Large language models in medicine: methods and applications

Yu Sun[1,#], Guiyan Liu[1,#], Tingyang Xu[2,*], Qifeng Bai[1,#,*]

[1]School of Basic Medical Sciences, Lanzhou University, Lanzhou University, Lanzhou 730000, Gansu, P. R. China

[2]Tencent AI Lab, Shenzhen, P. R. China

# These authors contributed equally to this work

* To whom correspondence should be addressed. Tel: +86-931-8912011; Fax: +86-931-8912011; Email: baiqf@lzu.edu.cn or tingyangxu@tencent.com

Abstract

The application of Large Language Models (LLMs) in the medical field signifies a revolutionary shift in medical informatics and patient care. The LLMs trained on vast and specialized medical corpora, bring precision and efficiency to medical diagnostics, treatment planning, and service delivery. This paper outlines the essential methodologies for preparing LLMs for deployment in healthcare as well as pre-training and fine-tuning processes. Innovative pre-training methods are introduced, such as the multimodal and Contrastive Language-Image Pre-training (CLIP)-based approaches that enhance the model understanding across different data types. Besides, fine-tuning techniques such as Supervised Fine-Tuning (SFT), Instruction Fine-Tuning (IFT), and Parameter-Efficient Tuning (PET) are discussed, which can tailor the LLMs to specific medical tasks with relatively low resource requirements. Moreover, our paper delves into advanced prompting strategies, including zero/few-shot, Chain-of-Thought (CoT), and self-consistency prompting, which refine the model's capacity to handle complex medical queries, as well as Retrieval-Augmented Generation (RAG) and interactive web search technologies in improving the accuracy and reliability of model responses, which can enhance the applicability of LLMs in real-world medical practice. At last, our article depicts the applications of multimodal models in medicine and LLMs in medical texts and images. Our paper not only provides the common methods and applications of LLMs in medicine but also the future directions, ethics, and privacy in this rapidly advancing field.

Keywords: Artificial intelligence; Deep learning; Large language models; Multi-modal learning

1. Introduction

The fusion of artificial intelligence (AI) with medical science marks the dawn of a new era in healthcare, defined by unparalleled precision and operational efficiency. At the forefront of this transformative shift are Large Language Models (LLMs), which are meticulously trained on extensive and specialized datasets. Key players such as PubMedBERT [1] and ClinicalBERT [2] undergo thorough pre-training regimens that immerse them deeply in the realm of medical knowledge. This crucial groundwork involves the analysis of extensive medical texts, including PubMed literature and clinical case studies, equipping LLMs with the sophisticated understanding required for intricate medical language processing. Additionally, fine-tuning methodologies like Parameter-Efficient Tuning and Instruction Fine-Tuning further sharpen these models for specific applications, significantly enhancing their efficiency and versatility [3]. These techniques are particularly designed to optimize resource use while enhancing the precision of outputs, an essential trait for deployment in resource-limited settings such as small-scale clinics or in emergency response scenarios.

In practical scenarios, the deployment of LLMs within the medical domain is catalyzing a revolution in patient care and clinical decision-making. The LLMs excel in a variety of functions, from automating diagnostics to crafting personalized treatment plans [4-7]. Notably, Retrieval-Augmented Generation (RAG) technology bolsters the trustworthiness of LLMs outputs by integrating up-to-date information from medical databases, ensuring that the provided diagnostics and recommendations adhere to the latest research and clinical protocols. This feature can gain useful and recent information to deal with complex medical tasks [8]. Furthermore, the integration of web search technologies enables these models to sift through the extensive online medical corpus, anchoring their responses in the most pertinent information available. By enhancing the precision and swiftness of medical services, LLMs not only elevate the standard of care but also set the stage for future innovations in medical AI. The discussion also addresses challenges such as data privacy, potential biases in model training, and the necessity for ongoing updates, offering a comprehensive overview of the environment in which these models operate.

The integration of AI with medical science via Large Language Models (LLMs) marks a pivotal shift in healthcare. These models, refined through advanced training techniques, significantly enhance diagnostic precision and treatment efficacy. Supported by technologies like Retrieval-Augmented Generation, LLMs are transforming patient care in real-world settings [8]. To better understand the application of LLMs in medicine, our review elaborates on the pre-training and fine-tuning processes essential for preparing LLMs for healthcare, highlighting innovative techniques that enhance model understanding and operational efficiency as well as promote strategies, Retrieval-Augmented Generation, and web search that contribute to the improved functionality and reliability of LLMs in handling complex medical scenarios. Besides, to better understand the applications of LLMs in the medical area, LLMs are listed and discussed for medical texts, images, and multimodal models. The critical challenges such as data privacy and model bias are also described to provide a comprehensive outlook on the future integration prospects of LLMs in healthcare.

2. Methods of Medical Large Language Models

In the rapidly evolving landscape of artificial intelligence within healthcare, the integration of large language models (LLMs) marks a transformative era for medical informatics and patient care. These sophisticated models, equipped with vast data and advanced computational capabilities, are set to revolutionize areas such as medical diagnostics, treatment planning, and service delivery. This section explores the research progress of pre-training and fine-tuning, preparing LLMs for effective deployment in the medical domain. It highlights innovative methods that enhance their performance and adaptability, such as Parameter-Efficient Tuning, Instruction Fine-Tuning, and advanced prompting techniques. Moreover, this section addresses emerging methodologies like Retrieval-Augmented Generation (RAG) and web search integrations, which further refine the utility and accuracy of LLMs in handling complex medical queries (see Figure 1). By delving into these technologies, we aim to highlight both the significant advancements and the ongoing challenges in optimizing LLMs to meet the specific demands of healthcare, ultimately paving the way for more personalized, efficient, and precise medical services.

2.1. Pre-training Models

Pre-trained models [9], especially large language models (LLMs), have become a cornerstone of artificial intelligence applications in the medical domain. Specialized large language models such as PubMedBERT [1], ClinicalBERT [2], BlueBERT [10], BioBERT [11], SciBERT [12], BEHRT [13], UmlsBERT [14], GatorTron [15] and MEDITRON [16] are typically pre-trained on specialized medical corpora. These corpora include medical literature from resources like PubMedQA [17], MIMIC-III [18], and webMedQA [19]. Through extensive pre-training, these models gain a deep understanding of the complexities of medical language, providing a solid foundation for application and fine-tuning in specific medical tasks.

The primary objectives of pre-training include masked language modeling, next sentence prediction, and next token prediction tasks. In these tasks, BERT series models [20-22] generally focus on the first two objectives, while GPT series models [3, 23] tend to emphasize the latter. Once pre-trained, these models are fine-tuned for specific application scenarios such as Question Answering (QA) and Named Entity Recognition (NER), to tailor their performance. The effectiveness of these models is assessed through standards such as the Biomedical Language Understanding Evaluation (BLUE) and the Biomedical Language Understanding and Reasoning Benchmark (BLURB), to validate their efficacy and accuracy in practical applications.

In addition to these traditional pre-training techniques, we have explored some innovative pre-training methods aimed at further enhancing the models' mastery of medical knowledge: Multimodal pre-training, particularly CLIP-based pre-training [24, 25], is becoming increasingly important in the medical field. This approach integrates multiple data modalities, including text, images, and structured data, enabling models to develop more comprehensive data representations as well as to deepen the understanding of complex relationships between medical language and imagery. Especially in matching medical images (such as X-rays, MRI, and CT scans) with corresponding clinical reports, CLIP's pre-training and fine-tuning for specific medical datasets significantly enhance its performance in tasks such as clinical diagnosis, disease classification, and treatment planning. The semantic understanding and adaptive learning capabilities of these deep learning models not only enhance their language and visual comprehension but also enable them to handle complex medical scenarios efficiently and accurately, paving the way for enhanced quality and efficiency in medical services.

Overall, pre-trained models play a crucial role in the medical domain. They not only accurately understand and generate medical language, but also provide more effective and refined AI support for the medical field by adopting cutting-edge pre-training methods, thereby driving technological advancement and improving the quality of healthcare services.

2.2. Fine-tuning Methods

Fine-tuning medical large language models (LLMs) is a crucial step in adapting general LLMs to the medical domain. Due to the high costs and time required to train medical LLMs from scratch, researchers have developed various fine-tuning methods to imbue general LLMs with domain-specific medical knowledge. The main fine-tuning methods include Supervised Fine-Tuning (SFT) and Parameter-Efficient Tuning.

