Application of large language models in disease diagnosis and treatment

Xintian Yang[1], Tongxin Li[1], Qin Su[1], Yaling Liu[1], Chenxi Kang[1], Yong Lyu[1], Lina Zhao[2], Yongzhan Nie[1], Yanglin Pan[1]

[1]State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers and National Clinical Research Center for Digestive Diseases, Xijing Hospital of Digestive Diseases, Fourth Military Medical University, Xi'an, Shaanxi 710032, China;
[2]Department of Radiotherapy, Xijing Hospital, Fourth Military Medical University, Xi'an, Shaanxi 710032, China.

Abstract

Large language models (LLMs) such as ChatGPT, Claude, Llama, and Qwen are emerging as transformative technologies for the diagnosis and treatment of various diseases. With their exceptional long-context reasoning capabilities, LLMs are proficient in clinically relevant tasks, particularly in medical text analysis and interactive dialogue. They can enhance diagnostic accuracy by processing vast amounts of patient data and medical literature and have demonstrated their utility in diagnosing common diseases and facilitating the identification of rare diseases by recognizing subtle patterns in symptoms and test results. Building on their image-recognition abilities, multimodal LLMs (MLLMs) show promising potential for diagnosis based on radiography, chest computed tomography (CT), electrocardiography (ECG), and common pathological images. These models can also assist in treatment planning by suggesting evidence-based interventions and improving clinical decision support systems through integrated analysis of patient records. Despite these promising developments, significant challenges persist regarding the use of LLMs in medicine, including concerns regarding algorithmic bias, the potential for hallucinations, and the need for rigorous clinical validation. Ethical considerations also underscore the importance of maintaining the function of supervision in clinical practice. This paper highlights the rapid advancements in research on the diagnostic and therapeutic applications of LLMs across different medical disciplines and emphasizes the importance of policymaking, ethical supervision, and multidisciplinary collaboration in promoting more effective and safer clinical applications of LLMs. Future directions include the integration of proprietary clinical knowledge, the investigation of open-source and customized models, and the evaluation of real-time effects in clinical diagnosis and treatment practices.

Keywords: Large language models; Artificial intelligence; Diagnosis; Treatment planning; Clinical decision support

Introduction

Since the release of ChatGPT-3.5 in November 2022, the potential use of this large language model (LLM) in the medical field has attracted widespread interest. LLMs are trained and fine-tuned on the vast knowledge available on the internet, including medical knowledge. Consequently, these models possess superior natural language understanding, pattern recognition, and association analysis capabilities in comparison with traditional language tools.[1] Research has shown that LLMs can pass the United States (US) Medical Licensing Examination.[2] As LLMs rapidly evolve, the newer versions (e.g., GPT-4o and Claude 3.5-sonnet) can not only handle multimodal medical data, but also integrate the latest online research findings and private knowledge bases, providing doctors with diverse information support that may enhance their decision-making abilities.[3] Through interactive question-answering and probabilistic reasoning, in combination with assessments of individual patient characteristics, LLMs can infer and present the pros and cons of different diagnostic and treatment options, potentially facilitating disease diagnosis and treatment decisions.

LLMs can be categorized on the basis of their technical architectures (e.g., autoregressive models like generative pre-trained transformer [GPT] and masked models like bidirectional encoder representations from transformers [BERT]), accessibility (open-source or closed-source), or functionality (general-purpose or domain-specific).[4,5] To enhance the performance of LLMs in professional tasks, researchers employ various optimization techniques, including prompt engineering, post-training techniques, and multi-agent systems.[6,7] For multimodal tasks, such as image interpretation and speech recognition, multimodal LLMs (MLLMs) have been developed to integrate image

Access this article online

Quick Response Code:

Website: www.cmj.org

DOI: 10.1097/CM9.0000000000003456

Correspondence to: Yanglin Pan, State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers and National Clinical Research Center for Digestive Diseases, Xijing Hospital of Digestive Diseases, Fourth Military Medical University, 15 West Changle Road, Xi'an, Shaanxi 710032, China
E-Mail: yanglinpan@hotmail.com

Chinese Medical Journal 2025;138(2)
Received: 16-08-2024; **Online:** 26-12-2024 **Edited by:** Yuanyuan Ji

encoders (e.g., Vision Transformers) and audio encoders (e.g., Wav2Vec) and fuse multimodal features using input projectors.[8] This architecture enables MLLMs to process and understand multiple types of medical data simultaneously. MLLMs have gained significant research attention in the clinical field for their potential in facilitating diagnosis and treatment. The choice of the LLM and optimization technique depends on the researcher's expertise, available resources, and research objectives, leading to diverse approaches for applying LLMs in clinical diagnosis and treatment.

A substantial amount of progress has been achieved in the use of LLMs for the diagnosis and treatment of diseases. LLMs can identify complex symptom patterns on the basis of text information, while MLLMs can simultaneously interpret common medical images and sounds.[9] A growing body of clinical research evidence shows that LLMs are valuable tools for precise grading of common diseases and the diagnosis of rare diseases, predicting disease risks on the basis of individual patient information, and providing personalized treatment recommendations [Figure 1].[10] This review provides an overview of the progress in the use of LLMs for diagnosing and treating diseases across different body systems based on retrospective and prospective studies.

From the perspective of clinical diagnosis and treatment, despite the potential of LLMs in assisting clinical decision-making, significant challenges persist because of the complexity and diversity of diseases. Each patient presents with a unique profile consisting of the genetic background, lifestyle, environmental factors, medical history, and physiological status, which can result in varied symptoms and treatment responses for the same condition.[11] Moreover, personalized medical care should take other factors into consideration, including resource allocation, cost-effectiveness, and ethical considerations.[12,13] Although LLMs have demonstrated the capability to augment diagnosis and management decisions, even in complex

