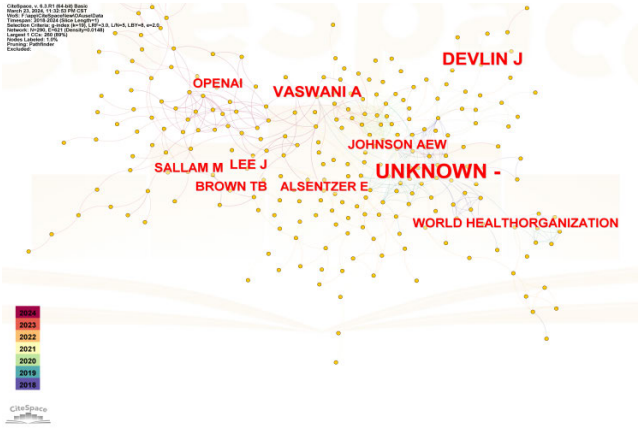

FIGURE 6. Co-citation network of authors (timespan: 2018–2024; slice length = 1; g-index = 25; LRF = 3; LBY = 5; e = 2; n = 353; E = 866).

Keywords	Year	Strength	Begin	End	2018 - 2024
big data	2018	1.64	2018	2019	
attitudes	2018	1.09	2018	2021	
acute physiology	2018	0.98	2018	2020	
knowledge	2019	3.35	2019	2020	
activity recognition	2019	1.11	2019	2020	
risk	2020	1.45	2020	2021	
information	2020	1.45	2020	2021	
depression	2020	1.41	2020	2022	
mortality	2020	1.12	2020	2022	
association	2020	0.94	2020	2021	
children	2020	0.95	2021	2022	
validation	2022	0.92	2022	2024	

FIGURE 7. Thematic trend analysis over seven years.

As shown in Figure 6, there are 353 co-occurrence nodes, 866 links, and a network density of 0.0139. Devlin J has been cited 63 times, followed by Vaswani A with 37 times. Additionally, Lee J has been cited 21 times.

Therefore, Devlin Jacob is the most influential researcher in LLMs, focusing on using Deep Learning techniques to improve natural language processing tasks. He proposed the BERT (Bidirectional Encoder Representations from Transformers) model, significantly contributing to the research and improvement of LLMs. In second place is Ashish Vaswani, an influential researcher in Deep Learning and natural language processing, who significantly contributed to the Transformer model's development. Joon Lee's research focuses on the data mining of clinical data. The study of these authors represents the mainstream trend of applying LLMs in the medical and healthcare field.

3) THEMATIC TRENDS

The phenomenon of keyword explosion refers to a sharp increase in the frequency of keywords within a relatively short period, reflecting the research frontier that has attracted considerable attention in a particular field during a specific period. This review uses Cite Space to measure topic trends over seven years statistically. Topic trends were evaluated by keyword burstiness (measured by changes in keyword frequency), as shown in Figure 7. The ''strength'' based on the statistical formula is used to measure the burstiness of the keywords.

As shown in Figure 7, these lines depict the evolving thematic trends of applying LLMs in healthcare over the past the past seven years. Specifically, from 2018 to 2020, ''big data,'' ''attitudes,'' ''acute physiology,'' ''knowledge,'' and ''activity recognition'' emerged as popular keywords. This reflects the early trends in research on applying LLMs in the medical and healthcare domain, during which studies discussed utilizing big data to improve decision-making and practices in the medical and healthcare field. From 2020 to 2022, ''risk,'' ''information,'' ''depression,'' ''mortality,'' ''association,'' and ''children'' were the most popular keywords. This indicates that researchers showed increased attention to understanding and preventing various health issues, particularly risk, mental health, and children's health. From 2022 to 2024, ''validation'' became a popular keyword, signaling a growing emphasis on validating medical technologies, data, models, and interventions when applying LLMs in healthcare.

FIGURE 8. Co-citation network of cited journals (timespan: 2018–2024; slice length = 1; g-index = 20; LRF = 3; LBY = 5; e = 2; n = 285; E = 428).

4) CITED JOURNALS

To better understand the academic research landscape and identify authoritative journals, this review utilizes Cite Space to analyze the cited journals of relevant studies published on applying LLMs in healthcare. As shown in Figure 8, there are 285 nodes and 428 connections.