Supervised Fine-Tuning (SFT) is a pivotal technique in modern artificial intelligence, specifically in the realm of medical large language models (LLMs). By utilizing high-quality, labeled medical datasets including physician-patient dialogues, medical Q&A pairs, and knowledge graphs, SFT finely tunes pre-trained language models to meet the unique demands of the medical field. This process involves precise adjustments and the generation of task-specific training samples that teach the model to produce contextually relevant medical information. Notably, during training, a mask vector is applied to focus the model's learning on the response sections of these samples, enhancing its grasp of medical terminology and improving both accuracy and adaptability in professional medical scenarios. SFT effectively transforms general language models into specialized tools capable of handling complex medical contexts and providing accurate clinical decision support. Despite its profound impact, SFT faces challenges related to data quality, model generalization, and ethical privacy issues. Addressing these concerns through ongoing research and optimization is essential to maximize the effectiveness, security, and transparency of these models in real-world applications, thereby setting the stage for future advancements in healthcare technology.

Instruction Fine-Tuning (IFT) represents a specialized adaptation of Supervised Fine-Tuning (SFT) aimed at enhancing the performance and controllability of large language models (LLMs) within the medical field. Unlike traditional SFT, IFT specifically focuses on improving a model's ability to comprehend and execute human instructions by employing training datasets composed of instruction-input-output triples, such as instruction-question-answer sequences. This approach not only boosts the model's proficiency in handling complex medical tasks but also ensures the precision and relevance of its outputs, vital for medical applications. IFT has led to the development of models like MedPaLM-2 [26], which demonstrate superior performance in medical question-answering and multi-turn dialogue scenarios. These models are instrumental in aiding medical professionals by accurately interpreting queries and producing responses aligned with specific instructions, thus improving patient education and self-management. Despite its benefits, IFT faces challenges related to the diversity and quality of instructional datasets, model adaptability, and generalization across varied medical contexts. Future research should focus on integrating IFT with other fine-tuning methods and updating models to keep pace with medical advancements, ensuring the continued relevance and accuracy of the information they provide.

Parameter-Efficient Tuning (PET) [27] techniques are increasingly vital in deploying large language models (LLMs) like GPT-3 within the medical field, particularly for handling complex text data and supporting clinical decisions. These methods focus on optimizing the performance of pre-trained models while minimizing resource usage by adjusting only a small subset of parameters. Techniques such as Low-Rank Adaptation (LoRA) [28], Adapters [29-31], and Prefix Tuning [32] substantially reduce computational demands and memory requirements. This is crucial for resource-limited medical institutions, allowing them to utilize advanced language models without incurring high costs. PET not only enhances resource efficiency but also provides the flexibility needed for diverse medical tasks, such as pathology report analysis and clinical trial document processing. By allowing rapid adaptation to specific tasks with minimal training, PET maintains task relevancy and improves accuracy, especially in handling medical terminologies and protocols. Additionally, PET supports model transparency and interpretability by preserving the original model structure, which is essential in medical applications and facilitates decision tracking and logical explanations of outputs. The ability of PET to rapidly adapt to changes in medical data and guidelines demonstrates its utility, as it reduces the time and resources required for model adjustments. Future research might explore the integration of various PET methods to optimize resource efficiency and performance, such as combining LoRA with Adapters for a balanced approach to feature preservation and task adaptability. Further studies could also focus on the automated selection of optimal PET techniques for specific medical tasks, enhancing the practical utility of medical LLMs. Overall, PET techniques offer a practical solution for deploying resource-efficient and adaptable medical language models, crucial for the dynamic and specialized field of medicine.

2.3. Prompting Methods

Prompting [33] methods represent a strategic and efficient approach to tailor general large language models (LLMs) to specific domains like medicine, circumventing the need for extensive additional training. Techniques such as Zero/Few-shot prompting, Chain-of-Thought (CoT) prompting, and Self-consistency prompting, along with Prompt Tuning, are crucial in aligning these powerful computational models to the nuanced requirements of medical diagnostics and treatment planning. These methods enhance the models' capability to process and interpret complex medical data, leading to more accurate and reliable healthcare solutions.

Zero/Few-shot Prompting: in the medical field, the application of large language models

(LLMs) is increasingly pivotal, especially in addressing complex medical issues and data. Leveraging pre-trained models in conjunction with zero/few-shot prompting techniques, these models adeptly tackle new tasks without the need for extensive labeled datasets. Zero-shot prompting [34] equips models to infer and execute tasks based on just a task description, using meticulously crafted prompts to navigate new challenges and utilizing their broad pre-trained knowledge bases for predictions. Conversely, few-shot prompting [35] equips models with a select number of examples or task demonstrations, enhancing their learning before task execution. This method proves particularly beneficial for intricate tasks such as medical question answering (QA), where models enhance downstream performance by examining a handful of high-quality medical QA pairs. Although these techniques greatly improve the models' proficiency in processing medical language, they also present challenges, including increased token usage for input, which can be restrictive for longer texts, and the potential for biases influenced by the selection of prompt examples. To optimize performance and reduce unintended biases, precise prompt engineering is essential. With continuous technological progress and innovations, LLMs are anticipated to play a more significant role in improving diagnostic accuracy, personalizing treatment plans, and comprehending complex medical reports. Overall, zero/few-shot prompting offers a flexible and effective strategy for deploying medical LLMs, enabling them to process and comprehend complex medical data and queries without extensive labeled datasets, and is set to markedly influence the future of medical AI.

Chain-of-Thought (CoT) Prompting: in the medical field, the incorporation of sophisticated prompting techniques with large language models (LLMs) significantly boosts their ability to address complex medical challenges. Chain-of-Thought (CoT) [36] prompting enhances the transparency and explainability of medical reasoning tasks by guiding models through structured, logical problem-solving steps, proving particularly beneficial in medical question-answering systems for diagnostics and treatment planning. Automatic Chain-of-Thought (Auto-CoT) [37] alleviates the labor-intensive process of crafting high-quality examples by autonomously generating reasoning chains, thereby increasing the robustness and accuracy of medical decision-making. Logical Chain-of-Thought (LogiCoT) [38] employs principles of symbolic logic to verify and refine each reasoning step, effectively reducing logical errors and hallucinations in intricate medical diagnostics. Condensed Symbolic (CoS) [39] prompting augments the model's capability to manage tasks involving complex spatial relationships through the use of succinct symbols rather than natural language, which is crucial for interpreting medical imagery or orchestrating surgical procedures. Tree-of-Thoughts (ToT) [40] prompting employs a structured hierarchy of reasoning steps to facilitate systematic exploration of potential solutions, thereby enabling more comprehensive treatment options and strategies. The integration of advanced prompting techniques in medical large language models (LLMs) significantly improves the precision of diagnostics and treatment recommendations. Furthermore, it enhances the interpretability and transparency of their decision-making processes. This advancement is set to substantially advance the field of medical AI, thereby enhancing the quality and efficiency of clinical healthcare services.

Self-consistency Prompting: the integration of Self-consistency [41] Prompting into medical large language models (LLMs) represents a significant breakthrough in enhancing diagnostic accuracy and consistency within healthcare. Building on the Chain-of-Thought (CoT) prompting foundation, this method significantly enhances response reliability by generating multiple answers to the same question and selecting the most consistent result. This is particularly crucial in the medical sector, where precise and reproducible diagnoses and treatment recommendations are vital. Self-consistency Prompting refines CoT by generating diverse reasoning chains from the model's decoder, adeptly navigating the complexities and variabilities of medical reasoning, and minimizing the risk of incorrect conclusions due to biased or incomplete analyses. By facilitating the synthesis and reconciliation of various reasoning paths, Self-consistency Prompting significantly bolsters the development of sophistication and reliable medical decision-support systems. This technique aligns with the broader goals of medical AI, promising safer, more effective, and consistent patient care, and is poised to profoundly influence the future trajectory of AI applications in medicine by enhancing the precision and reliability of the insights based on these powerful computational methods and tools.

The integration of these advanced prompting techniques has revolutionized the application of medical LLMs, exemplified by systems like MedPaLM [42] and MedPaLM-2 [26], which leverage these methods to achieve performances comparable to or even surpass human experts on medical QA datasets. The goal-oriented taxonomy in prompt engineering continues to refine these approaches, focusing on optimizing LLMs' performance by guiding them through structured, logical thinking processes. This not only showcases the broad impact of goal-oriented strategies in prompt engineering but also opens new avenues for further advancements, setting a promising direction for the future of AI in medicine. These developments promise to enhance the precision, efficiency, and reliability of medical care, aligning with broader healthcare goals and significantly influencing the trajectory of AI applications in the medical domain.