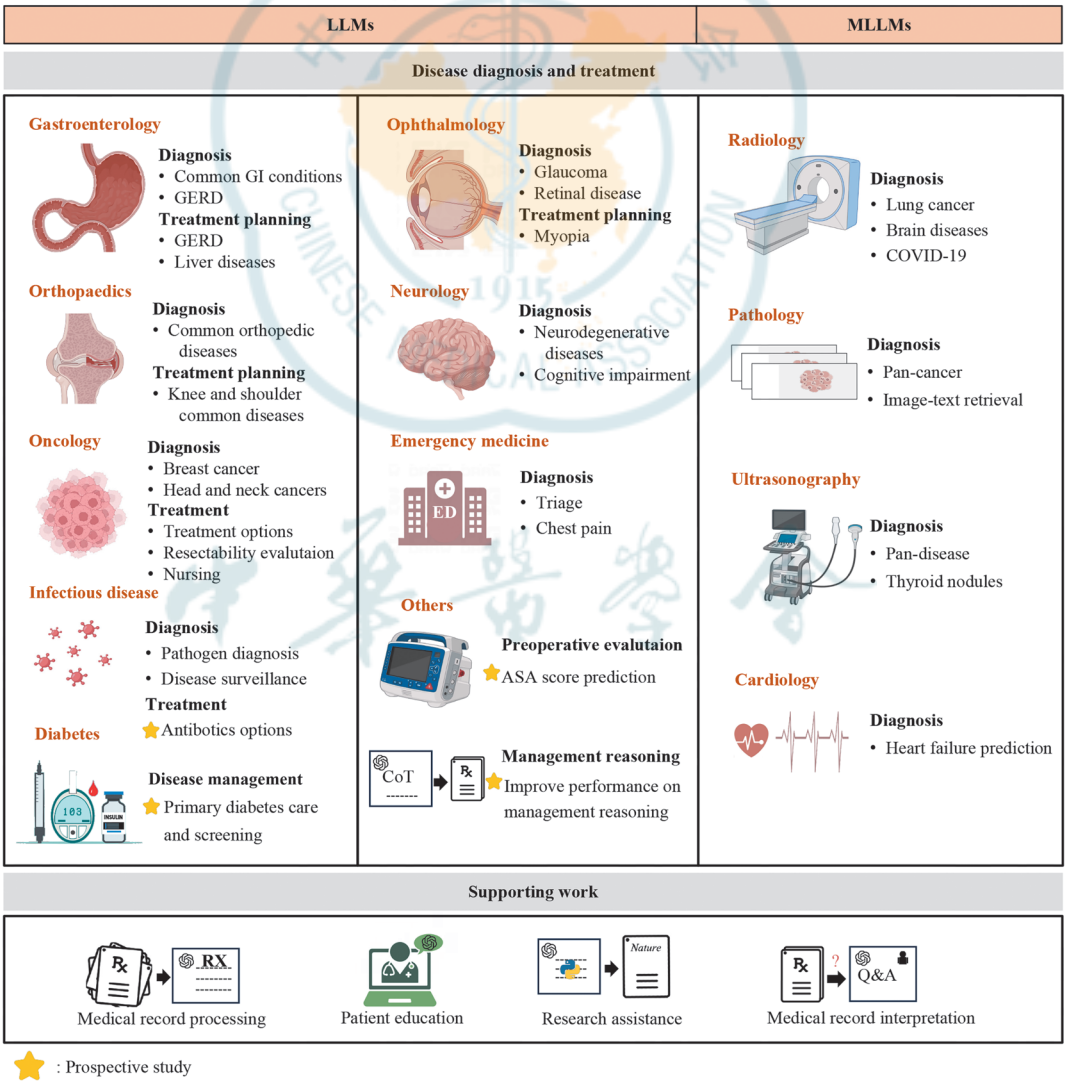

Figure 1: Studies of the applications of LLMs in diagnosis, treatment, and supporting work. ASA: American Society of Anesthesiologists; COPD: Chronic obstructive pulmonary disease; COVID-19: Coronavirus disease 2019; CT: Computed tomography; ECG: Electrocardiogram; GERD: Gastroesophageal reflux disease; GI: Gastrointestinal; HF: Heart failure; LLMs: Large language models; MCI: Mild cognitive impairment; MLLMs: Multimodal LLMs; NSCLC: Non-small cell lung cancer. Created by BioRender.com and Figdraw (www.figdraw.com).

cases,[14,15] further refinement is required to generate truly personalized recommendations that address individual medical requirements. Furthermore, although MLLMs are promising, we believe that they are still in the early stages of interpretation and integration of textual records with non-textual data (e.g., computed tomography [CT] images, breath sounds, and movement videos). The potential clinical benefits of such an integration warrant further investigation. This review also aims to provide a balanced perspective of the limitations and potential of LLMs in clinical practice.

Applications of LLMs in Disease Diagnosis

When LLMs first appeared, they excelled at processing long texts and disseminating common knowledge. Not surprisingly, considerable evidence demonstrates that LLMs hold promise in handling medical record management and interactively explaining medical conditions to patients. These are currently the most proficient and clinically relevant tasks attracting the attention of healthcare professionals. However, the diagnosis and treatment of diseases, which form the core of healthcare, are more complex and variable tasks. Encouragingly, recent research indicates that LLMs can assist healthcare professionals by identifying disease patterns and providing diagnostic suggestions through data analysis. Furthermore, MLLMs have the potential to interpret and integrate non-textual medical data, including images, sounds, and even videos.

The following sections explore the application of LLMs and MLLMs in diagnosis across different clinical departments, highlighting their impact and discussing how current research can further optimize their use in clinical practice.

Gastroenterology

Diagnosis of gastrointestinal (GI) diseases is one of the most important areas of interest for researchers.[16] To evaluate the performance of ChatGPT-3.5 in the diagnosis of GI disease, Lahat *et al*[17] conducted a comprehensive study using representative real-life questions from patients with GI diseases. The researchers submitted 45 diagnostic questions related to GI diseases such as gastritis, esophagitis, and cholelithiasis to GPT-3.5, and three experienced gastroenterologists rated the accuracy, clarity, and usefulness of the responses provided by GPT-3.5 (on a scale of 1–5). The results showed that for diagnostic questions, GPT-3.5's average scores were 3.7 ± 1.7 for accuracy, 3.7 ± 1.8 for clarity, and 3.5 ± 1.7 for usefulness.

The potential of GPT-4 in the differential diagnosis of gastroesophageal reflux disease (GERD) has also been evaluated.[18] The researchers submitted five diagnostic questions to GPT-4, focusing on differential diagnosis and complications of the disease. The results showed that 93.4% of the responses were considered completely appropriate or mostly appropriate.

To investigate the accuracy of LLMs' responses to liver cancer diagnostic questions, researchers posed 20 questions (including 13 diagnostic-related questions) to three LLMs: GPT-3.5, Gemini, and Bing.[19] The results showed that Gemini had the highest accuracy, followed by GPT-3.5 and Bing (60%, 45%, and 30%, respectively), indicating that general LLMs showed suboptimal ability to answer complex diagnosis-related questions when the basic diagnosis was clear.

According to a systematic analysis by Mauro *et al*,[20] accuracy in diagnosing digestive diseases ranged from 6.4% to 45.5% with GPT-3.5, and between 40.0% and 91.4% with GPT-4. However, all of the studies had a high risk of bias, and generalization of single-study results was not recommended. These limitations reflect the current lack of standardized research paradigms and robust evaluation metrics for studies on the use of LLMs in clinical diagnosis.

Neurology

In the field of neurological disease diagnosis, Koga *et al*[21] explored the ability of GPT-3.5, GPT-4, and Google Bard to predict neuropathological diagnoses from clinical summaries. Several LLMs were queried in 25 cases of neurodegenerative diseases and were asked to provide pathological diagnoses and reasons. In comparison to the final clinical diagnoses made by physicians, GPT-3.5, GPT-4, and Google Bard achieved accuracies of 76%, 84%, and 76%, respectively. These findings demonstrated the potential of LLMs in neuropathological diagnosis.

Additionally, to explore GPT-4's performance in the initial screening of mild cognitive impairment (MCI), Wang *et al*[22] collected the data of 174 participants from the Dementia Bank database. Their study found that GPT-4 had a sensitivity of 77.3–86.4% and a specificity of 83.3–94.9%. These findings emphasized that GPT-4 can effectively distinguish potential patients with MCI, but further optimization and standardized prompt engineering by clinicians are needed. Prompt engineering is the process of designing LLM prompts to elicit information that better meets user needs. This may improve the model performance without changing the model parameters, making it more user-friendly and easier to promote.

Emergency medicine

Timely triage and accurate diagnosis are the core responsibilities of emergency doctors in emergency departments (EDs). Research has shown that LLMs have the potential to improve triage efficiency and diagnostic accuracy in emergency patients. Berg *et al*[23] retrospectively explored the ability of GPT-3.5 and GPT-4 to generate early differential diagnoses in 30 patients in the ED. The results showed that without laboratory data, doctors correctly included the final diagnosis in 83% of the cases, whereas GPT-3.5 and GPT-4.0 had accuracy rates of 77% and 87%, respectively. With laboratory data, both doctors and GPT-4.0 had an accuracy rate of 87%, while GPT-3.5's accuracy improved to 97%.

To explore whether the GPT-4 can accurately assess clinical acuity in the ED, a cross-sectional study involving 251,401 adult ED visits was conducted to evaluate

GPT-4's potential to classify patient acuity levels based on 10,000 pairs of Emergency Severity Index scores.[24] The findings revealed that GPT-4 achieved an accuracy of 89%, and in a subset of 500 manually classified pairs, the accuracy of the LLM was comparable to that of the physician reviewer (88% *vs*. 86%).

Chest pain is a common presenting symptom in the ED and is often associated with serious conditions, but is usually benign. Heston and Lewis[25] evaluated GPT-4's ability to stratify risk in simulated chest pain cases. GPT-4's scores were compared with the Thrombolysis in Myocardial Infarction Risk Score (TIMI) and the history, electrocardiography (ECG) findings, age, risk factors, and troponin level (HEART) score, and they showed a high correlation ($r = 0.90$ and 0.93). However, this study also revealed that GPT-4's answers for a fixed score had a inconsistent rate of 45–48% when the LLM was queried repeatedly, highlighting the importance of physician supervision.

Infectious diseases

In infectious diseases, pathogen diagnosis is a key factor that often depends on the physician's experience and subjective judgment. Accurate determination of the presence of an infection and identification of the specific pathogen before successful cultivation are crucial for the subsequent treatment of patients. The applications of LLMs have been explored in the diagnosis of infectious diseases. Perret and Schmid[26] reported the performance of GPT-4 in detecting catheter-associated urinary tract infections in 50 virtual cases. The model demonstrated high performance with a sensitivity of 91% and specificity of 92%. In a prospective cohort study involving bloodstream infections, GPT-4 achieved an accuracy of 59% in 44 prospective cases.[27] This study is discussed in more detail later in the context of LLM applications in infectious disease treatment (see details in Table 1, which summarizes all available perspective studies in LLM-assisted disease diagnosis and treatment).

Additionally, Srikanth *et al*[28] reported that BERTweet, a BERT-based language model, can analyze and categorize information from social media to efficiently monitor the incidence of Lyme disease. This study highlights the broad potential applications of LLMs in public health and preventive medicine.

