Intelligent Medicine 5 (2025) 151–163

Contents lists available at ScienceDirect

Intelligent Medicine

journal homepage: www.elsevier.com/locate/imed

Review

Evaluating large language models and agents in healthcare: key challenges in clinical applications

Xiaolan Chen [1,#], Jiayang Xiang [2,#], Shanfu Lu [3,#], Yexin Liu [4], Mingguang He [1,5,6,*], Danli Shi [1,2,*]

[1] *School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China*
[2] *Department of Ophthalmology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China*
[3] *Perception Vision Medical Technologies Co. Ltd., Guangzhou, Guangdong 510530, China*
[4] *AI Thrust, The Hong Kong University of Science and Technology, Guangzhou, Guangdong 511453, China*
[5] *Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Kowloon, Hong Kong, China*
[6] *Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Hong Kong, China*

ARTICLE INFO

Keywords:
Large language model
Generative pre-trained transformer
Evaluation
Reasoning
Hallucination
Medical agent

ABSTRACT

Large language models (LLMs) have emerged as transformative tools with significant potential across healthcare and medicine. In clinical settings, they hold promises for tasks ranging from clinical decision support to patient education. Advances in LLM agents further broaden their utility by enabling multimodal processing and multitask handling in complex clinical workflows. However, evaluating the performance of LLMs in medical contexts presents unique challenges due to the high-risk nature of healthcare and the complexity of medical data. This paper provides a comprehensive overview of current evaluation practices for LLMs and LLM agents in medicine. We contributed 3 main aspects: First, we summarized data sources used in evaluations, including existing medical resources and manually designed clinical questions, offering a basis for LLM evaluation in medical settings. Second, we analyzed key medical task scenarios: closed-ended tasks, open-ended tasks, image processing tasks, and real-world multitask scenarios involving LLM agents, thereby offering guidance for further research across different medical applications. Third, we compared evaluation methods and dimensions, covering both automated metrics and human expert assessments, while addressing traditional accuracy measures alongside agent-specific dimensions, such as tool usage and reasoning capabilities. Finally, we identified key challenges and opportunities in this evolving field, emphasizing the need for continued research and interdisciplinary collaboration between healthcare professionals and computer scientists to ensure safe, ethical, and effective deployment of LLMs in clinical practice.

1. Introduction

In recent years, the emergence of large language models (LLMs) has catalyzed transformative advancements across diverse domains, spanning from natural language understanding to content generation. These models, with expanding capabilities, are increasingly finding integration into a wide spectrum of applications [1–3]. Notably, researchers have begun exploring their potential in the medical field [4,5] from aiding clinical decision-making to enhancing patient education and engagement [6].

Due to the limitations of general LLMs in medical applications, particularly in tasks like interpreting medical images or grasping clinical context [7,8] some studies have developed LLMs tailored specifically for medical applications, significantly improving their ability to address various tasks [4,9–12]. Additionally, in response to the multimodal, multitask demands of real-world medical needs, recent studies have developed artificial intelligence (AI) agent systems driven by LLMs, referred to as LLM agents. These systems use LLMs as the "brain" and integrate various expert AI models as tools, enabling them to autonomously understand user instructions, make decisions, and select appropriate tools to perform complex medical tasks. However, traditional AI evaluation methods based on single tasks and single dimensions are no longer sufficient to meet the new demands posed by the rapid development and increasing generality of medical LLMs and LLM agents. First, data bias remains a significant challenge. Some datasets are drawn from specific domains or populations, potentially misrepresenting real-world performance [13]. Second, evaluations across diverse healthcare applications are often broad but lack depth [14], failing to differenti-

[*] **Corresponding authors:** Mingguang He, Experimental Ophthalmology, School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China; Danli Shi, School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China.

E-mail addresses: mingguang.he@polyu.edu.hk (Mingguang He), danli.shi@polyu.edu.hk (Danli Shi).

[#] These authors contributed equally to this work.

https://doi.org/10.1016/j.imed.2025.03.002
Received 6 January 2025; Received in revised form 28 February 2025; Accepted 2 March 2025