2.4. Retrieval-Augmented Generation

The application of Large Language Models (LLMs) in the medical field has been somewhat limited due to challenges such as hallucinations, difficulties in updating knowledge, untraceable reasoning processes, and high costs. Accuracy is crucial in the medical domain, so LLMs often require validation by doctors with specialized knowledge. Particularly with Zero-shot and Few-shot samples, LLMs may produce factual errors that could lead to incorrect diagnoses or treatments for individuals without medical expertise. To address these issues, Retrieval-Augmented Generation (RAG) [43] technology has emerged as a key solution. RAG reduces modeling pain points and improves the trustworthiness of model outputs by integrating knowledge from external databases [8]. It retrieves content relevant to the problem and references it in the output, making the reasoning process more transparent and traceable. For example, generating answers by retrieving specialized medical guidelines and referencing the relevant parts can make the model output more credible and help users locate the corresponding guidelines. Especially for knowledge-intensive tasks, the use of specialized external databases can significantly improve the accuracy of model output. The application of RAG technology in various fields of clinical or medical sciences has great potential. Combining general-purpose

LLMs with semantic understanding capabilities and a specialized medical database may have no less potential than a model fine-tuned with specialized medical data.

Three main paradigms exist for RAG: Naive RAG [44], Advanced RAG [45, 46], and Modular RAG [47-49]. Naive RAG follows a conventional indexing, retrieval, and generation process but faces issues such as low retrieval precision and hallucinatory response generation. Advanced RAG introduces improvements in retrieval accuracy and post-retrieval processing, such as fine-tuning and dynamic embedding models, re-ranking, and prompt compression. Modular RAG extends the adaptability of RAG models by incorporating diverse methods to enhance functional modules, such as search and memory modules, and optimization techniques like hybrid search exploration and recursive retrieval.

In conclusion, the evolution from Naive RAG to Advanced RAG and Modular RAG demonstrates continuous progress in addressing the challenges faced by LLMs in the medical domain. These RAG models enhance the precision, credibility, and applicability of LLMs-generated responses by synergistically merging inherent knowledge with extensive, dynamic external knowledge repositories, paving the way for more effective and nuanced AI applications in healthcare.

2.5. Web Search

In the field of medical language modeling, integrating web search technology can significantly enhance the accuracy and timeliness of models. Although there is currently no practice of combining large language models with Web search technology, it is believed that the internet is the fastest channel for information updates. Compared to constantly maintaining static databases, data supported by the internet has clear advantages in terms of update speed and content breadth. However, this integration also indicates a potential need for stronger internet retrieval and data filtering capabilities. Selecting and refining data from a multitude of web pages is especially critical, which is particularly important in the medical field, as this area demands high accuracy and up-to-date information.

The introduction of interactive web search, as demonstrated in developments such as WebCPM [50], WebGLM [51], and WebGPT [52], marks a significant advancement in this area. WebCPM [50] introduces interactive web search in the Chinese Long-form Question Answering (LFQA) domain, enabling dynamic information retrieval akin to human web search behavior. WebGLM [51], a web-enhanced question-answering system, augments pre-trained language models with web search capabilities, focusing on efficiency and cost-effectiveness. WebGPT [52], on the other hand, fine-tunes GPT-3 [23] for long-form question answering using a text-based web browsing environment, showcasing the ability to navigate the web and collect references to support answers.

Unlike conventional non-interactive retrieval methods, interactive web search allows for a more dynamic and iterative process. Users can decompose complex questions into sub-questions, refine their searches based on the information gathered, and ask follow-up questions. This mimics the cognitive process of human problem-solving and ensures access to a broader range of information, thereby improving the interpretability and relevance of the model's outputs.

In the context of medical large language modeling, the integration of interactive web search can substantially enhance the model's ability to provide accurate and contextually relevant responses. For example, when faced with a complex medical query, the model can engage in real-time interaction with a search engine to gather the latest medical guidelines, research findings, or case studies. This not only ensures that the model's knowledge base is continually updated but also allows for a more nuanced understanding of the query.

Furthermore, the framework proposed in the WebCPM study, which includes a search model for information retrieval and a synthesis model for answer generation, can be adapted for medical LFQA. By fine-tuning pre-trained language models to mimic human web search behaviors and synthesize information from collected facts, a medical LFQA pipeline can be developed. This pipeline has the potential to generate responses that are comparable or even superior to human-written answers in the medical domain.

In conclusion, the combination of web search technology and medical large language modeling holds significant potential for improving the accuracy and reliability of medical LFQA systems. The interactive nature of web search, coupled with the advanced capabilities of RAG models, can lead to the development of more sophisticated and effective medical language models. The LLMs based on these methods and technologies can serve as valuable tools for healthcare professionals, researchers, and patients alike, providing access to timely and accurate medical information.

3. LLMs Applications in Medicine

The existing large language models (LLMs) such as ChatGPT [53-58], Bard [59, 60], LLaMA [61, 62], and so on perform very well in various general domain natural language processing (NLP), but their application in medicine has some limitations [4, 42, 63-66]. The summary of medical LLMs can help relevant medical workers better understand AI cutting-edge technologies and promote the application of LLMs in the medical field. In this paper, Table 1 lists the current common LLMs in the medical field.

3.1. Application of LLMs in Medical text

Medical text information transmits various information through written language, charts, data, and other forms, which is very important information in the medical field. It can make doctors and patients better understand the content of diagnosis and treatment, and facilitate the diagnosis and treatment of diseases. With the development of information technology, digitization and intelligent processing of medical texts have attracted more and more attention. Many LLMs have emerged for medical text information processing and extraction, which have improved the efficiency and accuracy of information retrieval and promoted the dissemination and application of medical knowledge.

Electronic health records (EHRs) [67, 68] contain a large amount of patient information and medical data, which is an important resource in the field of healthcare, but the data is often unstructured and difficult to be directly analyzed and extracted by doctors. To extract and understand key medical information, the researchers put forward the GatorTron model [69], which uses different datasets to extract clinical concepts and analyze similar semantics, and optimizes the language features of clinical texts, which makes it convenient for medical staff to quickly obtain and analyze patient information and assist clinical decision support. Thus, treatment plans can be developed faster and better while improving diagnostic accuracy.

Based on Transformer [70, 71], MedGPT [5], the first medical large language model in China, is built, which mainly aims at giving full play to the practical value of diagnosis and treatment in real medical scenes, and realizing the intelligent diagnosis and treatment ability of the whole process from disease prevention, diagnosis, treatment, and rehabilitation. The model is trained using valid interviews and medical examination data from the electronic health records (EHRs) dataset to make accurate disease diagnosis and design disease treatment plans for patients. Patients can deliver drugs to home through the Internet hospital, and MedGPT will take the initiative to provide patients with medication guidance and management, intelligent follow-up, rehabilitation guidance, and other intelligent disease diagnosis and treatment after receiving drugs.

A study fine-tuned and optimized the large language model LLaMA, by using a dataset of 100,000 doctor-patient conversations from HealthCareMagic, an online medical advice platform, to propose an autonomous ChatDoctor model [6] with a knowledge brain. An autonomous information retrieval mechanism that is added to the model by model fine-tuning and knowledge brain indoctrination strategies, can improve the ability of the model to understand patient inquiries and provide accurate recommendations. And the model based on this autonomous information retrieval mechanism can access and utilize the data of online resources and offline medical databases in real-time as well as answer the latest medical terms and medical questions about diseases. The comparison between ChatDoctor and ChatGPT shows that ChatDoctor performs better than ChatGPT in the answers to new medical terms, drug introduction, disease diagnosis and treatment, etc. However, the model is still in the research stage and is only used for academic research at present. There may be a risk of generating wrong answers when used in clinical practice. Therefore, it is necessary to further develop the model to reduce the possibility of errors and hallucinations, improve the accuracy and efficiency of the model for medical diagnosis, and reduce the task of medical staff.

LLMs have made remarkable progress in understanding and responding to human commands, but they usually perform better in common English language environments and are not specifically trained in the medical field. As a result, the accuracy of the model in auxiliary diagnosis and drug recommendation is insufficient. In order to solve the above challenges, the researchers collected the Chinese medical dialogue databases and fine-tuned the ChatGLM [72] to build a DoctorGLM model [7], which is mainly aimed at Chinese Q&A and dialogue. But it is still in its early stages and may produce incorrect answers, which is not suitable for clinical use and is expected to be further developed and improved.

Due to the huge demand for LLMs in the medical field and the poor performance of the existing models in the medical field, especially in the Chinese environment, researchers proposed the HuatuoGPT model [73], which is specially used for medical consultation. The model combines the advantages of ChatGPT data and doctors' real-world data while mitigating their weaknesses, allowing the model to diagnose like a doctor while having medical knowledge, and providing accurate information to patients. The model-generated answers combine the fluency and information richness of ChatGPT with the professionalism and interactivity of a doctor. It is worth noting that this model has not yet been put into clinical use, and needs further research and development before practical application.