Cardiology

ECG plays an important role in the timely detection of cardiovascular events. Researchers leveraged an LLM with a public dataset on an ECG-report alignment task based on BioClinical BERT.[29] Then, the network was fine-tuned for heart failure risk prediction using cohorts focusing on patients with hypertension and myocardial infarction. The concordance index (C-index) score of the model for hypertension was 0.63, and for myocardial infarction, it was 0.58, demonstrating BERT's potential in terms of effectiveness and scalability for advancing risk assessment with complex clinical ECG data. In another study, researchers explored the use of LLMs with Retrieval-Augmented Generation (RAG) for zero-shot ECG diagnosis.[30] The study found that GPT-3.5 achieved an accuracy of 75.7% in diagnosing arrhythmias, indicating the strong zero-shot ability of GPT-3.5. RAG can improve the performance of LLMs in specialized tasks without changing the model parameters by constructing domain-specific knowledge bases.

Ophthalmology

The diagnostic capabilities of LLMs on the basis of case texts have been evaluated in ophthalmology. By evaluating GPT-4's performance in diagnosing and managing glaucoma and retinal diseases through case questions, a study demonstrated that LLMs can outperform or match fellowship-trained specialists in accuracy and completeness. This underscores the potential of LLMs as effective diagnostic tools in specialized ophthalmic fields.[31]

Unlike text-based LLMs, MLLMs integrate processing modules for various types of information, such as images and audio. These models fuse features from multimodal data, enabling them to handle a range of multimodal tasks with considerable capabilities. MLLMs' capability in medical image description has recently garnered significant attention from researchers.[32]

Saif *et al*[33] evaluated the diagnostic accuracy of GPT-4V, an MLLM, in recognizing glaucoma using 400 test fundus images from the Retinal Fundus Glaucoma Challenge (REFUGE) dataset. Without specific training, GPT-4V achieved 90% accuracy and 94.44% specificity, but showed a lower sensitivity of 50%. Image preprocessing techniques were further explored: cropping images to focus on the optic disc and peripapillary area significantly improved the sensitivity to 87.50%, while further application of contrast-limited adaptive histogram equalization (CLAHE) resulted in 62.50% sensitivity. However, these preprocessing steps reduced specificity. This study demonstrates the potential of MLLMs in medical image diagnostics, suggesting that they may require less training data and could lead to innovative medical support tools, particularly in resource-constrained settings.

Radiology

To evaluate the potential of LLMs in radiological diagnostics, Dehdab *et al*[9] conducted a retrospective study to assess the performance of GPT-4V in diagnosing chest CT scans, specifically to identify coronavirus disease 2019 (COVID-19), non-small cell lung cancer (NSCLC), and control cases. The results showed that GPT-4V had an overall diagnostic accuracy of 56.8%. The sensitivity for identifying NSCLC was 27.3%, and the specificity was 60.5%; for COVID-19, the sensitivity was 13.6%, and the specificity was 64.3%; and for the control cases, the sensitivity and specificity were 31.8% and 95.2%, respectively. These findings indicate that GPT-4V shows variable diagnostic performance in chest CT interpretation. The variable diagnostic performance of MLLMs is mainly due to constraints such as insufficient and non-diverse

Table 1: Details of prospective studies related to LLM-assisted disease diagnosis and treatment.

References	Design	Specific LLM	Patients	Intervention and/or controls	Primary outcomes	Main findings
Maillard *et al*[27]	Cohort	GPT4.0	44 patients with positive blood culture	LLM-proposed management plan	Rates of optimal, satisfactory, or harmful answers	LLM plans were optimal in 1 (2%), satisfactory in 17 (39%), harmful in 7 (16%)
Turan *et al*[44]	Cohort	GPT4.0	2851 anesthesia outpatients	LLM-assigned ASA scores	ASA scores	High concordance between LLM predictions and anesthesiologists' evaluations
Li *et al*[62]	Cohort	DeepDR-LLM*	487 diabetes patients with referable DR	PCP recommendations + DeepDR-LLM *vs*. PCP alone	Diabetes management adherence	LLM group showed better self-management and adherence to referrals
Goh *et al*[68]	RCT	GPT4.0	92 physicians and residents	GPT-4 + conventional resources *vs*. conventional alone	Management question score	GPT-4 group showed superior management reasoning

*DeepDR-LLM: An integrated image–language system, combining a LLM module and image-based deep learning (DeepDR-Transformer), to provide individualized diabetes management recommendations. ASA: American Society of Anesthesiologists; DR: Diabetic retinopathy; LLM: Large language model; PCP: Primary care physician; RCT: Randomized controlled trial.

training data and model architectures that may not be optimized for image processing. Additionally, challenges in effectively integrating multimodal information and computational resource limitations contribute to the suboptimal performance in understanding images.

To explore the diagnostic capability of MLLMs in integrating text and image information in disease diagnosis, Hirosawa *et al*[34] tested their diagnostic accuracy using 363 case descriptions with images and compared them to GPT-4 without vision. The results showed that the final diagnosis was included in 85.1% of GPT-4V's top 10 differential diagnosis lists, similar to GPT-4's 87.9%. Similarly, Horiuchi *et al*[35] found lower diagnostic accuracies for GPT-4 and GPT-4V than for radiologists when comparing textual input with direct image input. In another study, Sebastian *et al*[36] compared radiology reports prepared by GPT-4 with those written by human radiologists using chest X-ray data. Among the reports prepared by GPT-4, those based on text-based and image-text combined artificial intelligence (AI) systems performed comparably well. The text-based LLM slightly outperformed the radiologist scoring (16.95 *vs*. 15.54), whereas the image–text combined MLLM outperformed most automated evaluation metrics. In contrast, image-only AI-generated reports consistently scored the lowest across all assessments. In another study, researchers created a 3D brain CT dataset consisting of 18,885 text-scan pairs. They then used Clinical Visual Instruction Tuning to train and fine-tune the BrainGPT model.[37] A Turing test was conducted to evaluate BrainGPT and showed that approximately 74% of BrainGPT-generated reports were indistinguishable from those written by humans, demonstrating the model's strong natural language processing capabilities and clinical potential in generating radiology reports.

To improve the reasoning ability of MLLMs in radiological diagnostics, David *et al*[38] introduced a multi-agent system combining CLIP and GPT-4 with multi-agent architectures and prompt-engineering methods. The study showed that the multi-agent system outperformed existing zero-shot methods in chest X-ray diagnosis tasks, particularly for rare diseases, and achieved comparable or superior performance to some supervised methods across multiple datasets. Multi-agent systems can incorporate multiple intelligent agents and information-processing tools, yielding superior performance on complex tasks.

These results suggest that although AI-generated reports approach human-level quality, they have not surpassed radiologist reports. Integrating images with text did not significantly improve GPT-4's accuracy, likely because of the reliance of MLLMs on text features, which limited their diagnostic accuracy in non-textual data scenarios.

Ultrasonography

In a small study conducted by Sultan *et al*,[39] GPT-4V demonstrated high accuracy in analyzing and interpreting thyroid and renal ultrasound images, and it was able to identify and mark lesions on the images. In another study, researchers investigated the diagnostic performance of LLaVA-Ultra, a multimodal Chinese language and visual assistant developed for ultrasound medicine.[40] In terms of performance, LLaVA-Ultra outperformed existing state-of-the-art models on the US-Hospital dataset, achieving an accuracy of 62.0% and an F1 score of 72.2%. It also excelled in CT, magnetic resonance imaging (MRI), and X-ray subsets, achieving up to 83.8% accuracy and an F1 score of 93.4%.

In a retrospective study, researchers integrated LLMs combined with image-to-text conversion technology with

a traditional deep learning (DL) model for the diagnosis of thyroid nodules.[41] The study indicated that while LLMs may underperform in comparison with DL models in medical image diagnosis, their transparency and interpretability offer significant value for medical education and clinical decision-making. MLLMs utilize transformer-based architectures and require diverse multimodal datasets, whereas DL models typically employ task-specific architectures, such as convolutional neural networks (CNNs), and require only single-modality data. DL models often achieve high accuracy for specific tasks and are relatively easy to deploy locally. In contrast, MLLMs offer coverage across multiple medical systems, zero-shot learning capabilities, and enhanced human–computer interactions. MLLMs generally have higher computational demands and may face additional privacy considerations, whereas DL models are typically more resource-efficient and straightforward to implement for privacy protection. These differences between MLLMs and DL models reflect their distinct approaches to image-based diagnosis, with each offering unique characteristics that can contribute to diagnostic capabilities in medical imaging.

Pathology

In the field of pathology, the emergence of fundamental MLLMs has contributed significantly to the use of generalized AI for pathological image analysis. PathChat, a multimodal language model that was developed and evaluated for pathology,[6] combines a visual encoder pre-trained through self-supervised learning with the Llama 2 LLMs consisting of 1.3 billion parameters, and fine-tuned on over 456,000 vision-language instructions. The results showed that PathChat performed well on multiple-choice diagnostic questions, with an accuracy of 78.1% without clinical context. With the addition of clinical context, PathChat's accuracy further increased to 89.5%. Xu *et al*[42] also developed a foundational whole-slide pathology model, Prov-GigaPath, which was pre-trained on 171,189 whole-slide images and demonstrated state-of-the-art performance across various digital pathology tasks.