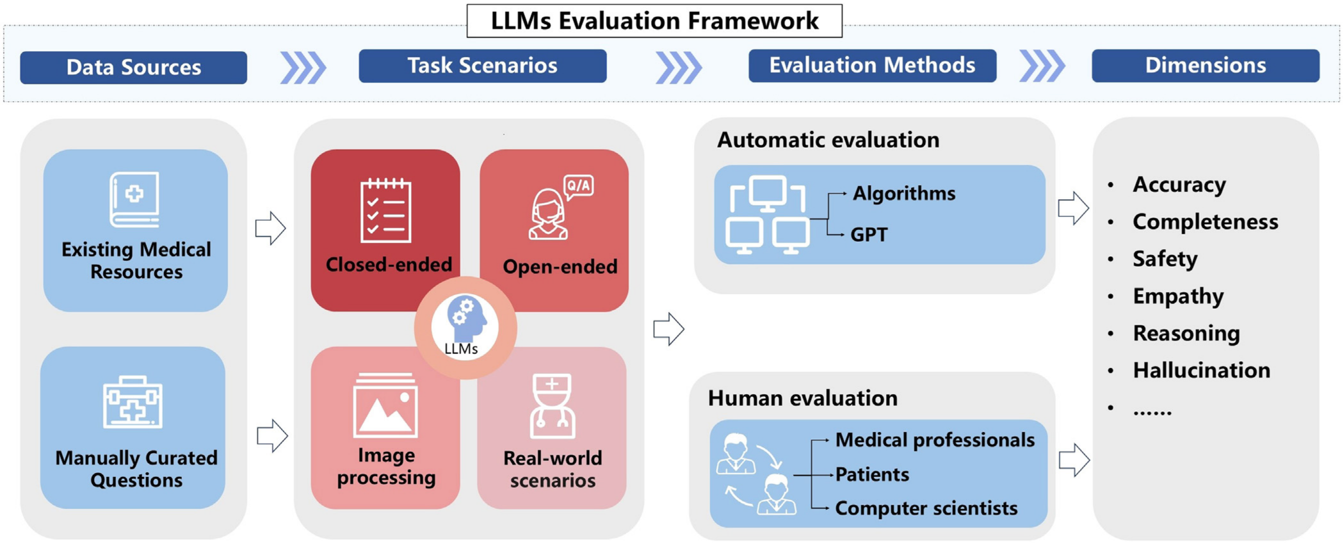


Figure 1. Illustration of the potential LLM evaluation framework in medicine. LLM: large language model.

ate the strengths and weaknesses of LLMs in practical scenarios. Third, prior evaluation methods primarily focus on accuracy [15], overlooking other critical attributes such as hallucination assessment, logical reasoning, and the likelihood of generating harmful content. As LLMs and LLM agents evolve, it is essential to establish standardized evaluation criteria and benchmarks [16].

This paper aims to provide a comprehensive overview of the landscape of medical LLM and LLM agent evaluation (Figure 1), synthesizing insights from existing studies and addressing key challenges and opportunities. Specifically, we first provide a broad perspective on their capabilities and challenges through exploration of data sources and task scenarios. Second, we synthesize diverse evaluation methodologies employed in medical LLM evaluation, from objective accuracy metrics to more nuanced human-centric evaluations, as well as traditional and LLM agent-specific evaluation dimensions. Third, key challenges and opportunities in medical LLM and LLM agent evaluation were identified to underscore the need for continued research and innovation in the field. By providing a coherent understanding of the latest techniques in medical evaluation of LLMs and LLM agents, we hope to offer conceptual advancements to facilitate the responsible integration of LLMs into medical practice.

2. Methods

This research followed the systematic review guidelines set forth by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses. A systematic search was carried out across PubMed, Google Scholar, and Web of Science databases for peer-reviewed journal articles and conference proceedings published between 1 January 2023 and 13 November 2024. Specific keywords were employed, including “Large Language Model,” “ChatGPT,” “AI Agent,” “LLM Agent,” “Medical,” “Medicine,” “Evaluation” and “Assess.” Studies were included if they had applied LLMs in the medical field and performed an adequate assessment of their performance. Studies were excluded if they were not relevant to the medical application or demonstrated methodological limitations, particularly those lacking formal evaluation protocols, statistical validation, or sample sizes smaller than 20. Representative studies from different task scenarios as well as evaluation methods were selected and cited as examples. To minimize recency bias and capture emerging work, an updated search was conducted on 25 February 2025, identifying 5 additional eligible peer-reviewed articles and 4 relevant preprints meeting inclusion criteria. After filtering, a total of 256 studies were included in the literature review (Supplementary Figure 1).

3. Data sources

Before applying LLMs in medical applications, a thorough evaluation of the LLMs is crucial. The complexity and diversity of medical data pose a significant challenge when constructing appropriate test sets. Currently, such datasets can be broadly divided into 2 main categories: existing medical resources and manually curated question sets.