The most cited journal is ARXIV (107 citations), a preprint service platform covering multiple disciplines, including physics, mathematics, computer science, and biology. It allows researchers to share their research findings before peer review, and this form of open access enables rapid dissemination and exchange of scientific research within the

scientific community. Following that is the journal PLOS ONE (88 citations), an open-access scientific journal covering various disciplines such as life sciences, medicine, engineering, and computer science. The Journal of Biomedical Informatics (74 citations) is dedicated to publishing high-quality original research in medical informatics, focusing on areas such as biomedical informatics and clinical informatics related to healthcare. These highly cited journals serve as important indicators in the research field of applying LLMs in the medical and healthcare domain and profoundly impact the research.

VIII. CONSTRUCTION OF LARGE LANGUAGE MODELS

This review provides a comprehensive review and summary of medical LLMs. Building on existing literature, it begins by introducing the underlying mechanisms of LLMs, starting with general models. It then transitions to medical-specific language models, discussing aspects such as corpora, paradigms, training processes, and fine-tuning, as well as an overview of the evaluation frameworks for these models.

Next, based on the practical needs of medical Artificial Intelligence, the paper outlines the applications of medical LLMs in areas such as disease diagnosis and prediction, decision support, medical knowledge dissemination, medical assistance, image analysis, biomedicine, and medical education. Several typical medical LLMs are analyzed in detail. In addition, bibliometric analysis and discussion were conducted, revealing the increasing importance of LLM applications in healthcare.

However, there are several challenges and issues with medical LLMs that require further optimization and improvement. This paper summarizes these challenges and provides corresponding opinions and suggestions. Future directions include exploring new healthcare approaches by integrating LLMs with the internet, and integrating LLMs with other digital health technologies, such as wearable devices and telemedicine platforms. It is also crucial to continuously develop more transparent and equitable datasets for diverse populations, aiming to address data quality issues and minimize biases and discrimination risks in generated content.

The rapid rise of research on large AI models, along with the accelerated transformation of related industries, presents new opportunities for the development of medical LLMs. Thus, although medical LLMs have made rapid progress in recent years, more in-depth research is needed to advance their development in the medical field.

ACKNOWLEDGMENT

This study received professional consultation from the First Clinical Hospital of Beijing University of Chinese Medicine.

REFERENCES

[1] L. Yang, Y. Chen, Q. Qiu, and J. Wang, "Risk control of mission-critical systems: Abort decision-makings integrating health and age conditions," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 6887–6894, Oct. 2022.

[2] T. Sun, X. He, and Z. Li, "Digital twin in healthcare: Recent updates and challenges," *Digit. Health*, vol. 9, Jan. 2023, Art. no. 20552076221149651.

[3] C. K. Wee, X. Zhou, R. Gururajan, X. Tao, J. Chen, R. Gururajan, N. Wee, and P. D. Barua, "Automated triaging medical referral for otorhinolaryngology using data mining and machine learning techniques," *IEEE Access*, vol. 10, pp. 44531–44548, 2022.

[4] A. Patnaik and P. K. Krishna, "Intelligent decision support system in healthcare using machine learning models," *Recent Patents Eng.*, vol. 18, no. 5, Jul. 2024, Art. no. e060623217715.

[5] A. Batool and Y.-C. Byun, "Enhanced sentiment analysis and topic modeling during the pandemic using automated latent Dirichlet allocation," *IEEE Access*, vol. 12, pp. 81206–81220, 2024.

[6] H. A. Younis, T. A. E. Eisa, M. Nasser, T. M. Sahib, A. A. Noor, O. M. Alyasiri, S. Salisu, I. M. Hayder, and H. A. Younis, "A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: Applications, considerations, limitations, motivation and challenges," *Diagnostics*, vol. 14, no. 1, p. 109, Jan. 2024.

[7] S. Swain and K. Muduli, "Uncovering the issues associated with AI and other disruptive technology enabled operational practices in healthcare sectors in India," *Recent Patents Eng.*, vol. 18, no. 5, Jul. 2024, Art. no. e130223213615.

[8] P. Malik and S. Singh, "Deep learning approaches and biomarkers in medical diagnosis," *Recent Patents Eng.*, vol. 18, no. 3, Apr. 2024, Art. no. e300123213249.

[9] H. Guo, P. Liu, R. Lu, F. Yang, H. Xu, Y. Zhuang, G. Huang, S. Song, and K. He, "Research on a massively large artificial intelligence model and its application in medicine," *SCIENTIA SINICA Vitae*, vol. 54, no. 3, pp. 482–506, 2024.