Zheng *et al*[43] conducted a benchmark evaluation of the performance of the PathCLIP model in pathological image analysis. PathCLIP is a contrastive language-image pretraining model specifically designed for pathology and trained on a dataset containing over 200,000 image–text pairs. PathCLIP is beneficial for facilitating image–text retrieval and assisting in diagnosis in pathology.

Others

In addition to the aforementioned fields, the diagnostic potential of LLMs in other areas has been explored. In a perioperative evaluation, a prospective multicenter study by Turan *et al*[44] evaluated the accuracy of GPT-4 in predicting the American Society of Anesthesiologists (ASA) score for perioperative assessment. This study included 2851 patients and compared the assessment of GPT-4 with those of expert anesthesiologists. The results showed that GPT-4 had over 90% accuracy in cases with ASA scores I–III, but showed significant differences in the evaluation of complex health conditions for ASA IV scores. This study indicated that while GPT-4 shows high predictive ability in some aspects, it still has room for improvement in dealing with complex cases. Additionally, LLMs have demonstrated potential in evaluating postoperative complications by monitoring patient data and detecting early signs of complications, thereby enhancing recovery quality.[45] Research showed that GPT-4 accurately replicates the Clavien–Dindo Classification definitions and analyzes complications with an accuracy rate of up to 97%. When handling real-world discharge summaries, GPT-4 achieved a Cohen's κ value of 0.92, demonstrating near-perfect consistency with human assessments.

In addition, Duong and Solomon[46] found no significant difference ($P = 0.83$) in GPT-3.5's performance in answering medical genetics questions in comparison with human respondents, with accuracy rates of 68.2% for GPT-3.5 and 66.6% for humans. While LLMs have limitations in calculating recurrence risks for individual genetic disorders, their potential in transmitting genetic information remains strong.[47] However, in pediatric diagnosis, GPT-3.5 made errors in 83 out of 100 cases, showing limited diagnostic value, although newer models like GPT-4 require further investigation in this regard.[48] For orthopedic diagnoses, the addition of specific symptoms improved diagnostic accuracy, with varied results observed across five common conditions.[49]

For clinical supporting tasks, Huang *et al*[50] found that GPT-3.5 achieved 89% accuracy in extracting structured data from over 1000 case reports. In another study, Amin *et al*[51] explored the effectiveness of four LLMs, including GPT-4, in simplifying radiology reports, such as CT and MRI reports. LLMs showed promise in reducing the reading difficulty of medical reports, making it easier for patients to understand their test results. LLMs are also being increasingly used in medical research to draft documents, edit papers, and generate statistical codes, thereby enhancing research efficiency and streamlining data analysis processes.

Applications of LLMs in Disease Treatment

The existing clinical research on the use of LLMs for therapeutic applications has been primarily based on retrospective or observational studies across multiple clinical disciplines. LLMs have shown significant potential in generating simulated treatment plans based on patient medical records, interpreting diagnostic data, and providing treatment recommendations to patients.

Oncology

Cancer treatment strategies are usually tailored on the basis of the specific cancer type, genetic profile, and stage of the disease, necessitating increasingly personalized therapeutic approaches.[52] Researchers have explored the use of LLMs to develop customized treatment plans for patients across various cancer types.

An observational study compared the consistency of treatment recommendations for 20 cases of breast cancer

between GPT-3.5 and a multidisciplinary team (MDT).[53] In specific treatment options, GPT-3.5 demonstrated high consistency with the MDT in surgery, chemotherapy, and radiotherapy recommendations (95%, 94.5%, and 95% respectively), whereas the consistency was lower in gene therapy recommendations (70%). Schmidl *et al*[54] compared GPT-3.5 and GPT-4 with an MDT in providing treatment recommendations for 20 patients with head and neck cancer. GPT-3.5 and GPT-4 provided general answers for surgery, chemotherapy, and radiation therapy with moderate consistency in comparison with the MDT. In the proposed treatment plans, GPT-3.5 aligned closely with the recommendations of the MDT in 90% of the cases. In contrast, GPT-4 presented a more diverse range of treatment options, averaging 5.1 more options than GPT-3.5 and significantly exceeding the MDT's scope.

For personalized cancer treatment, Manuela *et al*[15] evaluated four LLMs in planning treatments for 10 fictional cancer cases. The LLMs' F1 scores (0.04–0.19) were inferior to those of human experts, and their recommendations were more easily identified as AI-generated (median, 7.5 *vs*. 2.0 for manually annotated cases). The LLMs showed promise in providing complementary insights, with at least one LLM generating a helpful recommendation for each case and even offering two unique and useful treatment strategies. By combining recommendations from different LLMs, the F1 score improved to 0.29, indicating the scope for future enhancement. The study concluded that although LLMs show potential in cancer treatment planning, further improvements are needed to match expert-level performance. In a separate study, Krückel *et al*[55] found that GPT-3.5 was capable of offering personalized advice for gynecological malignancies on the basis of patient-specific symptoms, such as recommendations for pleural or ascites drainage and interdisciplinary care.

To deal with the difficulties in explaining treatment to patients, a cross-sectional study evaluated the capability of LLMs to address patient queries within the realm of radiation oncology.[56] The results showed that when comparing answers specific to treatment modalities, LLMs exhibited similar performance to expert answers in terms of accuracy, completeness, and conciseness. Lawson *et al*[57] also discussed the potential of ChatGPT and Bard in explaining neuro-oncology treatment plans and found that they could reduce information pressure on healthcare workers and enhance patient self-management.

Gastroenterology

LLMs have also been evaluated for their potential in the assessment of treatment plans across a range of GI disorders, including liver diseases, esophageal conditions, and other digestive tract ailments.

Giuffrè *et al*[7] reported that LLMs can provide treatment protocols for chronic hepatitis C combined with human immunodeficiency virus (HIV) infection, taking into account the different conditions of patients. Henson *et al*[18] found that GPT-4 performed well in answering questions related to the treatment of GERD, with 93.9% of its responses being either mostly or completely appropriate, and 75.8% containing specific guidance. The GPT-4 responses were well understood by 97.7% of the patients. A study showed that GPT-4 was 83% accurate in answering six treatment-related questions out of 15 for irritable bowel syndrome.[58] Of the 20 questions for inflammatory bowel disease, GPT-4 showed 85% accuracy for seven treatment-related questions. Adi *et al*[17] further evaluated GPT-4's performance in answering 110 real patient questions about GI health. For 42 treatment-related questions, GPT-4 scored 3.9 ± 0.8, 3.9 ± 0.9, and 3.3 ± 0.9 on average (out of 5) for accuracy, clarity, and effectiveness, respectively, but the quality of the answers varied significantly across treatment topics.

In addition to evaluating the treatment of GI diseases using general-purpose commercial large models, Jin *et al*[59] tried to use RAG to develop "LiVersa", a specific LLM for liver disease, and evaluated its performance in hepatitis B treatment and hepatocellular carcinoma surveillance issues. They found that LiVersa's outputs were more accurate but were rated less comprehensive and safe compared to those of GPT-4.

Geriatrics

The management of chronic diseases is a critical issue, especially in an aging society. Recent studies have explored the potential of LLMs in providing clinical support for various chronic conditions that commonly affect older adults.

Rao *et al*[60] investigated the application of GPT-3.5 in medication management for older patients. They found that GPT-3.5 consistently advised reducing medication in older patients with impaired activities of daily living and no history of cardiovascular disease. In contrast, for older patients with a history of cardiac events, GPT-3.5 exhibited greater caution in adjusting medication regimens, with 56% of the responses advising against reducing medications. Expanding the scope beyond medication management, Imtiaz *et al*[61] reported that GPT-3.5 often outperformed Microsoft Bing in providing treatment options for chronic obstructive pulmonary disease, a condition prevalent in the elderly, although the results indicated the need for updates to align with the latest medical guidelines. Similarly, another study on myopia, which can progress with age, found that GPT-4 significantly outperformed both GPT-3.5 and Google Bard in providing treatment and prevention information. Collectively, these results indicated that LLMs may provide useful clinical support for managing chronic diseases that commonly affect the elderly population.

A notable prospective study on diabetes, a condition prevalent among the elderly, explored the integration of LLMs with image-based DL.[62] This study evaluated the effectiveness of a system named DeepDR-LLM in 487 patients newly diagnosed with diabetes and referable diabetic retinopathy (DR). The authors compared the outcomes between patients receiving care from primary care physicians (PCPs) alone and those benefiting from PCPs augmented by the DeepDR-LLM. The primary focus was on adherence to diabetes management,

including self-management behaviors and DR referral compliance. The results showed that patients in the PCP + DeepDR-LLM group exhibited better self-management behaviors and were more likely to adhere to DR referrals. This study indicates the promising prospects for AI-enhanced healthcare in improving chronic disease management and treatment adherence in the elderly population.