3.1. Existing medical resources

Medical examinations, designed to assess the competency of healthcare professionals, offer readily available benchmark datasets. These exams, crafted through generations of expertise and accompanied by standardized answers, provide a substantial volume of validated material for evaluating LLMs. Diverse medical exams from different countries have been used to assess LLMs’ general medical knowledge capabilities, including the United States Medical Licensing Examination (USMLE) [13,17], the National Medical Licensing Examination in China [18], the National Pharmacist Licensing Examination in China, National Nurse Licensing Examination in China [19], Chinese Master’s Degree Entrance Examination [20] and more. For more focused assessment in specific subspecialties in medicine, exams such as the Ophthalmic Knowledge Assessment Program examination [21], the Basic Science and Clinical Science Self-Assessment Program [22], the oral and written board examinations for the American Board of Neurological Surgery (ABNS) [23], Otolaryngology-Head and Neck Surgery Certification Examinations [24], the Royal College of General Practitioners Applied Knowledge Test [25], European Board of Radiology exam [26], and others are utilized. These datasets provide a wealth of data aligned with specialty knowledge and practices for the assessment of the depth and breadth of knowledge possessed by LLMs.

In addition to examinations, medical literature serves as an important knowledge repository, including peer-reviewed journal articles and conference papers [27–29]. These databases offer cutting-edge medical insights and research findings, contributing to assessing whether the LLM is adept at rapidly updating medical knowledge.

3.2. Manually curated questions

While the range of exam questions and academic materials is extensive, their ability to reflect the dynamic capabilities required for real-world interactions remains limited. Therefore, some studies have turned

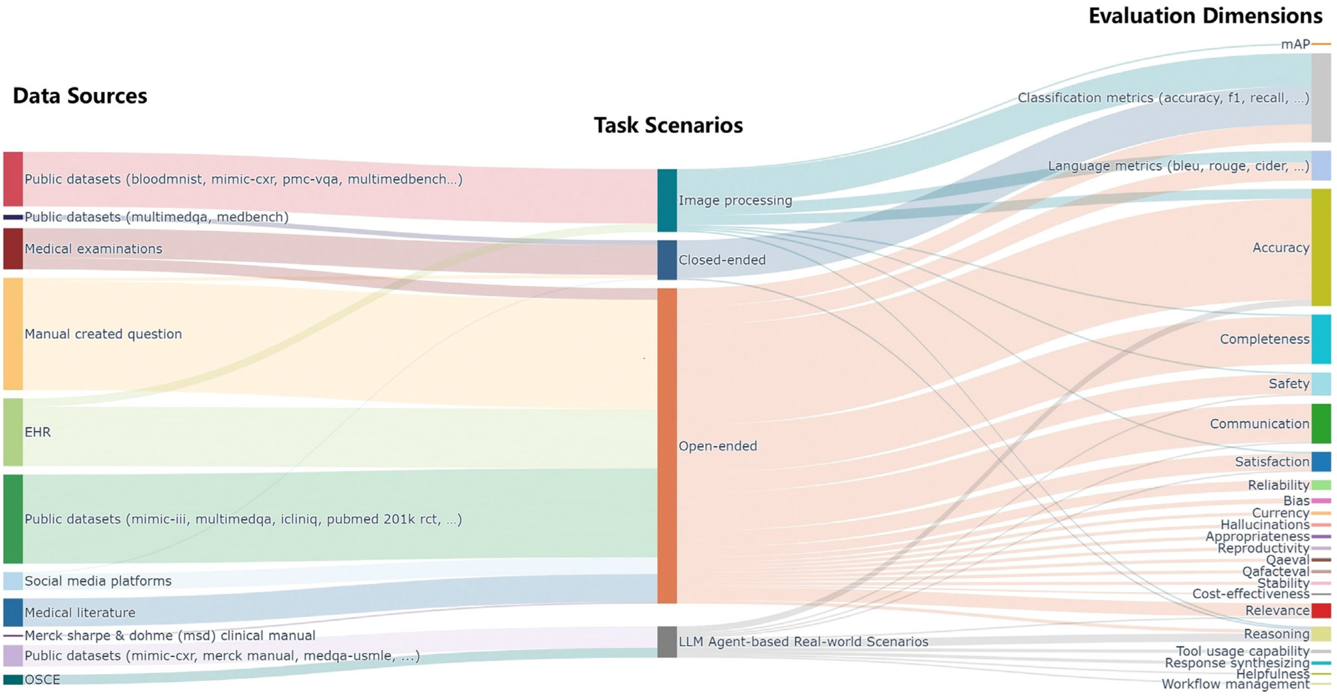


Figure 2. Summary of data sources and evaluation dimensions related to different task scenarios in LLM medical evaluation. LLM: large language model; EHR: electronic health record; OSCE: objective structured clinical examination.

to the use of real-world data, manually created or collected by healthcare professionals for evaluation. For instance, Singhal et al [4] proposed MultiMedQA as the evaluation dataset for evaluating their developed models. Other studies leveraged real-world interactions and discussions from medical forums and social media to evaluate the conversational and consultative skills of LLM [30–32]. Medical images, such as X-rays, magnetic resonance imaging (MRI), computed tomography (CT) in radiology [33], fundus photographs, fundus fluorescein angiography (FFA), and optical coherence tomography (OCT) images in ophthalmology [34], are essential for tasks like disease diagnosis and image analysis. These clinically derived image data often come with expert medical reports and are crucial resources for constructing multimodal datasets, thus facilitating the testing of LLMs' ability to handle complex visual and textual information.