To better serve Chinese users, the researchers developed the MedicalGPT-zh model [74] (later renamed MING), which is based on the Transformer architecture. The model is pre-trained through a large amount of Chinese medical dialogue data to learn the context, terminology, and communication patterns in the medical dialogue. The model can understand and answer medical-related consultation questions through NLP technology, handle complex Chinese medical conversations, and be applied to a variety of scenarios, such as online medical consultation, patient education, health guidance, and so on. The study points out that the development of MING is expected to improve the efficiency and accessibility of medical services, and help promote the development of NLP technology to play a greater role in the field of Chinese medical dialogue in the future.

Based on the LLaMA model, ChatMed [75] series of Chinese medical LLMs improve the performance of Chinese medical dialogue system by integrating medical knowledge to improve the ability to answer medical advice. The ChatMed series models include ChatMed-Consult and ShenNong-TCM-LLM, etc. The ChatMed-Consult model is mainly aimed at online consultation, understanding users' health consultation needs through dialogue, and providing corresponding information and suggestions. ShenNong-TCM-LLM model mainly focuses on the field of Traditional Chinese Medicine (TCM), assisting TCM diagnosis and treatment recommendations, TCM education, and popularization. Another study also constructed HuaTuo model [76] (later renamed BenTsao) based on the LLaMA model by integrating structured and unstructured medical knowledge from the CMeKG and fine-tuning it using knowledge-based instruction data. This model is mainly used in medical research and cannot provide medical advice at present. As a traditional medical system, TCM has rich theoretical knowledge and practical experience. Integrating TCM knowledge into LLMs can improve the accessibility and efficiency of TCM services. Therefore, the researchers constructed the Zhongjing model [77], which is based on the LLaMA model. It enhances the TCM ability of the model through expert feedback and multi-round dialogue in the real world, and ensures its accuracy in the use of terms and concept understanding in the field of TCM, to provide personalized TCM suggestions and treatment plans for patients, and improve the popularity of TCM services.

GLM-130B [78], a bilingual (English and Chinese) pre-training large language model jointly developed by Tsinghua University and Zhipu.AI, has the function of question and answer in the vertical field of medicine, which supports intelligent question and answer of medical and health questions. At the same time, it has developed the auxiliary diagnosis and treatment functions such as generating TCM prescriptions according to symptoms and providing medical explanations for prescriptions.

LLMs have made great progress in the field of medical Q&A, but most of them perform well in multiple-choice questions, and there are still some defects in answering open-ended questions compared with clinicians. The researchers proposed an improved model Med-PaLM 2 [26] based on PaLM 2 [79, 80], which improves the performance of medical reasoning through specific fine-tuning and prompting strategies in the medical field. Compared with clinicians' answers to patients' open-ended questions, the answer of this model can better solve the patients' problems. However, the model still has some limitations: the consistency between the output of the model and the high-quality medical answers expected by patients is not high; the method of evaluating the model needs to cover more dimensions, such as whether the answer reflects humanistic care. At present, LLMs can provide a wide range of health advice in single-round conversations, but the questioning ability in multiple rounds of conversations is insufficient. Researchers proposed a BianQue model [81] based on ChatGLM, which is trained using data from multiple rounds of health conversation to improve the active questioning ability of LLMs. However, the model is limited to academic research and cannot be used in practical clinical applications.

3.2. Application of LLMs in Medical Images

Medical images are an indispensable part of the clinical diagnosis and treatment process. They provide doctors with visual evidence of the disease and can help doctors make more accurate judgments. With the development of information technology, the types and applications of medical images are also expanding, including but not limited to X-rays, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound imaging, and pathological tissue images.

The amount of data in medical images are very large, so it is necessary to separate the regions of interest (such as organs, diseased parts, etc.) from the background, that is, medical image segmentation technology [82-84]. A study proposed a new image segmentation task and model Segment Anything Model (SAM) [85]. However, at present, SAM only performs well in 2D image segmentation, and there are still many problems in processing 3D medical images. To make SAM deal with 3D images better, researchers proposed SAM-Med3D [86], a comprehensively improved 3D architecture model, which is trained on large-scale 3D medical image datasets, and improves the efficiency of the model in 3D medical image segmentation tasks.

A study has proposed the CSCA U-Net model [87], which introduces the channel and spatial compound attention mechanism. It can better capture the channel dependence and spatial relationship in the image, so that it can identify and segment the region of interest in the medical image more accurately, and it enables physicians to propose treatment plans suitable for patients. Some researchers have proposed a RETFound model [88] based on self-supervised learning

(SSL) technology. The model realizes the diagnosis and prognosis of eye diseases and the prediction of complex systemic diseases (such as heart failure and myocardial infarction) by pre-training retinal images and then fine-tuning specific tasks. However, the model currently uses fewer datasets, and it is expected to introduce more data to adjust the model in the future, to improve the performance of model.

Pathological assessment is the gold standard for the diagnosis of many diseases, especially cancer, which mainly relies on doctors' analysis of hematoxylin-eosin (HE) staining and immunohistochemistry (IHC) staining images. A study proposed a new self-supervised learning framework, PathoDuet [89], which enhances the model's understanding of pathological images through cross-scale localization and cross-stain transfer. Studies have shown that the PathoDuet model is effective in most tasks and is expected to be applied in clinical practice in the future. Other LLMs related to medical images are more often combined with text information for output, mostly in the form of multimodal models, which will be introduced in detail later.

3.3. Application of multimodal model in medicine

Most of the previous medical models were developed for a single task, such as the characteristics of a particular disease or image, while medical data are usually multimodal, including images, texts, and other data. These different types of data usually require different processing and analysis methods, so the clinical application of LLMs needs to focus more on the construction of multimodal models [90]. Multimodal models can better understand various forms of medical content, to provide better services for doctors and patients, and improve the accuracy of diagnosis and treatment. To meet these requirements, researchers proposed OpenMEDLab [91], which is an open-source, multimodal basic model platform, using the latest deep learning techniques and algorithms to build the model, so that the model can better meet the actual medical research and clinical applications. Another study proposed a new paradigm of medical artificial intelligence (AI), called generalist medical AI (GMAI) [92]. The GMAI model can perform a variety of tasks with little or no task-specific tagging data. Through self-supervised learning on large and diversified datasets, the model can flexibly accept different medical contents, including images, electronic health records, laboratory results, charts, or texts, and output the results in accurate medical language. However, because it needs a lot of data for training, the cost is high, and it may produce complex output, which makes it impossible for doctors and patients to determine its correctness.

Based on LLaMA-7B, some researchers have combined NLP and computer vision (CV) technology to build an open-source and parameter-efficient biomedical model Visual Med-Alpaca [93]. The model can process and understand visual information and generate biomedical-related text and image content, but the model data is limited to English diagnostic reports. To promote the research and development of the multimodal models of Chinese medicine, a study proposed the XrayGLM model [94], which constructs a Chinese X-rays-diagnostic report dataset by using ChatGPT and public chest X-rays images and texts. It is the first large Chinese multimodal medical model that can view chest X-rays. The XrayGPT model [95] proposed by another study can interpret chest X-rays in a conversational way and answer related questions, providing a new possibility for automated X-rays analysis.

4. Discussion

Although LLMs can assist in medical diagnosis and help doctors solve clinical problems, they still face many challenges in clinical practice scenarios. These challenges include complex data processing, possible hallucinations of models, and potential ethical and privacy issues. The following text will discuss the specific problems in the clinical application of LLMs.

4.1. Complex data processing

The medical field has a very huge data resources, including medical professional books, clinical guidelines, and data in medical records [96]. These data record in detail the characteristics of the disease, diagnostic basis, treatment plan, and individual information of patients such as medical history, medical information, examination results, and imaging reports. There are both structured data and unstructured data in these data. LLMs need to integrate and process these data effectively for more accurate model training. However, when the amount of data used for training is too large, it may lead to an increase in training costs, and the model will also produce complex results due to excessive information input, which not only increases the difficulty of clinical application, but also may affect the efficiency of decision-making. Therefore, optimizing the training dataset and improving the data quality becomes the key step to improving the practicability of the model.

4.2. Hallucination and accuracy

The output results of the LLMs have stability issues, which are easy to produce inaccurate or untrue results (hallucination). One of the reasons may be that the data on which the model relies for training is not novel enough [97]. Due to rapid medical progress, the original data are difficult to ensure the stability and timeliness of the model output, leading to hallucinations. On the other hand, the model training data has not been verified for accuracy. To gain better insight into patients' demands and provide accurate answers, it is necessary to carry out in-depth training on the model, build updated data sets, and use doctor-patient interaction in the real world as training materials to reduce the risk of model output misinformation[6].