Infectious diseases

The use of pathogen-sensitive drugs is crucial for the treatment of infectious diseases. Studies have also reported the potential applications of LLMs in this field.

Researchers found that the empirical antimicrobial therapy recommendations prospectively provided by GPT-4 for patients with positive blood cultures were deemed appropriate in 64% of cases, but caused harm in 2% of cases.[27] For the final antibiotic therapy, the suggestions by GPT-4 were considered adequate in 91% of patients, but still resulted in harmful effects in 5% of cases. This prospective study highlights the limitations and potential risks of using AI for medical consultations. Another study found that GPT-3.5 generally aligns with diagnoses in providing antimicrobial treatment advice when presented with eight hypothetical infection scenarios.[63]

Orthopedics

Accurate diagnosis and appropriate treatment of orthopedic diseases are essential for patient recovery and long-term well-being.[64] The performance of LLMs in the treatment and management of joint conditions and osteoarthritis has been evaluated.

A study evaluated the performance of GPT-4 in providing treatment recommendations for common knee and shoulder orthopedic conditions based on 20 MRI reports.[65] When assessing the clinical usefulness and relevance of the treatment recommendations, the doctors reported "agree" and "strongly agree" for 60% and 20% of the recommendations, respectively. However, doctors also pointed out that GPT-4 does not fully consider specific patient situations and the urgency of treatment. Another study reported that a GPT-4-based tool using RAG and instruction prompting significantly surpassed other general LLMs, such as GPT-4 and GPT-3.5, in the management of osteoarthritis.[66]

Other fields

In addition to the aforementioned fields, researchers have explored the application of LLMs to other aspects of clinical treatment. A recent randomized controlled trial (RCT) investigated GPT-4's influence on physicians' management reasoning.[14] This study involved 92 practicing attending physicians and residents, using virtual case scenarios. Participants were randomly assigned to two groups: one using GPT-4 with conventional resources and the other using conventional resources alone. The primary outcome measure was the intergroup difference in the total scores when answering management questions. The results showed that physicians using GPT-4 alongside traditional resources achieved superior performance in management reasoning in comparison with those using conventional resources alone. This finding suggests that integrating advanced LLMs such as GPT-4 into clinical practice may enhance physicians' decision-making capabilities and improve patient care outcomes. The use of virtual case scenarios in this study allowed for a controlled environment to assess the impact of GPT-4 on management reasoning while also providing a standardized set of clinical challenges across participants. This methodological approach enhanced the reliability and generalizability of the findings and offered valuable insights into the practical applications of LLMs in real-world clinical settings.

In addition to exploring new applications, researchers have examined various strategies to improve the output quality of LLMs. Notably, studies have shown that LLMs can provide more accurate responses when given the opportunity to self-correct, highlighting the potential of using chain-of-thought prompting with LLMs in clinical support systems.[67] This finding suggests that allowing LLMs to engage in a more iterative and reflective process could substantially enhance their effectiveness in healthcare applications.

Although the abovementioned studies have shown the beneficial potential of LLMs in assisting disease treatment, real-world case studies can be helpful for a more in-depth analysis of how LLMs contribute to personalized treatment plans. Two examples are presented for further reference [Supplementary Table 1, http://links.lww.com/CM9/C286]. The first example was of an obese patient with hypertriglyceridemia-induced acute pancreatitis. In such cases, physicians may benefit from LLM assistance in addressing several challenging aspects of care. These challenges include managing intractable pain, implementing early enteral nutrition in cases complicated by gastric outlet obstruction, treating infected walled-off necrosis, managing refractory hypertriglyceridemia, and preventing recurrent episodes. Doctors can refer to LLM-generated suggestions to make informed clinical decisions. Another illustrative example is a patient with advanced esophageal squamous cell carcinoma complicated by esophageal obstruction. In this scenario, LLMs can simulate a multidisciplinary tumor board by integrating perspectives from various fields such as radiation oncology, medical oncology, radiology, thoracic surgery, nutrition, and endoscopy. This virtual consultation can provide comprehensive, multifaceted, and timesaving treatment recommendations. Considering the specificities of individual cancer patients, qualified oncologists should make the final medical decisions. The valuable insights and suggestions provided by LLMs can be helpful in this process.

To date, studies (especially RCTs) related to LLM-assisted treatment are limited. This can be attributed to two factors. First, unlike diagnostic reasoning, which often involves a classification task with a single correct answer, treatment reasoning is more complex and lacks definitive answers. This process requires balancing trade-offs between inherently risky options.[14] Second, assessing

treatment efficacy is more difficult and time-consuming. Although LLM assistance may offer substantial benefits, these effects are often minimal and a large sample size is required to demonstrate the superiority of LLM-assisted treatment planning. Thus, the existing literature cannot clarify whether LLM assistance can improve disease treatment outcomes, such as prolonging survival in malignant tumors or improving the symptoms of common diseases (e.g., functional GI disorders), in comparison with standard treatments. Based on the aforementioned research results, further exploration of the impact of LLM-assisted treatment on patient-centered outcomes in well-designed clinical studies can be expected to yield meaningful findings.

Limitations

Performance and reasoning limits

A key issue associated with the use of LLMs in clinical practice is the accuracy of their inquiries and answers. Hallucinations in LLMs may produce incorrect or even harmful information, although this issue has shown improvement in newer iterations of models.[68] Nevertheless, the performance drift observed in some LLMs still lacks sufficient theoretical explanation, potentially raising concerns about the safety and stability of LLMs in clinical applications.[69]

Currently, MLLMs such as GPT-4V demonstrate some ability to interpret medical X-ray images but still lack sufficient capabilities with some imaging modalities such as MRI.[70] This limitation is likely due to insufficient and non-diverse training data and model architectures not optimized for image processing. In addition, MLLMs tend to prioritize textual information over visual inputs, which limits their diagnostic accuracy in image-based medical tasks. Challenges in effectively integrating multimodal information and computational resource limitations further hinder the ability of these models to understand images. To enhance the utility of MLLMs in medical image interpretation, optimizing and fine-tuning the image feature-processing modules on the basis of multimodal data is crucial. Future developments should focus on improving data integration and cross-modal information synthesis capabilities to overcome the current limitations and achieve a better understanding of medical images for more accurate diagnoses.

Moreover, LLMs have been reported to show insufficient reasoning abilities, which may limit their performance in challenging cases.[71] LLMs often lack the ability to organize multi-round inquiries for diagnostic purposes owing to the limited information provided by patients in each turn, thus constraining their application in medical consultation.[72] Clinical diagnosis and treatment require high reasoning and logical capabilities to analyze multi-system symptoms and disease progression. The limitations of LLMs in reasoning may also lead to difficulties in discerning the authenticity of patient descriptions or cause overreliance on previous diagnostic results. Fortunately, LLMs with enhanced reasoning capabilities are continuously emerging. The recently released OpenAI O1 model exemplifies this trend. Such advancements are expected to further improve diagnostic prediction and treatment planning in future research.

Limits in accessing and integrating clinical information

The effectiveness of LLMs in medical diagnosis relies heavily on access to comprehensive, accurate, and timely clinical information. However, the current healthcare landscape presents significant challenges in this regard. Numerous laboratory and radiological examinations generate diverse types of medical data. The existence of disparate, often non-interoperable, medical information systems creates information silos within and across healthcare institutions.

This fragmentation of clinical data also poses a substantial obstacle to the seamless integration of LLMs into healthcare settings. To address the challenges of implementing LLMs in medical diagnostics effectively, a multifaceted and comprehensive strategy is required. The primary task is to promote unified data exchange standards, real-time synchronization mechanisms, and open application programming interfaces (APIs) to achieve seamless integration of existing electronic health-record systems. Simultaneously, strengthening data security and privacy protection measures is also crucial. Research has shown that LLMs have the potential to organize medical reports and records;[50,51] thus, they can be used to structure complex medical information. Optimizing human–machine interfaces, such as intelligent voice assistants and augmented reality (AR) technology, will significantly enhance the collaboration efficiency between doctors and LLMs. Furthermore, promoting industry standardization, improving legal and ethical frameworks, and enhancing AI-related training for medical personnel are indispensable. However, the implementation of these measures requires collaborative efforts from medical institutions, technology companies, regulatory bodies, and academia to fully leverage the potential of LLMs in medical diagnostics and ultimately improve the quality and efficiency of healthcare services.