However, collecting appropriate data can be challenging due to scarce resources in certain disciplines, as well as ethical considerations and data privacy issues. In response, some studies have turned to expert-crafted questions carefully formulated based on clinical expertise [35–37]. For example, Marshall et al [38] have constructed datasets centered around symptoms, examinations, and treatments of uveitis to evaluate ChatGPT's proficiency in handling specialized content. Another study involved collaboration with a panel of 8 board-certified clinicians and 2 healthcare practitioners, who generated a dataset of 314 clinical questions spanning 9 medical specialties, to assess the performance of LLMs [39]. Although these approaches may result in datasets with a limited number of questions, they provide highly specialized and practical insights into LLMs, reflecting the nuanced understanding required in real-world clinical practice. Additionally, the manual creation of test questions guarantees their exclusion from the training data, ensuring its integrity and preventing contamination.

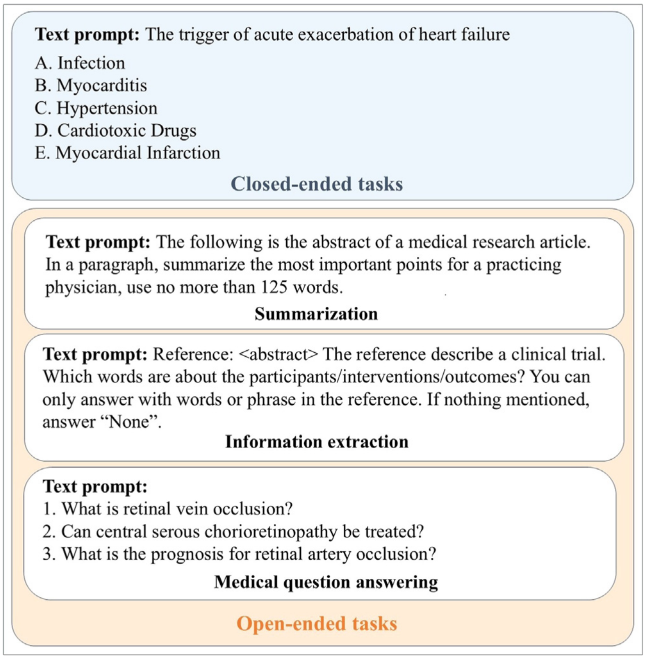
4. Task scenarios

With the rapid development of LLMs, they have demonstrated extensive application prospects in the medical field. The evaluation protocol of the models varies depending on the specific nature of each task. This section aims to review the recent advancements of evaluating LLMs in different medical scenarios, particularly focusing on tasks such as medical closed-ended tasks, open-ended medical question answering (QA), image processing, and real-world multitask scenarios based on LLM agents. Figure 2 demonstrates the main data sources and evaluation dimensions related to different task scenarios in the medical evaluation of LLMs and LLM agents. Supplementary Table 1 summarized the major studies on the evaluation of LLMs and LLM agents in different scenarios.

4.1. Closed-ended tasks

Closed-ended medical questions are commonly used in medical education. Compared to open-ended questions, closed-ended questions typically have definite answers, reducing the possibility of subjective judgment (Figure 3). This enables easy quantification of the performance of LLMs on multiple-choice questions (MCQ), making them suitable for large-scale evaluation and comparison between different models. For instance, Royer et al [16] developed an open-source evaluation toolkit called MultiMedEval to comprehensively assess the performance of LLMs on medical MCQ tasks. They conducted experiments on 3 datasets covering a wide range of medical domains, including general medical knowledge, radiological image interpretation, and yes/no questions from literature. They compared the performance of closed-source and open-source models based on the bilingual evaluation understudy (BLEU) scores. Liu et al [40] evaluated models such as GPT-4 on the CMedExam dataset and found that its 61.6% accuracy was still significantly lower than human level (71.6%). The limitations mentioned above can be primarily attributed to the insufficient coverage of medical domain knowledge by LLMs, their limited understanding of professional medical terminology, and the inadequacies of current evaluation metrics. Li et al [20] assessed ChatGPT's reliability and practicality in medical education by testing its performance on the 2021–2023 Chinese Master of Clinical Medicine comprehensive exams. The results found that both ChatGPT-3.5 and ChatGPT-4 passed the admission threshold but showed biases, with high accuracy in humanities subjects (93.75%) and lower accuracy in pathology (37.5% for ChatGPT-3.5).