4.3. Ethics and privacy

When applying LLMs in the medical field, it is necessary to pay high attention to ethics and personal privacy protection [97]. The development of medical LLMs needs a lot of research to ensure the safety, effectiveness, reliability, and accuracy of its output. Because model training needs to input a large amount of data, which may involve the privacy information of patients, so the processing of these data should strictly comply with the privacy protection principle. In addition, due to the differences in training datasets, models may generate biased recommendations based on race, region, gender, and other characteristics, resulting in unfair results [98]. Therefore, the use of LLMs in different clinical environments requires professional evaluation, standardization, and supervision of medical LLMs to reduce over-dependence on LLMs.

5. Conclusions

LLMs in medicine are based on deep learning and adjust the model through Pre-training, Fine-tuning, Prompting strategies, Retrieval-Augmented Generation, and Web Search (see Figure 1). LLMs train the model by constantly collecting medical data to strengthen the understanding and processing ability of the model to medical texts, images, and other types of data, to assist doctors in disease diagnosis, and provide individual diagnosis and treatment programs for patients. In this review, we introduce the methods of medical LLMs in detail and summarize their wide applications in the medical field, such as the question-and-answer model based on medical text, medical image processing analysis model, and multimodal model integrating various types of data. In addition, we explore many challenges faced by medical LLMs, including the processing of complex data, possible hallucination of models, and potential ethical and privacy concerns. The application of medical LLMs is still in its stages and faces many challenges. For the research and application of medical LLMs, we should pay attention to the following points in the future: (1) Ensure that the dataset used has been strictly clinically verified to ensure the accuracy of the results. (2) The model should only serve as a means to assist doctors in decision-making, considering the complexity of using LLMs for diagnosing and treating diseases, its output should not be used as the only basis for providing medical advice. (3) Efforts should be made to develop more specialized models applicable to subfields in medicine (such as rehabilitation medicine, sports medicine, etc.). With the development of the basic model, medical LLMs are expected to play a more significant role in clinical practice in the future, assisting doctors in diagnosis and treatment, effectively alleviating the pressure on medical resources, and facilitating patients to seek medical treatment. This progress depends on the close cooperation and contact between AI researchers and clinical doctors to overcome the challenges in the application of the model, aiming to integrate models more accurately into the medical field and promote innovative development of medical LLMs.

List of abbreviations

AI: Artificial Intelligence

BLURB: Biomedical Language Understanding and Reasoning Benchmark

BLUE: Biomedical Language Understanding Evaluation

CoT: Chain-of-Thought

CLIP: Contrastive Language-Image Pre-training

CT: computed tomography

CV: Computer Vision

EHRs: Electronic Health Records

GMAI: Generalist Medical AI

GPT: Generative Pre-trained Transformer

HE: Hematoxylin-Eosin

IFT: Instruction Fine-Tuning

IHC: immunohistochemistry

LLMs: Large Language Models

LoRA: Low-Rank Adaptation

MIMIC-III: Medical Information Mart for Intensive Care III

MRI: magnetic resonance imaging

NLP: Natural Language Processing

PET: Parameter-Efficient Tuning

QA: Question Answering

RAG: Retrieval-Augmented Generation

SAM: Segment Anything Model

SFT: Supervised Fine-Tuning

SSL: Self-Supervised Learning

TCM: Traditional Chinese Medicine

ToT: Tree-of-Thought

CRediT authorship contribution statement

Yu Sun: Conceptualization, Data curation, Visualization, Methodology, Writing - original draft, Writing - reviewing and editing. **Guiyan Liu**: Conceptualization, Data curation, Validation, Methodology, Writing - original draft, Writing - reviewing and editing. **Tingyang Xu**: Conceptualization, Validation, Data curation, Resources, Methodology, Writing - original draft, Supervision, Writing - reviewing and editing. **Qifeng Bai**: Conceptualization, Validation, Data curation, Resources, Methodology, Writing - original draft, Supervision, Writing - reviewing and editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We are grateful to “Tencent AI Lab Rhino-Bird Focused Research Program (Tencent AI Lab RBFR2023006, Tencent AI Lab RBFR2022006)” who provides the grants for this manuscript. We also thank the Supercomputing Center of Lanzhou University and Gansu Computation Center for supporting this paper.

References

1. Gu, Y., et al., *Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing.* ACM Trans. Comput. Healthcare, 2021. **3**(1): p. Article 2.
2. Huang, K., J. Altosaar, and R. Ranganath *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. 2019. arXiv:1904.05342 DOI: 10.48550/arXiv.1904.05342.
3. Achiam, J., et al., *Gpt-4 technical report.* arXiv preprint arXiv:2303.08774, 2023.
4. Singhal, K., et al., *Large language models encode clinical knowledge.* Nature, 2023. **620**(7972): p. 172-180.
5. Kraljevic, Z., et al., *MedGPT: Medical Concept Prediction from Clinical Narratives.* ArXiv, 2021. **abs/2107.03134**.
6. Li, Y., et al., *ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge.* Cureus, 2023. **15**(6): p. e40895.
7. Xiong, H., et al., *DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task.* ArXiv, 2023. **abs/2304.01097**.
8. Xiong, G., et al., *Benchmarking Retrieval-Augmented Generation for Medicine.* ArXiv, 2024. **abs/2402.13178**.
9. Zhou, H., et al., *A survey of large language models in medicine: Principles, applications, and challenges.* arXiv preprint arXiv:2311.05112, 2023.
10. Peng, Y., S. Yan, and Z. Lu *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. 2019. arXiv:1906.05474 DOI: 10.48550/arXiv.1906.05474.
11. Lee, J., et al. *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*. 2019. arXiv:1901.08746 DOI: 10.48550/arXiv.1901.08746.
12. Beltagy, I., K. Lo, and A. Cohan, *SciBERT: A pretrained language model for scientific text.* arXiv preprint arXiv:1903.10676, 2019.
13. Li, Y., et al., *BEHRT: Transformer for Electronic Health Records.* Scientific Reports, 2020. **10**: p. 7155.

14. Michalopoulos, G., et al. *UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus*. 2020. arXiv:2010.10391 DOI: 10.48550/arXiv.2010.10391.
15. Yang, X., et al. *GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records*. 2022. arXiv:2203.03540 DOI: 10.48550/arXiv.2203.03540.
16. Chen, Z., et al. *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models*. 2023. arXiv:2311.16079 DOI: 10.48550/arXiv.2311.16079.
17. Jin, Q., et al., *Pubmedqa: A dataset for biomedical research question answering.* arXiv preprint arXiv:1909.06146, 2019.
18. Johnson, A.E., et al., *MIMIC-III, a freely accessible critical care database.* Scientific data, 2016. **3**(1): p. 1-9.
19. He, J., M. Fu, and M. Tu, *Applying deep matching networks to Chinese medical question answering: a study and a dataset.* BMC medical informatics and decision making, 2019. **19**: p. 91-100.
20. Devlin, J., et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv:1810.04805 DOI: 10.48550/arXiv.1810.04805.
21. Liu, Y., et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv:1907.11692 DOI: 10.48550/arXiv.1907.11692.
22. Joshi, M., et al. *SpanBERT: Improving Pre-training by Representing and Predicting Spans*. 2019. arXiv:1907.10529 DOI: 10.48550/arXiv.1907.10529.
23. Brown, T., et al., *Language models are few-shot learners.* Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
24. Zhang, S., et al. *BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs*. 2023. arXiv:2303.00915 DOI: 10.48550/arXiv.2303.00915.
25. Wang, Z., et al. *MedCLIP: Contrastive Learning from Unpaired Medical Images and Text*. 2022. arXiv:2210.10163 DOI: 10.48550/arXiv.2210.10163.
26. Singhal, K., et al., *Towards Expert-Level Medical Question Answering with Large Language Models.* CoRR, 2023. **abs/2305.09617**.
27. Xu, L., et al., *Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment.* arXiv preprint arXiv:2312.12148, 2023.
28. Hu, E.J., et al., *Lora: Low-rank adaptation of large language models.* arXiv preprint arXiv:2106.09685, 2021.
29. Houlsby, N., et al. *Parameter-efficient transfer learning for NLP*. in *International conference on machine learning*. 2019. PMLR.
30. Lin, Z., A. Madotto, and P. Fung, *Exploring versatile generative language model via parameter-efficient transfer learning.* arXiv preprint arXiv:2004.03829, 2020.
31. He, J., et al., *Towards a unified view of parameter-efficient transfer learning.* arXiv preprint arXiv:2110.04366, 2021.
32. Li, X.L. and P. Liang, *Prefix-tuning: Optimizing continuous prompts for generation.* arXiv preprint arXiv:2101.00190, 2021.
33. Sahoo, P., et al., *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications.* arXiv preprint arXiv:2402.07927, 2024.
34. Radford, A., et al., *Language models are unsupervised multitask learners.* OpenAI blog, 2019. **1**(8): p. 9.