Lack of general evaluation standards

As highlighted in a systematic review by Mauro *et al*,[20] the current landscape of LLMs presents significant challenges in terms of their comparability and assessment across various clinical diagnostic studies. An important contributing factor to this issue is the lack of standardized evaluation criteria. The inherent diversity of evaluation methods for different diseases and stages of clinical progression poses substantial challenges in this regard. Evaluation metrics that rely on human experts mainly include diagnostic accuracy, readability, or subjective scores based on specific task scenarios.[17,19,41,48] Studies with metrics relying on expert evaluation are often limited by small sample sizes for LLM testing due to the scarcity of experienced professionals.[73] Meanwhile, human-based evaluation metrics may introduce bias and variability into assessments. On the other hand, automatic evaluation

metrics based on lexical overlaps, such as Bilingual Evaluation Understudy (BLEU) and metrics for the evaluation of translation with explicit ordering (METEOR), struggle with weak correlations to human evaluations.[74,75] These discrepancies may raise questions regarding the effectiveness of automatic evaluation metrics. Furthermore, the disparate experimental designs and evaluation criteria employed across studies render comparisons between research findings difficult. This lack of standardization highlights the need for more robust and universally applicable evaluation frameworks.

To address these challenges and further enhance the reliability and applicability of LLMs in medical diagnostics, the development of more standardized test-question datasets and an increase in high-quality RCTs using consistent evaluation methodologies represent promising avenues. These approaches can potentially resolve the current limitations, provide more robust and comparable evidence for the efficacy of LLMs in healthcare settings, and facilitate more meaningful cross-study comparisons, thereby accelerating their integration into clinical practice.

Legal and ethical risks

The potential legal risks of AI in healthcare have been discussed extensively. The primary concerns include who should use them, how they should be used, and who should be responsible. As mentioned above, the answers provided in the diagnosis of chest CT images and antibiotic treatment recommendations are prone to errors.[9,27] Although some LLMs have performed well in medical consultations, users without a clinical knowledge background may still face adverse health effects and legal risks when using LLMs to seek healthcare advice.

Furthermore, GPT-4 has been reported to potentially leak user information, raising concerns about information safety.[76] Some researchers have explored deploying local models to annotate radiology reports to address patient privacy constraints.[77] However, the cost and technological requirements of deployment and maintenance of high-performance LLMs make it challenging to deploy them locally in hospitals to protect patient information privacy. Considering these issues, a promising approach involves leveraging capable third-party service providers operating under robust regulatory frameworks. These providers can implement federated learning techniques or stringent confidentiality measures, potentially mitigating data security concerns while harnessing the power of advanced language models for healthcare applications.

Issues related to fairness and bias

The deployment and service of LLMs typically require a high-quality information infrastructure, potentially creating barriers for residents in developing regions to access AI-supported healthcare, which may exacerbate inequalities in medical resource distribution. If LLMs are widely used in future clinical work, local governments will need to implement effective regulations for service providers to ensure stable accessibility policies and access channels in case of sudden shortages of medical resources and healthcare panic.

In addition, because model training is based on historical data, LLMs may perpetuate and amplify biases. A study has shown that some models exhibit conversational biases across different races and genders, possibly stemming from limitations in training data.[13] To address these concerns, developers must undertake more rigorous data-cleaning processes to eliminate biased data from training sets. Additionally, implementing robust bias-detection mechanisms for model outputs could contribute to reducing the generation of biased content. These measures are crucial to ensure that AI-supported healthcare systems promote equity and avoid reinforcing existing societal disparities.

Perspectives

Development trends and technological approaches

LLMs became the focus of AI-assisted diagnostics and therapeutics after the release of GPT-3.5 and GPT-4. Currently, the best general LLMs are closed-source models released by companies such as OpenAI and Anthropic. These companies are actively collaborating with biomedical researchers to develop AI tools for clinical healthcare and drug discovery. Some open-source models, such as Llama 3.1 and Qwen 2.5, are also gaining attention because of their open technology and good performance and are being used by an increasing number of researchers to develop medical applications.[78] Customized LLMs are expected to be deployed and assist medical staff in handling clinical work.

Considering the current scarcity of training data, one effective method for improving the performance of LLMs in clinical diagnosis and treatment is post-training technology, including fine-tuning, RAG, and reinforced learning [Figure 2].[7] Besides post-training technology, multi-agent systems can excel in specific tasks but may face challenges in multidisciplinary clinical applications and widespread deployment.[38] Using pretrained models, fine-tuning can effectively improve LLMs' performance in domain-specific tasks with less data. Prompt-engineering techniques can potentially further enhance the performance of general LLMs without changing their parameters. Notably, these approaches are not exclusive or contradictory but can be developed and used collaboratively and iteratively. Researchers can select the appropriate technical routes based on their tasks, technological capabilities, and hardware resources.

Future of MLLMs

With the development of MLLMs, researchers are increasingly exploring their diagnostic potential in medical imaging.[9] In comparison with DL models, MLLMs are attracting researchers' attention for generating medical reports and interpreting medical images due to their text-generation capabilities. Studies on treatment aspects are currently rare in comparison with those on diagnostics. This disparity could be related to the longer treatment cycles in clinical trials. Moreover, the medical risks and

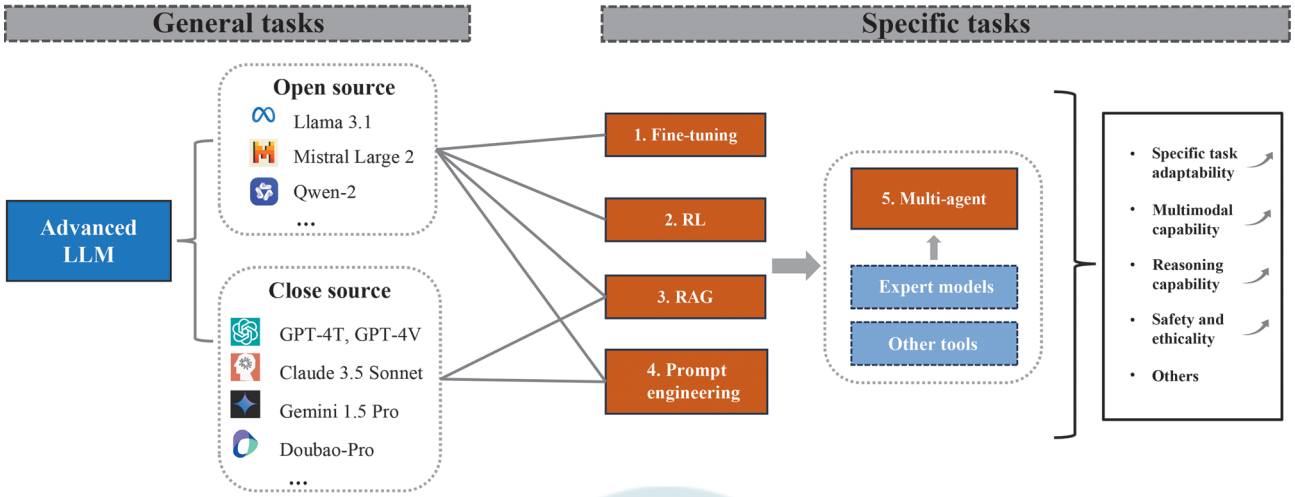

Figure 2: Roadmap of common techniques for customizing LLMs. LLMs: Large language models; RAG: Retrieval-Augmented Generation; RL: Reinforcement learning.

ethical concerns associated with conducting clinical trials based on MLLM-generated treatment plans may make researchers cautious and conservative.

Since clinical diagnosis and treatment using MLLMs are complex and multidisciplinary applications, future development requires enhanced collaboration among AI researchers, medical professionals, and policy regulators.

Updating medical knowledge and addressing emerging pandemics

Medical research produces vast amounts of new knowledge and insights annually, which can lead to the obsolescence of previous medical findings. LLMs can continuously update their knowledge base and learn the latest medical research, thereby providing cutting-edge medical advice.

The emergence and spread of new infectious diseases in recent years, such as COVID-19 and Ebola hemorrhagic fever, have severely threatened public health, causing global public health organizations to worry about future "Disease X" scenarios.[79] In this context, systems based on rapid updates to the existing knowledge bases and the deployment of new LLMs show promise as public health knowledge training tools or diagnostic aids in the future. Such systems may potentially provide "Disease X"-related medical information and advice to help people quickly address future pandemics.

LLMs have emerged as a transformative technology for the diagnosis and treatment of various diseases. We believe that under reasonable regulatory frameworks and multidisciplinary collaboration, LLMs can continue to evolve, offering more equitable, safe, high-quality, and accessible clinical diagnostic and treatment services.

Funding

This research was supported by grants from the National Key Research & Development Program of China (No. 2022YFC2505100) and the National Natural Science Foundation of China (No. 81970557).