Another reason for using MCQs in evaluating LLMs is the broad knowledge coverage of existing datasets. These datasets span a wide



Text prompt: The trigger of acute exacerbation of heart failure
A. Infection
B. Myocarditis
C. Hypertension
D. Cardiotoxic Drugs
E. Myocardial Infarction

Text prompt: The following is the abstract of a medical research article. In a paragraph, summarize the most important points for a practicing physician, use no more than 125 words.

Text prompt: Reference: <abstract> The reference describe a clinical trial. Which words are about the participants/interventions/outcomes? You can only answer with words or phrase in the reference. If nothing mentioned, answer "None".

Text prompt:
1. What is retinal vein occlusion?
2. Can central serous chorioretinopathy be treated?
3. What is the prognosis for retinal artery occlusion?

Figure 3. Examples of closed-ended tasks and various open-ended tasks.

range of topics and difficulty levels, making them valuable resources for constructing domain-specific datasets. This is particularly advantageous when assessing the performance of language models in specialized fields, such as medicine, where comprehensive evaluation requires testing across diverse knowledge areas. Suchman et al [41] evaluated the performance of GPT-3 and GPT-4 on the American College of Gastroenterology self-assessment tests. A score of 70% or higher is generally required to pass the evaluation. However, ChatGPT-3.5 and ChatGPT-4 scored 65.1% and 62.4%, respectively, both failing to meet the passing threshold. Therefore, ChatGPT is currently not suitable for gastroenterology medical education in its present form. Bhayana et al [42] evaluated ChatGPT's performance on radiology board-style exam questions without images. ChatGPT correctly answered 69% of the questions, essentially passing the exam. Gupta et al [43] evaluated ChatGPT's performance on the Plastic Surgery Inservice Training Examination to determine if it could serve as an assistive tool for resident education. The model achieved an accuracy of 54.95% on this task. By testing ChatGPT's performance on various questions related to general knowledge, information clarification, case-based learning, and evidence-based medicine promotion, the results suggest that ChatGPT could potentially become an effective tool for plastic surgery resident education.

Despite closed-ended questions providing an objective and quantifiable means to evaluate LLMs' medical knowledge, there are limitations, including a focus on procedural knowledge, a lack of deep assessment of complex situations, and a failure to reflect the models' performance in real-world scenarios. Li et al [44] found that LLMs showed sensitivity to answer positions in bilingual MCQs, especially a tendency to choose answers located in the first position, revealing potential biases, suggesting that current MCQ benchmarks may not accurately reflect the true capabilities of LLMs. Therefore, a comprehensive evaluation framework should include both closed-ended and open-ended tasks to assess the full capabilities and limitations of medical LLMs.

4.2. *Open-ended tasks*

Open-ended tasks refer to generating answers diversely. Compared to the MCQ tasks that have only a few simple and mechanical answers like "yes or no," open-ended tasks require improved natural language processing (NLP) ability and feasibility of LLMs to fully express opinions (Figure 3). For this reason, the evaluation dimensions become more complicated in open-ended tasks accordingly. In medical fields, there are 3 main open-ended tasks: summarization, information extraction, and medical QA.

4.2.1. *Summarization*

Text summarization related to the medical field requires extracting key information from various medical data sources, such as medical literature and electronic health records (EHRs), and then generating a short, concise summary of the given medical text. LLMs can help summarize medical research evidence [45,46]. Tang et al [29] investigated the capabilities and limitations of LLMs in summarizing reviews across 6 clinical domains. Automatic metrics were used to assess the lexical and semantic similarity, and human evaluation was conducted on coherence, factual consistency, comprehensiveness, and harmfulness. The finding revealed that LLMs could be susceptible to generating incorrect information and lead to potential harm due to misinformation. Hake et al [47] evaluated ChatGPT's ability to summarize abstracts from more types of clinical research, such as case series, observational studies, randomized controlled trials (RCT), and more. They solely relied on human assessment, encompassing quality, accuracy, bias, and relevance, and showed LLMs' acceptable performance in these aspects.