[10] M. Cascella, F. Semeraro, J. Montomoli, V. Bellini, O. Piazza, and E. Bignami, "The breakthrough of large language models release for medical applications: 1-year timeline and perspectives," *J. Med. Syst.*, vol. 48, no. 1, pp. 1–11, Feb. 2024.

[11] J. Surovková, S. Haluzová, M. Strunga, R. Urban, M. Lifková, and A. Thurzo, "The new role of the dental assistant and nurse in the age of advanced artificial intelligence in telehealth orthodontic care with dental monitoring: Preliminary report," *Appl. Sci.*, vol. 13, no. 8, p. 5212, Apr. 2023.

[12] D. Tian, S. Jiang, L. Zhang, X. Lu, and Y. Xu, "The role of large language models in medical image processing: A narrative review," *Quant. Imag. Med. Surg.*, vol. 14, no. 1, pp. 1108–1121, Jan. 2024.

[13] H. Xiong, H. Chen, L. Xu, H. Liu, L. Fan, Q. Tang, and H. Cho, "A survey of data element perspective: Application of artificial intelligence in health big data," *Frontiers Neurosci.*, vol. 16, Oct. 2022, Art. no. 1031732.

[14] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomed. Eng.*, vol. 6, no. 12, pp. 1346–1352, Aug. 2022.

[15] F. H. Xue, H. B. Jiang, and D. Tang, "Review of deep learning applications in healthcare," *Comput. Sci.*, vol. 50, no. 4, pp. 1–15, 2023.

[16] Z. S. Hu, R. Yang, and J. H. Zhu, "Research and development of large language models in the medical field," *Artif. Intell. View*, vol. 4, no. 4, pp. 10–19, 2023.

[17] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[19] C. Chakraborty, S. Pal, M. Bhattacharya, S. Dash, and S.-S. Lee, "Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science," *Frontiers Artif. Intell.*, vol. 6, Oct. 2023, Art. no. 1237704.

[20] J. Z. Luo, Y. L. Sun, and Z. Z. Qian, "Overview and prospect of artificial intelligence large models," *Radio Eng.*, vol. 53, no. 11, pp. 2461–2472, 2023.

[21] Y. Wang, Q. Li, Z. Dai, and Y. Xu, "Current status and trends in large language modeling research," *Chin. J. Eng.*, vol. 46, no. 8, pp. 1411–1425, 2024.

[22] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," *Engineering*, vol. 25, pp. 51–65, Jun. 2023.