References

1. Varghese J, Chapiro J. ChatGPT: The transformative influence of generative AI on science and healthcare. J Hepatol 2024;80:977–980. doi: 10.1016/j.jhep.2023.07.028.
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, *et al*. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. doi: 10.1371/journal.pdig.0000198.
3. Zhang N, Sun Z, Xie Y, Wu H, Li C. The latest version ChatGPT powered by GPT-4o: What will it bring to the medical field? Int J Surg 2024;110:6018–6019. doi: 10.1097/JS9.0000000000001754.
4. Raiaan MAK, Mukta MSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, *et al*. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access 2024;12:26839–26874. doi: 10.1109/ACCESS.2024.3365742.
5. Wei C, Wang YC, Wang B, Kuo CCJ. An overview of language models: Recent developments and outlook. APSIPA Trans Signal Inf Process 2024;13. doi: 10.1561/116.00000010.
6. Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, Chow AK, *et al*. A multimodal generative AI copilot for human pathology. Nature 2024;634:466–473. doi: 10.1038/s41586-024-07618-3.
7. Giuffrè M, Kresevic S, Pugliese N, You K, Shung DL. Optimizing large language models in digestive disease: Strategies and challenges to improve clinical outcomes. Liver Int 2024;44:2114–2124. doi: 10.1111/liv.15974.
8. Zhang D, Yu Y, Li C, Dong J, Su D, Chu C, *et al*. MM-LLMs: Recent advances in multimodal large language models. arXiv Preprint 2024. arXiv: 2401.13601. doi: 10.48550/arXiv.2401.13601.
9. Dehdab R, Brendlin A, Werner S, Almansour H, Gassenmaier S, Brendel JM, *et al*. Evaluating ChatGPT-4V in chest CT diagnostics: A critical image interpretation assessment. Jpn J Radiol 2024;42:1168–1177. doi: 10.1007/s11604-024-01606-3.
10. Liang L, Chen Y, Wang T, Jiang D, Jin J, Pang Y, *et al*. Genetic transformer: An innovative large language model driven approach for rapid and accurate identification of causative variants in rare genetic diseases. medRxiv 2024. doi: 10.1101/2024.07.18.24310666.
11. Liu X, Luo X, Jiang C, Zhao H. Difficulties and challenges in the development of precision medicine. Clin Genet 2019;95:569–574. doi: 10.1111/cge.13511.
12. Kasztura M, Richard A, Bempong NE, Loncar D, Flahault A. Cost-effectiveness of precision medicine: A scoping review. Int J Public Health 2019;64:1261–1271. doi: 10.1007/s00038-019-01298-x.
13. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, *et al*. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. Lancet Digit Health 2024;6:e12–e22. doi: 10.1016/S2589-7500(23)00225-X.

14. Goh E, Gallo R, Strong E, Weng Y, Kerman H, Freed J, *et al*. Large language model influence on management reasoning: A randomized controlled trial. medRxiv 2024. doi: 10.1101/2024.08.05.24311485.
15. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, *et al*. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023;6:e2343689. doi: 10.1001/jamanetworkopen.2023.43689.
16. Li K, Ruan G, Liu S, Xu T, Guan K, Li J, *et al*. Eosinophilic gastroenteritis: Pathogenesis, diagnosis, and treatment. Chin Med J 2023;136:899–909. doi: 10.1097/CM9.0000000000002511.
17. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: Are we there yet? Diagnostics (Basel) 2023;13:1950. doi: 10.3390/diagnostics13111950.
18. Henson JB, Glissen Brown JR, Lee JP, Patel A, Leiman DA. Evaluation of the potential utility of an artificial intelligence chatbot in gastroesophageal reflux disease management. Am J Gastroenterol 2023;118:2276–2279. doi: 10.14309/ajg.0000000000002397.
19. Cao JJ, Kwon DH, Ghaziani TT, Kwo P, Tse G, Kesselman A, *et al*. Large language models' responses to liver cancer surveillance, diagnosis, and management questions: Accuracy, reliability, readability. Abdom Radiol (NY) 2024;49:4286–4294. doi: 10.1007/s00261-024-04501-7.
20. Giuffrè M, Kresevic S, You K, Dupont J, Huebner J, Grimshaw AA, *et al*. Systematic review: The use of large language models as medical chatbots in digestive diseases. Aliment Pharmacol Ther 2024;60:144–166. doi: 10.1111/apt.18058.
21. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol 2024;34:e13207. doi: 10.1111/bpa.13207.
22. Wang C, Liu S, Li A, Liu J. Text dialogue analysis for primary screening of mild cognitive impairment: Development and validation study. J Med Internet Res 2023;25:e51501. doi: 10.2196/51501.
23. Berg HT, van Bakel B, van de Wouw L, Jie KE, Schipper A, Jansen H, *et al*. ChatGPT and generating a differential diagnosis early in an emergency department presentation. Ann Emerg Med 2024;83:83–86. doi: 10.1016/j.annemergmed.2023.08.003.
24. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Kornblith AE, *et al*. Use of a large language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open 2024;7:e248895. doi: 10.1001/jamanetworkopen.2024.8895.
25. Heston TF, Lewis LM. ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. PLoS One 2024;19:e0301854. doi: 10.1371/journal.pone.0301854.
26. Perret J, Schmid A. Application of OpenAI GPT-4 for the retrospective detection of catheter-associated urinary tract infections in a fictitious and curated patient data set. Infect Control Hosp Epidemiol 2024;45:96–99. doi: 10.1017/ice.2023.189.
27. Maillard A, Micheli G, Lefevre L, Guyonnet C, Poyart C, Canouï E, *et al*. Can chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. Clin Infect Dis 2024;78:825–832. doi: 10.1093/cid/ciad632.
28. Boligarla S, Laison EKE, Li J, Mahadevan R, Ng A, Lin Y, *et al*. Leveraging machine learning approaches for predicting potential Lyme disease cases and incidence rates in the United States using Twitter. BMC Med Inform Decis Mak 2023;23:217. doi: 10.1186/s12911-023-02315-z.
29. Chen C, Li L, Beetz M, Banerjee A, Gupta R, Grau V. Large language model-informed ECG dual attention network for heart failure risk prediction. arXiv Preprint 2024. arXiv: 2403.10581. doi: 10.48550/arXiv.2403.10581.
30. Yu H, Guo P, Sano A. Zero-shot ECG diagnosis with large language models and retrieval-augmented generation. Mach Learn Health 2023;225:650–663.
31. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol 2024;142:371–375. doi: 10.1001/jamaophthalmol.2023.6917.
32. Topol EJ. As artificial intelligence goes multimodal, medical applications multiply. Science 2023;381:adk6139. doi: 10.1126/science.adk6139.
33. AlRyalat SA, Musleh AM, Kahook MY. Evaluating the strengths and limitations of multimodal ChatGPT-4 in detecting glaucoma using fundus images. Front Ophthalmol (Lausanne) 2024;4:1387190. doi: 10.3389/fopht.2024.1387190.
34. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Evaluating chatgpt-4's diagnostic accuracy: Impact of visual data integration. JMIR Med Inform 2024;12:e55627. doi: 10.2196/55627.
35. Horiuchi D, Tatekawa H, Oura T, Oue S, Walston SL, Takita H, *et al*. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4v-based ChatGPT, and radiologists in challenging neuroradiology cases. Clin Neuroradiol 2024;34:779–787. doi: 10.1007/s00062-024-01426-y.
36. Ziegelmayer S, Marka AW, Lenhart N, Nehls N, Reischl S, Harder F, *et al*. Evaluation of GPT-4's chest X-Ray impression generation: A reader study on performance and perception. J Med Internet Res 2023;25:e50865. doi: 10.2196/50865.
37. Li CY, Chang KJ, Yang CF, Wu HY, Chen W, Bansal H, *et al*. Towards a holistic framework for multimodal large language models in three-dimensional brain CT report generation. arXiv Preprint 2024. arXiv: 240702235.
38. Bani-Harouni D, Navab N, Keicher M, eds. MAGDA: Multi-agent guideline-driven diagnostic assistance. In: International workshop on foundation models for general medical AI. Springer: 2024. doi: 10.1007/978-3-031-73471-7_17.
39. Sultan LR, Mohamed MK, Andronikou S. ChatGPT-4: A breakthrough in ultrasound image analysis. Oxford University Press; 2024: umae006. doi: 10.1093/radadv/umae006.
40. Guo X, Chai W, Li SY, Wang G, eds. LLaVA-ultra: Large Chinese language and vision assistant for ultrasound. ACM Multimedia 2024. doi: 10.48550/arXiv.2410.15074
41. Wu SH, Tong WJ, Li MD, Hu HT, Lu XZ, Huang ZR, *et al*. Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. Radiology 2024;310:e232255. doi: 10.1148/radiol.232255.
42. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, *et al*. A whole-slide foundation model for digital pathology from real-world data. Nature 2024;630:181–188. doi: 10.1038/s41586-024-07441-w.
43. Zheng S, Cui X, Sun Y, Li J, Li H, Zhang Y, *et al*. Benchmarking pathCLIP for pathology image analysis. J Imaging Inform Med 2024. doi: 10.1007/s10278-024-01128-4.
44. Turan EI, Baydemir AE, Özcan FG, Şahin AS. Evaluating the accuracy of ChatGPT-4 in predicting ASA scores: A prospective multicentric study ChatGPT-4 in ASA score prediction. J Clin Anesth 2024;96:111475. doi: 10.1016/j.jclinane.2024.111475.
45. Staubli SM, Walker HL, Saner F, Salinas CH, Broering DC, Malagò M, *et al*. Decoding the Clavien-Dindo classification: Artificial intelligence (AI) as a novel tool to grade postoperative complications. Ann Surg 2024. doi: 10.1097/SLA.0000000000006399.
46. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. Eur J Hum Genet 2024;32:466–468. doi: 10.1038/s41431-023-01396-8.
47. Alkuraya IF. Is artificial intelligence getting too much credit in medical genetics? Am J Med Genet C Semin Med Genet 2023;193:e32062. doi: 10.1002/ajmg.c.32062.
48. Barile J, Margolis A, Cason G, Kim R, Kalash S, Tchaconas A, *et al*. Diagnostic accuracy of a large language model in pediatric case studies. JAMA Pediatr 2024;178:313–315. doi: 10.1001/jamapediatrics.2023.5750.
49. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, *et al*. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: Exploratory study. J Med Internet Res 2023;25:e47621. doi: 10.2196/47621.
50. Huang J, Yang DM, Rong R, Nezafati K, Treager C, Chi Z, *et al*. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med 2024;7:106. doi: 10.1038/s41746-024-01079-8.
51. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology 2023;309:e232561. doi: 10.1148/radiol.232561.
52. Barr T, Ma S, Li Z, Yu J. Recent advances and remaining challenges in lung cancer therapy. Chin Med J 2024;137:533–546. doi: 10.1097/CM9.0000000000002991.
53. Griewing S, Gremke N, Wagner U, Lingenfelder M, Kuhn S, Boekhoff J. Challenging ChatGPT 3.5 in senology-an assessment of