As for summarizing EHRs, LLMs have the potential to be applied in generating medical text to ease the documentation burden for physicians [48–50]. Dubinski et al [51] investigated the time consumption and factual correctness of neurosurgical discharge summaries and operative reports ChatGPT generated. The result showed significant time reduction and a high degree of factual correctness with the assistance of ChatGPT. Zaretsky et al [52] further assessed the ability of LLMs to transform discharge summaries into patient-friendly language and format from readability, accuracy, and completeness. These findings suggested that despite not being perfect and containing a few inappropriate omissions or insertions, LLMs have the potential to enhance the efficiency of generating medical documents. In addition to generating discharge summaries and various other documents, EHRs play a crucial role in medical examination reports, which often involve complex terminology. LLMs can summarize the key information and convey the message in plain language. Lyu et al [53] and Chung et al [54] evaluated ChatGPT's performance in summarizing radiology reports, including CT and MRI. Lyu et al [53] assessed completeness and correctness while Chung et al [54] assessed readability, factual correctness, ease of understanding, completeness, potential for harm and overall quality. They demonstrated the novel feasibility of using LLMs to generate patient-friendly summaries of radiology reports. Van et al [55] evaluated LLMs' performance in clinical text summarization across multiple tasks, including radiology reports, patient questions, progress notes, and doctor-patient dialogues from similarity, completeness, correctness, and conciseness. They provided evidence of LLMs outperforming medical experts in clinical text summarization across multiple tasks.

4.2.2. *Information extraction*

Information extraction is basic but essential in medical fields, especially in the biomedical domain. Numerous studies showed LLMs have the potential to assist researchers or patients search and acquire knowledge from a large amount of biomedical data.

Named entity recognition (NER) is one of the information extraction tasks that involves identifying named entities such as genes, proteins, diseases, and more, in the given input text. Many studies evaluated LLMs' performances on this task in the first place [56,57]. Evaluation metrics such as precision, recall, and F1 score were widely used in the evaluation of this task. In the EHR field, Gu et al [58] trained an LLM to extract and quantify stroke severity from EHR based on Chinese clinical NER. This model demonstrated a high F1-score of 0.990, which ensured the reliability of the model in accurately extracting the entities for the subsequent automatic NIHSS scoring. In addition, Guevara et al [59] used LLMs to extract social determinants of health (SDoH) categories from real-world clinic notes. Their fine-tuned models achieved higher F1 scores than ChatGPT for most SDoH classes and highlighted the potential of LLMs to improve real-world data collection and identification of SDoH from the EHR.

Relation extraction refers to extracting relations between named entities in a given text, such as relations between genes and diseases, or genes and proteins. Researchers also evaluate the performance of LLMs using statistical measures such as accuracy, precision, recall, and F1 scores. Most researchers evaluate LLMs on relation extraction through datasets related to drug-drug interaction, chemical-disease relation, gene-disease associations [56] and drug-target interaction [57,60]. Cinquin et al [61] developed a fine-tuned generative pre-trained transformer model named ChIP-GPT, which was able to extract data from biomedical database records, especially to identify cell lines and the gene. ChIP-GPT demonstrates 90%–94% accuracy when trained with 100 examples.

4.2.3. Medical question answering

Patients often have numerous questions and concerns about their health. A number of studies evaluated LLMs' open-minded responses across various disciplines, encompassing questions such as the concept of disease, etiology, examination, diagnosis, prevention, treatment, and care. The majority of LLMs' evaluations concentrated on the reliability of their responses, which may impact their healthcare decision to a great extent.

The evaluation of LLMs in medical QA tasks involves diverse sources of questions, including author-designed questions [62–64], questions from professional societies and institutions [65,66], social media [67,68] and validated real or simulated clinical patient cases [69–71]. When it comes to specific applications in various clinical disciplines, clinicians paid more attention to the LLMs' application in their clinical sub-specialty. Take ophthalmology, for example, Ali [72] evaluated ChatGPT's performance on queries related to lacrimal drainage disorders from correctness, but the performance of ChatGPT in this context, at best, can be considered average. The study highlighted that there is a need for it to be specifically trained for individual medical subspecialties. Potapenko et al [73] consulted ChatGPT for common retinal diseases, including age-related macular degeneration, diabetic retinopathy, retinal vein occlusion, retinal artery occlusion, and central serous chorioretinopathy and evaluated the accuracy of responses. ChatGPT provided highly accurate responses to most questions except for questions dealing with treatment options. Rasmussen et al [74] evaluated the accuracy of responses to typical patient-related questions on vernal keratoconjunctivitis. The result also demonstrated that responses to treatment/prevention questions obtained lower scores than the rest.