35. Mann, B., et al., *Language models are few-shot learners.* arXiv preprint arXiv:2005.14165, 2020.

36. Wei, J., et al., *Chain-of-thought prompting elicits reasoning in large language models.* Advances in neural information processing systems, 2022. **35**: p. 24824-24837.

37. Zhang, Z., et al., *Automatic chain of thought prompting in large language models.* arXiv preprint arXiv:2210.03493, 2022.

38. Liu, H., et al. *LogiCoT: Logical Chain-of-Thought Instruction Tuning*. in *The 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.

39. Hu, H., et al., *Chain-of-symbol prompting elicits planning in large langauge models.* arXiv preprint arXiv:2305.10276, 2023.

40. Long, J., *Large language model guided tree-of-thought.* arXiv preprint arXiv:2305.08291, 2023.

41. Wang, X., et al., *Self-consistency improves chain of thought reasoning in language models.* arXiv preprint arXiv:2203.11171, 2022.

42. Zhou, H., et al., *A Survey of Large Language Models in Medicine: Progress, Application, and Challenge.* ArXiv, 2023. **abs/2311.05112**.

43. Gao, Y., et al., *Retrieval-augmented generation for large language models: A survey.* arXiv preprint arXiv:2312.10997, 2023.

44. Kumar, M., et al., *Artificial Hallucinations by Google Bard: Think Before You Leap.* Cureus, 2023. **15**(8): p. e43313.

45. Ilin, I., *Advanced rag techniques: an illustrated overview*. 2023.

46. Johnson, A.E.W., et al., *MIMIC-IV, a freely accessible electronic health record dataset.* Scientific Data, 2023. **10**(1): p. 1.

47. Lin, X.V., et al., *Ra-dit: Retrieval-augmented dual instruction tuning.* arXiv preprint arXiv:2310.01352, 2023.

48. Yu, W., et al., *Generate rather than retrieve: Large language models are strong context generators.* arXiv preprint arXiv:2209.10063, 2022.

49. Shao, Z., et al., *Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy.* arXiv preprint arXiv:2305.15294, 2023.

50. Qin, Y., et al., *Webcpm: Interactive web search for chinese long-form question answering.* arXiv preprint arXiv:2305.06849, 2023.

51. Liu, X., et al. *Webglm: Towards an efficient web-enhanced question answering system with human preferences*. in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023.

52. Nakano, R., et al., *Webgpt: Browser-assisted question-answering with human feedback, 2021.* URL https://arxiv. org/abs/2112.09332, 2021.

53. Schopow, N., G. Osterhoff, and D. Baur, *Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review.* JMIR Med Inform, 2023. **11**: p. e48933.

54. Qureshi, R., et al., *Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?* Syst Rev, 2023. **12**(1): p. 72.

55. Zangrossi, P., et al., *Large language model, AI and scientific research: why ChatGPT is only the beginning.* J Neurosurg Sci, 2024.

56. Tessler, I., et al., *Advancing Medical Practice with Artificial Intelligence: ChatGPT in Healthcare.* Isr Med Assoc J, 2024. **26**(2): p. 80-85.

57. Touvron, H., et al., *Llama 2: Open Foundation and Fine-Tuned Chat Models.* ArXiv, 2023. **abs/2307.09288**.
58. Leiter, C., et al., *ChatGPT: A Meta-Analysis after 2.5 Months.* ArXiv, 2023. **abs/2302.13795**.
59. Nicholson, A.E., et al., *BARD: A Structured Technique for Group Elicitation of Bayesian Networks to Support Analytic Reasoning.* Risk Analysis, 2020. **42**: p. 1155 - 1178.
60. Abi-Rafeh, J., et al., *Complications Following Body Contouring: Performance Validation of Bard, a Novel AI Large Language Model, in Triaging and Managing Postoperative Patient Concerns.* Aesthetic Plast Surg, 2024.
61. Sandmann, S., et al., *Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks.* Nat Commun, 2024. **15**(1): p. 2050.
62. Touvron, H., et al., *LLaMA: Open and Efficient Foundation Language Models.* ArXiv, 2023. **abs/2302.13971**.
63. Blease, C., et al., *Generative Language Models and Open Notes: Exploring the Promise and Limitations.* JMIR Med Educ, 2024. **10**: p. e51183.
64. Ashraf, H. and H. Ashfaq, *The Role of ChatGPT in Medical Research: Progress and Limitations.* Ann Biomed Eng, 2024. **52**(3): p. 458-461.
65. Abi-Rafeh, J., et al., *Large Language Models and Artificial Intelligence: A Primer for Plastic Surgeons on the Demonstrated and Potential Applications, Promises, and Limitations of ChatGPT.* Aesthet Surg J, 2024. **44**(3): p. 329-343.
66. Ufuk, F., *The Role and Limitations of Large Language Models Such as ChatGPT in Clinical Settings and Medical Journalism.* Radiology, 2023. **307**(3): p. e230276.
67. Melton, G.B., et al., *Electronic Health Records*, in *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, E.H. Shortliffe and J.J. Cimino, Editors. 2021, Springer International Publishing: Cham. p. 467-509.
68. Tsai, C.H., et al., *Effects of Electronic Health Record Implementation and Barriers to Adoption and Use: A Scoping Review and Qualitative Analysis of the Content.* Life (Basel), 2020. **10**(12).
69. Yang, X., et al., *A large language model for electronic health records.* NPJ Digit Med, 2022. **5**(1): p. 194.
70. Lin, T., et al., *A Survey of Transformers.* AI Open, 2021. **3**: p. 111-132.
71. Vaswani, A., et al. *Attention is All you Need*. in *Neural Information Processing Systems*. 2017.
72. Du, Z., et al. *GLM: General Language Model Pretraining with Autoregressive Blank Infilling*. in *Annual Meeting of the Association for Computational Linguistics*. 2021.
73. Zhang, H., et al. *HuatuoGPT, towards Taming Language Model to Be a Doctor*. in *Conference on Empirical Methods in Natural Language Processing*. 2023.
74. Liao, Y., et al., *MING-MOE: Enhancing Medical Multi-Task Learning in Large Language Mode ls with Sparse Mixture of Low-Rank Adapter Experts*.
75. Zhu, W. and X. Wang, *ChatMed: A Chinese Medical Large Language Model*. GitHub.
76. Wang, H., et al., *HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge.* ArXiv, 2023. **abs/2304.06975**.
77. Yang, S., et al. *Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue*. in *AAAI Conference on Artificial Intelligence*. 2023.
78. Zeng, A., et al., *GLM-130B: An Open Bilingual Pre-trained Model.* ArXiv, 2022. **abs/2210.02414**.

79. Chowdhery, A., et al., *PaLM: Scaling Language Modeling with Pathways.* J. Mach. Learn. Res., 2022. **24**: p. 240:1-240:113.

80. Anil, R., et al., *PaLM 2 Technical Report.* ArXiv, 2023. **abs/2305.10403**.

81. Chen, Y., et al., *BianQue: Balancing the Questioning and Suggestion Ability of Health LLMs with Multi-turn Health Conversations Polished by ChatGPT.* ArXiv, 2023. **abs/2310.15896**.

82. Pu, Q., et al., *Advantages of transformer and its application for medical image segmentation: a survey.* Biomed Eng Online, 2024. **23**(1): p. 14.

83. Shi, J., et al. *Predictive Accuracy-Based Active Learning for Medical Image Segmentation*. 2024.

84. Ma, J., et al., *Segment anything in medical images.* Nature Communications, 2024. **15**(1): p. 654.

85. Kirillov, A., et al., *Segment Anything.* 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023: p. 3992-4003.

86. Wang, H., et al. *SAM-Med3D*. 2023.

87. Xin, S., et al., *CSCA U-Net: A channel and space compound attention CNN for medical image segmentation.* Artificial Intelligence in Medicine, 2024. **150**: p. 102800.

88. Zhou, Y., et al., *A foundation model for generalizable disease detection from retinal images.* Nature, 2023. **622**(7981): p. 156-163.

89. Hua, S., et al., *PathoDuet: Foundation Models for Pathological Slide Analysis of H&E and IHC Stains.* ArXiv, 2023. **abs/2312.09894**.

90. Topol, E.J., *As artificial intelligence goes multimodal, medical applications multiply.* Science, 2023. **381**(6663): p. adk6139.