concordance with breast cancer tumor board decision making. J Pers Med 2023;13:1502. doi: 10.3390/jpm13101502.
54. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, *et al*. Assessing the role of advanced artificial intelligence as a tool in multidisciplinary tumor board decision-making for primary head and neck cancer cases. Front Oncol 2024;14:1353031. doi: 10.3389/fonc.2024.1353031.
55. Krückel A, Brückner L, Psilopatis I, Fasching PA, Beckmann MW, Emons J. Evaluation of ChatGPT's potential in tailoring gynecological cancer therapies. In Vivo 2024;38:1649–1659. doi: 10.21873/invivo.13614.
56. Yalamanchili A, Sengupta B, Song J, Lim S, Thomas TO, Mittal BB, *et al*. Quality of large language model responses to radiation oncology patient care questions. JAMA Netw Open 2024;7:e244630. doi: 10.1001/jamanetworkopen.2024.4630.
57. Lawson McLean A, Wu Y, Lawson McLean AC, Hristidis V. Large language models as decision aids in neuro-oncology: A review of shared decision-making applications. J Cancer Res Clin Oncol 2024;150:139. doi: 10.1007/s00432-024-05673-x.
58. Kerbage A, Kassab J, El Dahdah J, Burke CA, Achkar JP, Rouphael C. Accuracy of ChatGPT in common gastrointestinal diseases: Impact for patients and providers. Clin Gastroenterol Hepatol 2024;22:1323–1325e3. doi: 10.1016/j.cgh.2023.11.008.
59. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, *et al*. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology 2024;80:1158–1168. doi: 10.1097/HEP.0000000000000834.
60. Rao A, Kim J, Lie W, Pang M, Fuh L, Dreyer KJ, *et al*. Proactive polypharmacy management using large language models: Opportunities to enhance geriatric care. J Med Syst 2024;48:41. doi: 10.1007/s10916-024-02058-y.
61. Imtiaz A, King J, Holmes S, Gupta A, Bafadhel M, Melcher ML, *et al*. ChatGPT *versus* Bing: A clinician assessment of the accuracy of AI platforms when responding to COPD questions. Eur Respir J 2024;63:400163. doi: 10.1183/13993003.00163-2024.
62. Li J, Guan Z, Wang J, Cheung CY, Zheng Y, Lim LL, *et al*. Integrated image-based deep learning and language models for primary diabetes care. Nat Med 2024;30:2886–2896. doi: 10.1038/s41591-024-03139-8.
63. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: The end of the consulting infection doctor? Lancet Infect Dis 2023;23:405–406. doi: 10.1016/S1473-3099(23)00113-5.
64. Vuurberg G, Hoorntje A, Wink LM, van der Doelen BFW, van den Bekerom MP, Dekker R, *et al*. Diagnosis, treatment and prevention of ankle sprains: Update of an evidence-based clinical guideline. Br J Sports Med 2018;52:956. doi: 10.1136/bjsports-2017-098106.
65. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, *et al*. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep 2023;13:20159. doi: 10.1038/s41598-023-47500-2.
66. Chen X, Wang L, You M, Liu W, Fu Y, Xu J, *et al*. Evaluating and enhancing large language models' performance in domain-specific medicine: Development and usability study with DocOA. J Med Internet Res 2024;26:e58158. doi: 10.2196/58158.
67. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, *et al*. Benchmarking large language models' performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine 2023;95:104770. doi: 10.1016/j.ebiom.2023.104770.
68. Kern FB, Wu CT, Chao ZC. Assessing novelty, feasibility and value of creative ideas with an unsupervised approach using GPT-4. Br J Psychol 2024. doi: 10.1111/bjop.12720.
69. Li D, Gupta K, Bhaduri M, Sathiadoss P, Bhatnagar S, Chong J. Comparing GPT-3.5 and GPT-4 accuracy and drift in radiology diagnosis please cases. Radiology 2024;310:e232411. doi: 10.1148/radiol.232411.
70. Deng J, Heybati K, Shammas-Toma M. When vision meets reality: Exploring the clinical applicability of GPT-4 with vision. Clin Imaging 2024;108:110101. doi: 10.1016/j.clinimag.2024.110101.
71. Liu X, Wu Z, Wu X, Lu P, Chang KW, Feng Y. Are LLMs capable of data-based statistical and causal reasoning? Benchmarking advanced quantitative reasoning with data. arXiv Preprint 2024. arXiv: 2402.17644. doi: 10.48550/arXiv.2402.17644.
72. Chen Y, Wang Z, Xing X, Xu Z, Fang K, Wang J, *et al*. Bianque: Balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT. arXiv Preprint 2023. arXiv: 2310.15896. doi: 10.48550/arXiv.2310.15896.
73. Shea YF, Lee CMY, Ip WCT, Luk DWA, Wong SSW. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. JAMA Netw Open 2023;6:e2325000. doi: 10.1001/jamanetworkopen.2023.25000.
74. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, *et al*. Evaluating large language models on medical evidence summarization. NPJ Digit Med 2023;6:158. doi: 10.1038/s41746-023-00896-7.
75. Wang LL, Otmakhova Y, DeYoung J, Truong TH, Kuehl BE, Bransom E, *et al*. Automated metrics for medical multi-document summarization disagree with human evaluations. arXiv Preprint 2023. arXiv: 2305.13693. doi: 10.48550/arXiv.2305.13693.
76. Pelrine K, Taufeeque M, Zając M, McLean E, Gleave A. Exploiting novel gpt-4 apis. arXiv Preprint 2023. arXiv: 2312.14302. doi: 10.48550/arXiv.2312.14302.
77. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. Radiology 2023;309:e231147. doi: 10.1148/radiol.231147.
78. Shi Y, Shu P, Liu Z, Wu Z, Li Q, Li X. Mgh radiology llama: A llama 3 70b model for radiology. arXiv Preprint 2024. arXiv: 2408.11848. doi: 10.48550/arXiv.2408.11848.
79. Simpson S, Kaufmann MC, Glozman V, Chakrabarti A. Disease X: Accelerating the development of medical countermeasures for the next pandemic. Lancet Infect Dis 2020;20:e108–e115. doi: 10.1016/S1473-3099(20)30123-7.

How to cite this article: Yang XT, Li TX, Su Q, Liu YL, Kang CX, Lyu Y, Zhao LN, Nie YZ, Pan YL. Application of large language models in disease diagnosis and treatment. Chin Med J 2025;138:130–142. doi: 10.1097/CM9.0000000000003456