The dimensions used to assess performance in medical QA tasks vary across the literature. However, relying solely on metrics like BLEU, consensus-based image description evaluation (CIDEr), recall-oriented understudy for gisting evaluation (ROUGE), and others for automatic evaluation of these tasks has its limitations. Interestingly, a few studies have employed GPT-4 to automatically evaluate responses of other models across multiple dimensions [75]. This introduced a novel and user-friendly automated assessment method. However, further exploration is needed to determine the reliability and consistency of GPT-4 compared to human assessments. Currently, human evaluation is necessary due to the inability to parse scripts to properly identify the correctness of responses with the latest updates in medical knowledge and analyze the implications of these responses for patients. In addition to the crucial evaluation metric of accuracy or correctness, which is prioritized in most studies, there are other valuable evaluation metrics such as completeness [31,64,65,67,68,76–80] and readability [64,76,77,81–83]. Some studies have also examined relatively uncommon dimensions, including safety and humanistic care. Cadamuro et al [84] evaluated whether the responses were safe, which referred to the potential negative consequences and detrimental effects of ChatGPT's response on the patient's health and well-being. Menz et al [85] explored the issue of the security of LLMs when LLMs were prompted to generate health disinformation and measured whether safeguards of LLMs prevented the generation of health disinformation. This study found that inconsistencies in the effectiveness of LLM safeguards to prevent the mass generation of health disinformation. Yeo et al [68] and Zhu et al [64] paid attention to LLMs' performance in emotional support (or humanistic care), which referred to offering psychological assistance to individuals facing particular emotional distress when diagnosed with cancer or other diseases. Their results showed that ChatGPT demonstrates empathy when answering patients' questions.

4.3. Image processing tasks

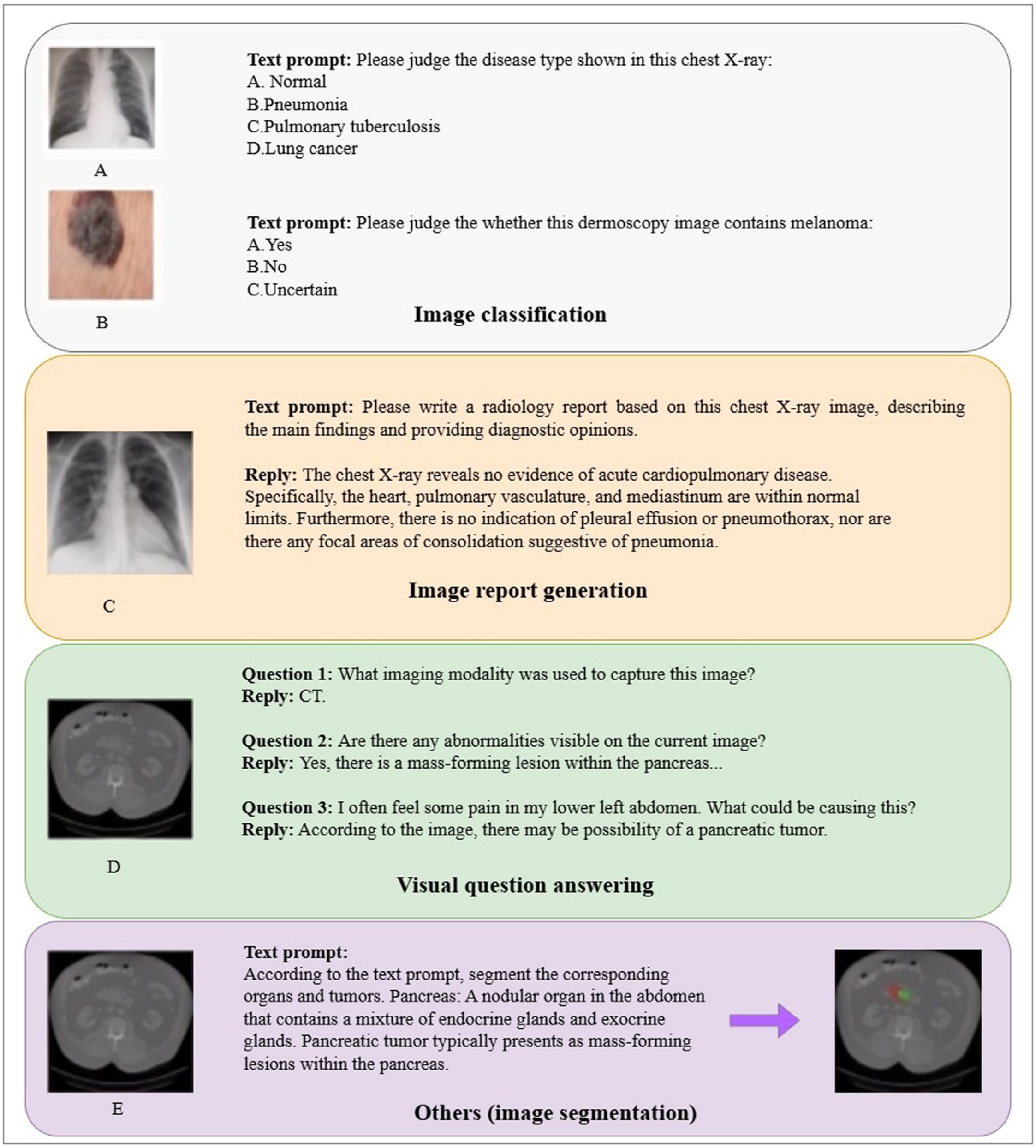
Medical diagnosis and treatment often rely on various types of images, including CT and MRI in radiology, as well as fundus photographs and OCT in ophthalmology. By jointly modeling image and text information, LLMs can not only better understand the content of medical images but also generate diagnostic reports based on images. This section will review the evaluation progress of LLMs in medical image processing tasks, including image classification, report generation, and visual question answering (VQA). Examples of various image processing tasks can be found in Figure 4.