91. Wang, X., et al., *OpenMEDLab: An Open-source Platform for Multi-modality Foundation Models in Medicine.* ArXiv, 2024. **abs/2402.18028**.

92. Moor, M., et al., *Foundation models for generalist medical artificial intelligence.* Nature, 2023. **616**(7956): p. 259-265.

93. Chang Shu, B.C., Fangyu Liu, Zihao Fu, Ehsan Shareghi, Nigel Collier. *Visual Med-Alpaca: A Parameter-Efficient Biomedical LLM with Visual Capabilities*. 2023; Available from: https://github.com/cambridgeltl/visual-med-alpaca/tree/main.

94. Rongsheng, W. and T. Tan, *XrayGLM: The first Chinese Medical Multimodal Model that Chest Radiogr aphs Summarization*. GitHub.

95. Thawakar, O., et al., *XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models.* ArXiv, 2023. **abs/2306.07971**.

96. Jie, R.T.B.Y.a.Y.G.x., *A Review on Research and Application of Medical Large Language Models.* Chinese Journal of Health Informatics and Management, 2023. **20**(06): p. 853-861.

97. Thirunavukarasu, A.J., et al., *Large language models in medicine.* Nat Med, 2023. **29**(8): p. 1930-1940.

98. He, K., et al., *A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics.* ArXiv, 2023. **abs/2310.05694**.

99. Shin, H.-C., et al. *BioMegatron: Larger Biomedical Domain Language Model*. 2020. Online: Association for Computational Linguistics.

100. Johnson, A.E.W., et al., *MIMIC-III, a freely accessible critical care database.* Scientific Data, 2016. **3**(1): p. 160035.

101. Pampari, A., et al. *emrQA: A Large Corpus for Question Answering on Electronic Medical Records*. in *Conference on Empirical Methods in Natural Language Processing*. 2018.

102. Zhang, S., et al., *Multi-Scale Attentive Interaction Networks for Chinese Medical Question Answer Selection.* IEEE Access, 2018: p. 1-1.

103. He, J., M. Fu, and M. Tu, *Applying deep matching networks to Chinese medical question answering: a study and a dataset.* BMC Med Inform Decis Mak, 2019. **19**(Suppl 2): p. 52.

104. Li, J., et al., *Huatuo-26M, a Large-scale Chinese Medical QA Dataset.* ArXiv, 2023. **abs/2305.01526**.

105. BYAMBASUREN Odmaa, Y.Y., SUlZhifang, DAl Damai, CHANG Baobao, LI Suiian, ZAN Hongying, *Preliminary Study on the Construction of Chinese Medical Knowledge Graph.* Journal of Chinese Information Processing, 2019. **33**(10): p. 1-9.

106. Gao, L., et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling.* ArXiv, 2020. **abs/2101.00027**.

107. Yuan, S., et al., *WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models.* AI Open, 2021. **2**: p. 65-68.

108. Paperno, D., et al., *The LAMBADA dataset: Word prediction requiring a broad discourse context.* ArXiv, 2016. **abs/1606.06031**.

109. Hendrycks, D., et al., *Measuring Massive Multitask Language Understanding.* ArXiv, 2020. **abs/2009.03300**.

110. Srivastava, A., et al., *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.* ArXiv, 2022. **abs/2206.04615**.

111. Xu, L., et al. *CLUE: A Chinese Language Understanding Evaluation Benchmark*. in *International Conference on Computational Linguistics*. 2020.

112. Xu, L., et al., *FewCLUE: A Chinese Few-shot Learning Evaluation Benchmark.* ArXiv, 2021. **abs/2107.07498**.

113. *The winograd schema challenge.* Proceedings of the International Conference on Knowledge Representation and Reasoning, 2012: p. 552--561.

114. Wang, A., et al., *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.* ArXiv, 2019. **abs/1905.00537**.

115. Kwiatkowski, T., et al., *Natural Questions: A Benchmark for Question Answering Research.* Transactions of the Association for Computational Linguistics, 2019. **7**: p. 453-466.

116. Geva, M., et al., *Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies.* Transactions of the Association for Computational Linguistics, 2021. **9**: p. 346-361.

117. Talmor, A., et al., *CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge.* ArXiv, 2019. **abs/1811.00937**.

118. Zhou, B., et al., *"Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding.* ArXiv, 2019. **abs/1909.03065**.

119. Jin, D., et al., *What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams.* Applied Sciences, 2021. **11**(14): p. 6421.

120. Pal, A., L.K. Umapathi, and M. Sankarasubbu. *MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*. in *ACM Conference on Health, Inference, and Learning*. 2022.

121. Jin, Q., et al. *PubMedQA: A Dataset for Biomedical Research Question Answering*. in *Conference on Empirical Methods in Natural Language Processing*. 2019.

122. Chen, S., et al., *MedDialog: A Large-scale Medical Dialogue Dataset.* ArXiv, 2020. **abs/2004.03329**.

123. Chen, W., et al., *A benchmark for automatic medical consultation system: frameworks, tasks and datasets.* Bioinformatics, 2022. **39**.

124. Zhang, N., et al. *CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark*. 2022. Dublin, Ireland: Association for Computational Linguistics.

125. Wasserthal, J., et al., *TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images.* Radiol Artif Intell, 2023. **5**(5): p. e230024.

126. Ji, Y., et al., *AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation.* ArXiv, 2022. **abs/2206.08023**.

127. Podobnik, G., et al., *HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset.* Med Phys, 2023. **50**(3): p. 1917-1927.

128. Payette, K., et al., *An automatic multi-tissue human fetal brain segmentation benchmark using the Fetal Tissue Annotation Dataset.* Sci Data, 2021. **8**(1): p. 167.

129. Wang, L., et al., *Benchmark on Automatic Six-Month-Old Infant Brain Segmentation Algorithms: The iSeg-2017 Challenge.* IEEE Transactions on Medical Imaging, 2019. **38**(9): p. 2219-2230.

130. Sun, Y., et al., *Multi-Site Infant Brain Segmentation Algorithms: The iSeg-2019 Challenge.* IEEE Trans Med Imaging, 2021. **40**(5): p. 1363-1376.

131. Mendrik, A.M., et al., *MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans.* Comput Intell Neurosci, 2015. **2015**: p. 813696.

132. Yue, M., et al. *MICCAI Grand Challenge on Multi-domain Cross-time-point Infant Cerebellum MRI Segmentation 2022 : Structured description of the challenge design*. 2022.

133. Heller, N., et al., *The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT.* ArXiv, 2023. **abs/2307.01984**.

134. Baid, U., et al., *The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification.* ArXiv, 2021. **abs/2107.02314**.

135. Quinton, F., et al., *A Tumour and Liver Automatic Segmentation (ATLAS) Dataset on Contrast-Enhanced Magnetic Resonance Imaging for Hepatocellular Carcinoma.* Data, 2023. **8**: p. 79.

136. Jha, D., et al. *Kvasir-SEG: A Segmented Polyp Dataset*. in *MultiMedia Modeling*. 2020. Cham: Springer International Publishing.

137. Vázquez, D., et al., *A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images.* J Healthc Eng, 2017. **2017**: p. 4037190.

138. Bernal, J., J. Sánchez, and F. Vilariño, *Towards automatic polyp detection with a polyp appearance model.* Pattern Recognition, 2012. **45**(9): p. 3166-3182.

139. Silva, J., et al., *Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer.* Int J Comput Assist Radiol Surg, 2014. **9**(2): p. 283-93.

140. Caicedo, J.C., et al., *Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl.* Nature Methods, 2019. **16**(12): p. 1247-1253.

141. Tschandl, P., C. Rosendahl, and H. Kittler, *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.* Scientific Data, 2018. **5**(1): p. 180161.

142. Shu, X., et al., *ECAU-Net: Efficient channel attention U-Net for fetal ultrasound cerebellum segmentation.* Biomedical Signal Processing and Control, 2022. **75**: p. 103528.

143. Gulshan, V., et al., *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs.* Jama, 2016.

144. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.

145. Lotz, J., et al., *Comparison of consecutive and restained sections for image registration in histopathology.* Journal of Medical Imaging, 2021. **10**: p. 067501 - 067501.

146. Liu, S., et al., *BCI: Breast Cancer Immunohistochemical Image Generation through Pyramid Pix2pix.* 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022: p. 1814-1823.

147. Ehteshami Bejnordi, B., et al., *Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer.* JAMA, 2017. **318**(22): p. 2199-2210.

148. Wang, Z., et al., *Foundation Model for Endoscopy Video Analysis via Large-scale Self-supervised Pre-train.* ArXiv, 2023. **abs/2306.16741**.

149. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS Med, 2015. **12**(3): p. e1001779.

150. Johnson, A.E.W., et al. *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. 2019.

151. Demner-Fushman, D., et al., *Preparing a collection of radiology examinations for distribution and retrieval.* J Am Med Inform Assoc, 2016. **23**(2): p. 304-10.

152. Johnson, A.E.W., et al., *MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports.* Scientific Data, 2019. **6**(1): p. 317.

Figures legends

Figure 1. Summary of the methods of the Medical Large Language Model.

Tables legends

Table 1. Summary of LLMs in the medical field, covering its specific tasks and datasets used for training and evaluation.

Table 1. Summary of LLMs in the medical field, covering its specific tasks and datasets used for training and evaluation.

Mode	Models	Tasks	Datasets
Text			
	GatorTron[69]	Quick access to information to aid clinical diagnosis.	The UF Health Integrated Data Repository (IDR), PubMed[99], Wikipedia[99], and Medical Information Mart for Intensive Care III (MIMIC-III)[100], emrQA dataset[101]
	MedGPT[5]	To predict and diagnose diseases and to design treatment plans for patients.	King's College Hospital (KCH), MIMIC-III[100]
	ChatDoctor[6]	Provide accurate advice to patients through self-searching of online and offline medical databases.	HealthCareMagic-100k (www.healthcaremagic.com), iCliniq (www.icliniq.com)
	DoctorGLM[7]	Chinese medical dialogue model.	Translated ChatDoctor's[6] database, Chinese Medical Dialogue (CMD), MedDialog
	HuatuoGPT[73]	Provide detailed, rich content while interacting and diagnosing like a doctor.	HealthCareMagic-100k, iCliniq, cMedQA2[102], webMedQA[103], and Huatuo-26M[104]
	MedicalGPT-zh (MING)[74]	Handle complex Chinese medical conversations and apply to a variety of scenarios, such as online medical consultation, patient education, health guidance, etc.	Chinese medical dialogue dataset constructed based on www.healthcaremagic.com, USMLE cases and other data
	ChatMed[75]	Answer patients' daily medical questions online.	ChatMed Consult Dataset, ChatMed TCM Dataset
	HuaTuo[76] (BenTsao)	Generate accurate and professional medical information in a Chinese context.	Chinese Medical Knowledge Graph (CMeKG)[105]
	Zhongjing[77]	Integrate TCM knowledge into the LLM to provide patients with personalized TCM advice and treatment plans.	CMeKG[105], Chinese Multi-turn Medical Question Answering (CMtMedQA), Huatuo-26M[104]
	GLM-130B[78]	Intelligent question and answer for medical and health problems to assist diagnosis and treatment.	Pile[106], Chinese WudaoCorpora[107], LAMBADA[108], MMLU[109], BIG-bench-lite[110], CLUE[111], FewCLUE[112], Winograd Schemas Challenge (WSC)[113], SuperGLUE[114], Natural Questions[115], StrategyQA[116], Commonsense QA[117], Multiple-choice Temporal Commonsense (MC-TACO)[118]

Table 1 (CONTINUED)

Mode	Models	Tasks	Datasets
Text	Med-PaLM 2[26]	Answer open-ended questions from patients.	MedQA[119], MedMCQA[120], PubMedQA[121], MMLU[109]
	BianQue[81]	The model was trained with multiple rounds of dialogue data to improve its ability of asking questions.	BianQueCorpus, MedDialog-CN[122], IMCS-V2[123], CHIPMDCFNPC[124], MedDG[124]
Image	SAM[85]	2D medical image segmentation.	Segment Anything 1B (SA-1B)
	SAM-Med3D[86]	3D medical image segmentation.	Totalseg-Test[125], AMOS-Val[126], BTCV, HaN-Seg[127], FeTA21[128], FeTA22[128], iSeg17[129], iSeg19[130], MRBrains13[131], MRBrains18, cSeg22[132], KiTS21-Val[133], BraTS21-Val[134], ATLAS[135], TDSC-ABUS
	CSCA U-Net[87]	Accurately identify and segment areas of interest in medical images, so that doctors can propose treatment plans suitable for patients.	Kvasir-SEG[136], CVC-ClinicDB[137], CVC-ColonDB[138], ETIS[139], CVC-T[137], 2018 Data Science Bowl (2018 DSB)[140], ISIC 2018[141], JSUAH-Cerebellum[142]
	RETFound[88]	The diagnosis and prognosis of eye diseases and the prediction of complex systemic diseases.	Moorfields Diabetic imAge dataSet (MEH-MIDAS), Kaggle EyePACS[143], Moorfields AlzEye study (MEH-AlzEye), UK Biobank[144]
	PathoDuet[89]	Understanding and analyzing pathological images.	TCGA, HyReCo[145], BCI[146], NCT-CRC-HE, CAMELYON16[147], IHC dataset
Multi-mode	OpenMEDLab[91]	The basic medical model can be applied to a variety of medical data and solve a variety of clinical and research problems.	SA-Med2D-20M, SNOW, Endo-FM[148], MedFM
	GMAI[92]	Use multiple datasets to learn and flexibly interpret different medical data.	MIMIC[46], UK Biobank[149], UniProt
	Visual Med-Alpaca[93]	Understand visual information and generate biomedical relevant text and image content.	roco-dataset, MEDIQA RQE, MedQA, MedDialog, MEDIQA QA, PubMedQA
	XrayGLM[94]	Provide diagnostic reports by viewing chest X-rays.	MIMIC-CXR[150], OpenI[151]
	XrayGPT[95]	Conversational medical AI for radiation image analysis.	MIMIC-CXR[152], OpenI[151]

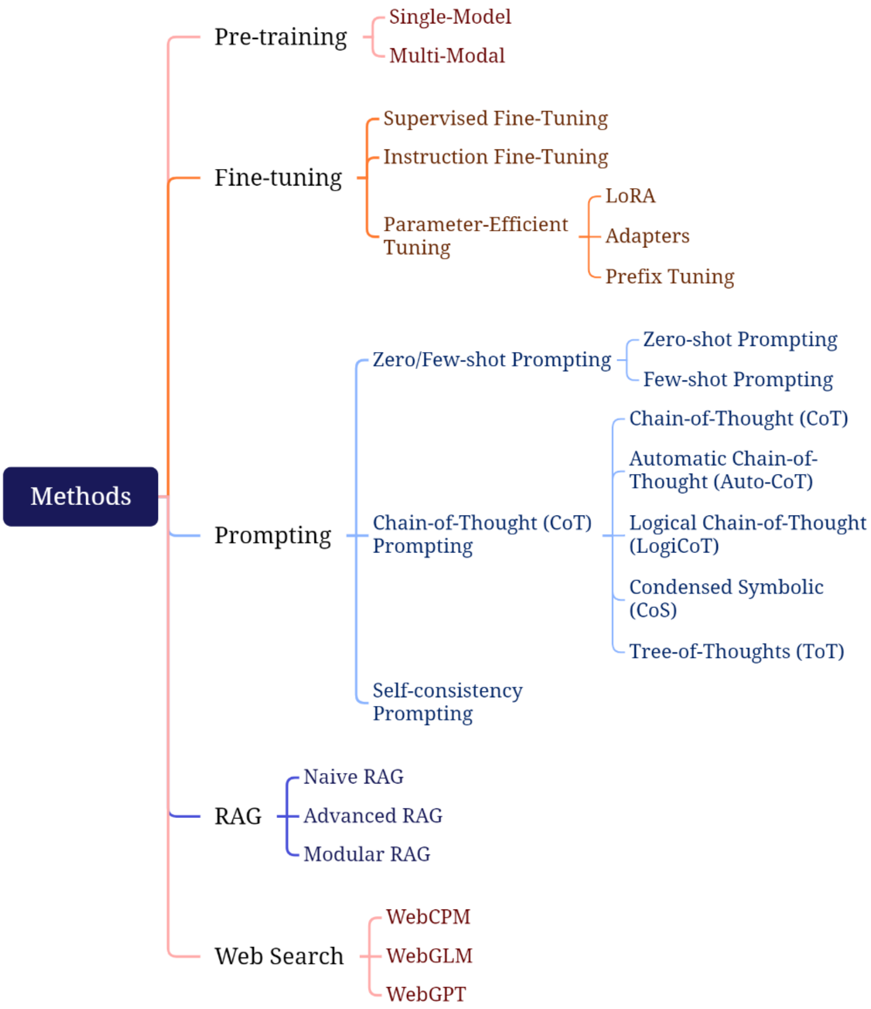

Figure 1. Summary of the methods of the Medical Large Language Model.

Conflict of Interest

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.