[23] T. Ruan, Y. A. Bian, and G. Y. Yu, "A review on research and application of medical large language models," *Chin. J. Health Informat. Manage.*, vol. 20, no. 6, pp. 853–861, 2023.
[24] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and É. Grave, "CCNet: Extracting high quality monolingual datasets from web crawl data," in *Proc. 12th Int. Conf. Lang. Resour. Eval. (LREC)*, Paris, France, Jan. 2019, pp. 1–9.
[25] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800GB dataset of diverse text for language modeling," 2021, *arXiv:2101.00027*.
[26] W. T. Shu, R. X. Li, and T. X. Sun, "Large language models: Principles, implementation, and progress," *J. Comput. Res. Develop.*, vol. 61, no. 2, pp. 351–361, 2024.
[27] A. Pal, L. Kumar Umapathi, and M. Sankarasubbu, "MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering," 2022, *arXiv:2203.14371*.
[28] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, and P. Payne, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Jul. 2023.
[29] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, May 2016.
[30] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 9049–9058.
[31] A. B. Abacha, Y. Mrabet, M. J. Sharp, T. R. Goodwin, S. E. Shooshan, and D. Demner-Fushman, "Bridging the gap between consumers' medication questions and trusted answers," in *MEDINFO 2019: Health and Wellbeing e-Networks for All*, vol. 264, L. Ohnomachado and B. Seroussi, Eds., Amsterdam, The Netherlands: IOS Press, 2019, pp. 25–29.
[32] Y. Cao, Y. Kang, C. Wang, and L. Sun, "Instruction mining: Instruction data selection for tuning large language models," 2023, *arXiv:2307.06290*.
[33] J. D. Zhang, S. H. Yang, and J. F. Liu, "AIGC empowering the revitalization of traditional Chinese medicine ancient books: A study on the construction of the huang-Di large language model," *Library Tribune*, vol. 44, no. 10, pp. 103–112, 2023.
[34] Y. Wang, Y. X. Song, and Y. F. Wang, "Ethical and governance issues of artificial intelligence in the health field: A multimodal large model guide," *Chin. Medical Ethics*, vol. 37, no. 9, pp. 1–58, 2024.
[35] C. Y. Zhao, G. B. Zhu, and J. Q. Wang, "The inspiration brought by ChatGPT to LLM and the new development ideas of multi-modal large model," *Data Anal. Knowl. Discovery*, vol. 7, no. 3, pp. 26–35, 2023.
[36] G. Wang, X. Liu, Z. Ying, G. Yang, Z. Chen, Z. Liu, M. Zhang, H. Yan, Y. Lu, Y. Gao, K. Xue, X. Li, and Y. Chen, "Optimized glycemic control of type 2 diabetes with reinforcement learning: A proof-of-concept trial," *Nature Med.*, vol. 29, no. 10, pp. 2633–2642, Oct. 2023.
[37] B. Sheng, Z. Guan, L.-L. Lim, Z. Jiang, N. Mathioudakis, J. Li, R. Liu, Y. Bao, Y. M. Bee, Y.-X. Wang, Y. Zheng, G. S. W. Tan, H. Ji, J. Car, H. Wang, D. C. Klonoff, H. Li, Y.-C. Tham, T. Y. Wong, and W. Jia, "Large language models for diabetes care: Potentials and prospects," *Sci. Bull.*, vol. 69, no. 5, pp. 583–588, Mar. 2024.
[38] B. Cheng, Y. Zhang, D. Cai, W. Qiu, and D. Shi, "Construction of traditional Chinese medicine knowledge graph using data mining and expert knowledge," in *Proc. Int. Conf. Netw. Infrastructure Digit. Content (ICNIDC)*, Guiyang, China, Aug. 2018, pp. 209–213.
[39] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-bench and chatbot arena," 2023, *arXiv:2306.05685*.
[40] P. Yu, H. Xu, X. Hu, and C. Deng, "Leveraging generative AI and large language models: A comprehensive roadmap for healthcare integration," *Healthcare*, vol. 11, no. 20, p. 2776, Oct. 2023.
[41] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–35, Sep. 2023.
[42] P. Ke, W. Q. Lei, and M. L. Huang, "Research progress of large language models represented by ChatGPT," *Bull. Nat. Natural Sci. Found. China*, vol. 37, no. 5, pp. 714–723, 2023.
[43] Y. H. Xiao and Y. D. Xu, "The application of large generative language models in the medical field: Opportunities and challenges," *J. Med. Informat.*, vol. 44, no. 9, pp. 1–11, 2023.
[44] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, J. Li, G. Chen, X. Wu, Z. Zhang, Q. Xiao, X. Wan, B. Wang, and H. Li, "HuatuoGPT, towards taming language model to be a doctor," 2023, *arXiv:2305.15075*.
[45] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digit. Health*, vol. 2, no. 2, Feb. 2023, Art. no. e0000198.
[46] D. P. Panagoulias, M. Virvou, and G. A. Tsihrintzis, "Augmenting large language models with rules for enhanced domain-specific interactions: The case of medical diagnosis," *Electronics*, vol. 13, no. 2, p. 320, Jan. 2024.
[47] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, and D. Shen, "DoctorGLM: Fine-tuning your Chinese doctor is not a herculean task," 2023, *arXiv:2304.01097*.
[48] J. Z. Yan, Y. X. He, and Z. Y. Luo, "Generative large language models in the medical domain: Potential and typical applications and challenges," *J. Med. Informat.*, vol. 44, no. 9, pp. 23–31, 2023.
[49] S. L. Zheng, X. T. Li, and M. Xu, "Automatic generation of healthcare examination summaries using large language models," *J. Chin. Comput. Syst.*, vol. 45, no. 11, pp. 1–8, 2024.
[50] A. Arora and A. Arora, "The promise of large language models in health care," *Lancet*, vol. 401, no. 10377, p. 641, Feb. 2023.
[51] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutiérrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, Jul. 2023.
[52] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Commun.*, vol. 5, no. 1, pp. 1–9, Jun. 2014.
[53] B. Meskó, "The impact of multimodal large language models on health care's future," *J. Med. Internet Res.*, vol. 25, Nov. 2023, Art. no. e52865.
[54] P. P. Ray and P. Majumder, "AI tackles pandemics: ChatGPT's game-changing impact on infectious disease control," *Ann. Biomed. Eng.*, vol. 51, no. 10, pp. 2097–2099, Oct. 2023.
[55] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, "Large language models generate functional protein sequences across diverse families," *Nature Biotechnol.*, vol. 41, no. 8, pp. 1099–1106, Aug. 2023.
[56] D. H. A. Mai, L. T. Nguyen, and E. Y. Lee, "TSSNote-CyaPromBERT: Development of an integrated platform for highly accurate promoter prediction and visualization of *Synechococcus* sp. and *Synechocystis* sp. through a state-of-the-art natural language processing model BERT," *Frontiers Genet.*, vol. 13, Nov. 2022, Art. no. 1067562.
[57] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, "Artificial intelligence in drug discovery and development," *Drug Discovery Today*, vol. 26, no. 1, pp. 80–93, Jan. 2021.
[58] M. Karabacak, B. B. Ozkara, K. Margetis, M. Wintermark, and S. Bisdas, "The advent of generative language models in medical education," *JMIR Med. Educ.*, vol. 9, Jun. 2023, Art. no. e48163.
[59] Y. Zhang and Y. Hao, "Traditional Chinese medicine knowledge graph construction based on large language models," *Electronics*, vol. 13, no. 7, p. 1395, Apr. 2024.
[60] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," 2022, *arXiv:2211.09085*.
[61] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245
[62] S. Huh, "Can we trust AI chatbots' answers about disease diagnosis and patient care?" *J. Korean Med. Assoc.*, vol. 66, no. 4, pp. 218–222, Apr. 2023.
[63] A. C. Fernandes and M. E. V. C. Souto, "Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine," *New England J. Med.*, vol. 388, no. 25, pp. 2399–2400, 2023.
[64] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," 2023, *arXiv:2303.13375*.

[65] J. Xiao, D. Xu, and H. Wang, "A survey of large language models in healthcare," *CAAI Trans. Intell. Syst.*, pp. 1–15, Sep. 2024.
[66] J. Devlin, M. W. Chang, and K. Lee, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North*, Minneapolis, MN, USA, 2019, pp. 1–16.
[67] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.
[68] E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin, and C. D. Manning, "BioMedLM: A 2.7B parameter language model trained on biomedical text," 2024, *arXiv:2403.18421*.
[69] D. Gu, Z. Huang, and K. Zhu, "Medical health large language models: Construction of knowledge system, service applications, and synergetic governance of risk management," *Inf. Sci.*, pp. 1–29, Sep. 2024.
[70] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Stroudsburg, PA, USA, 2022, pp. 1–39.
[71] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-HALT: Medical domain hallucination test for large language models," in *Proc. 27th Conf. Comput. Natural Lang. Learn. (CoNLL)*, Singapore, 2023, pp. 314–334.
[72] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P. C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, and A. Palepu, "Towards generalist biomedical AI," 2023, *arXiv:2307.14334*.
[73] K. M. Boehm, P. Khosravi, R. Vanguri, J. Gao, and S. P. Shah, "Harnessing multimodal data integration to advance precision oncology," *Nature Rev. Cancer*, vol. 22, no. 2, pp. 114–126, Feb. 2022.
[74] S. Niu, J. Ma, L. Bai, Z. Wang, L. Guo, and X. Yang, "EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102069.
[75] N. Luo, X. Zhong, L. Su, Z. Cheng, W. Ma, and P. Hao, "Artificial intelligence-assisted dermatology diagnosis: From unimodal to multimodal," *Comput. Biol. Med.*, vol. 165, Oct. 2023, Art. no. 107413.
[76] (2024). *Visual Med-Alpaca: A Parameter-Efficient Biomedical LLM With Visual Capabilities*. Cambridge Language Technology Lab. Accessed: Mar. 18, 2024. [Online]. Available: https://github.com/cambridgeltl/visual-med-alpaca
[77] (2024). *Openmedlab/XrayPULSE*. OpenMEDLab. Accessed: Mar. 18, 2024. [Online]. Available: https://github.com/openmedlab/XrayPULSE
[78] O. Thawkar, A. Shaker, S. Shaji Mullappilly, H. Cholakkal, R. Muhammad Anwer, S. Khan, J. Laaksonen, and F. Shahbaz Khan, "XrayGPT: Chest radiographs summarization using medical vision-language models," 2023, *arXiv:2306.07971*.
[79] R. Wang, Y. Duan, and J. Li. *XrayGLM*. Accessed: Mar. 18, 2024. [Online]. Available: https://github.com/WangRongsheng/XrayGLM
[80] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 359–377, Feb. 2006.
[81] P. Zhang, Y. Du, S. Han, and Q. Qiu, "Global progress in oil and gas well research using bibliometric analysis based on VOSviewer and CiteSpace," *Energies*, vol. 15, no. 15, p. 5447, Jul. 2022.
[82] B. Lei, Y. Ren, H. Luan, R. Dong, X. Wang, J. Liao, S. Fang, and K. Gao, "A review of optimization for system reliability of microgrid," *Mathematics*, vol. 11, no. 4, p. 822, Feb. 2023, doi: 10.3390/math11040822.

QIMING LIU received the Bachelor of Management degree from Zhejiang University of Finance and Economics, Hangzhou, China, in 2022. She is currently pursuing the master's degree with the Management Science and Engineering Department, Beijing Information Science and Technology University, Beijing, China.

Her research interests include health management, online healthcare services, and the application of large language models in the healthcare sector.

RUIRONG YANG received the master's degree from the First Clinical Medical School, Beijing University of Chinese Medicine and Medical School, University of Glasgow.

Her main research interests include the clinical study of psychosomatic diseases in Chinese medicine. Her current research focuses on sleep disorders, transmission of experience of famous, and old Chinese medicine doctors, and public health.

QIN GAO received the Bachelor of Management degree from Beijing Forestry University, in 2022, where she is currently pursuing the master's degree in applied statistics.

Her research interests include online healthcare, emergency resource allocation, and statistical modeling for decision-making.

TENGXIAO LIANG received the Ph.D. degree in medical science from Beijing University of Traditional Chinese Medicine, which mainly focuses on the treatment of acute and critical diseases and the diagnosis and treatment of complicated internal diseases through the integration of traditional Chinese and modern medicine.

He is currently a Professor and a Ph.D. Supervisor with Beijing University of Chinese Medicine, and the Chief Physician of the Emergency Department of Dongzhimen Hospital.

XIUYUAN WANG received the Bachelor of Management degree from Hohai University, Changzhou, China, in 2019, and the Master of Engineering Administration degree from Beijing Information Science and Technology University, Beijing, China, in 2023.

She is currently a Lecturer with the Department of Digital Economy, Changzhou Liu Guojun Vocational and Technical College, Changzhou. Her research interests include high-quality development of economy, comparative advantage evaluation in decision-making, and large multimodal models in education.

SHIJU LI received the master's degree in management from Beijing Jiaotong University, Beijing, China, where he is currently pursuing the Ph.D. degree with the Management Science and Engineering Department.

His main research interests include health management and supply chain management.

BINGYIN LEI received the Ph.D. degree in management from Renmin University of China, Beijing, China.

He is currently an Associate Professor with Beijing Information Science and Technology University, Beijing, where his primary research focus is on the digital economy, industrial ecology, and technological innovation. In addition to his academic role, he holds several distinguished positions, including a Secretary-General of the Resources and Environmental Protection Committee of the China Productivity Society, a Distinguished Researcher with the Center for Internet Governance, Tsinghua University, Beijing, and an Adjunct Professor with the School of Economics and Management, Beijing Jiaotong University.

KAIYE GAO (Member, IEEE) received the Ph.D. degree from the University of Science and Technology Beijing, Beijing, China (jointly trained with the University of Manchester, U.K.), in 2018.

He is currently a Professor and a Ph.D. Supervisor with Beijing Forestry University, Beijing, a Postdoctoral Fellow with the University of Chinese Academy of Sciences, Beijing, and a Postdoctoral Researcher with The Hong Kong Polytechnic University, Hong Kong. In recent years, he has published or accepted more than 60 articles in these fields, including 21 SCI journal articles. His publications have appeared in journals, such as IEEE Transactions on Systems, Man and Cybernetics: Systems, *Journal of Industrial Information Integration*, *Reliability Engineering and System Safety*, *Journal of Risk and Reliability*, *Systems Engineering-Theory and Practice*, *Journal of Systems and Management*, and *Journal of Systems Science and Mathematical Sciences*. His primary research interests include system risk and reliability.

Prof. Gao also serves as the Director of the System Reliability Branch of the Chinese Society of Systems Engineering, a Committee Member of the Chinese Command and Control Society (Safety Protection and Emergency Management, Reliability Science and Engineering), the Director of the Reliability Engineering Branch of the Chinese Society for Field Statistics, an Expert in the Sino-Polish University Alliance Talent Pool, and a Reviewer for several international renowned SCI journals, including IEEE Transactions Series and Reliability Engineering and Systems Safety.