4.3.1. Image classification

Medical image classification is a fundamental task for evaluating the ability of LLMs to understand medical image content and identify disease patterns. This task demands strong feature extraction capabilities from the model to distinguish subtle differences between images. Common performance metrics include accuracy, recall, precision, and F1-score. Royer et al [16] evaluated the multi-class/multi-label classification performance of LLMs using metrics such as F1, AUROC, and accuracy on 15 datasets, including MIMIC-CXR, Pad-UFES-20, and CBIS-DDSM. The results showed that LLMs performed well on different medical image tasks but still need improvement in recognizing fine-grained visual concepts. Van et al [86] compared LLMs like BiomedCLIP, OpenCLIP, OpenFlamingo, LLaVA, and ChatGPT-4 with traditional convolutional neural network (CNN) models on medical image classification tasks using brain MRI, blood smear microscopy images, and chest X-rays. Results indicated that while CNN models excelled due to specialized training, vision-language models (VLMs) showed promise in zero-shot and few-shot scenarios without prior training. Wan et al [87] proposed the Med-UniC framework, which employs cross-lingual text regularization techniques. This framework not only exhibits outstanding performance on multiple medical image classification datasets, such as CheXpert, RSNA, and COVIDx, but also achieves remarkable results in zero-shot classification tasks. In the cross-lingual evaluation on the CXP500 and PDC datasets, Med-UniC, utilizing both English and Spanish prompts, achieved an improvement of over 20% in F1-scores, demonstrating its effectiveness and adaptability when processing non-English community images and prompts. These results underscore the importance of reducing community bias to enhance the diagnostic quality and task performance of the model in clinical applications. Moreover, these findings further highlight the robust feature extraction and cross-lingual understanding capabilities of VLMs.

4.3.2. Report generation

Medical report generation aims to automate the process of generating diagnostic reports based on medical images. Given medical images, LLMs are required to generate diagnostic reports that conform to clinical norms, including major sections such as imaging findings, diagnostic opinions, and differential diagnoses. It requires not only accurately extracting key findings from images but also standardized descriptions using professional terminology and reasoning about possible diagnoses.



Text prompt: Please judge the disease type shown in this chest X-ray:
A. Normal
B.Pneumonia
C.Pulmonary tuberculosis
D.Lung cancer

Text prompt: Please judge the whether this dermoscopy image contains melanoma:
A.Yes
B.No
C.Uncertain

Image classification

Text prompt: Please write a radiology report based on this chest X-ray image, describing the main findings and providing diagnostic opinions.

Reply: The chest X-ray reveals no evidence of acute cardiopulmonary disease. Specifically, the heart, pulmonary vasculature, and mediastinum are within normal limits. Furthermore, there is no indication of pleural effusion or pneumothorax, nor are there any focal areas of consolidation suggestive of pneumonia.

Image report generation

Question 1: What imaging modality was used to capture this image?
Reply: CT.

Question 2: Are there any abnormalities visible on the current image?
Reply: Yes, there is a mass-forming lesion within the pancreas...

Question 3: I often feel some pain in my lower left abdomen. What could be causing this?
Reply: According to the image, there may be possibility of a pancreatic tumor.

Visual question answering

Text prompt:
According to the text prompt, segment the corresponding organs and tumors. Pancreas: A nodular organ in the abdomen that contains a mixture of endocrine glands and exocrine glands. Pancreatic tumor typically presents as mass-forming lesions within the pancreas.

Others (image segmentation)

Figure 4. Examples of various image processing tasks. A. Image classification task: Chest X-ray showing pneumonia infection. B. Image classification task: Dermoscopic images of melanoma. C. Image report generation task: Normal chest X-ray images. D, E. Visual question answering & Image segmentation tasks: Abdominal CT scans showing internal organs.

Currently, the evaluation on this task is mainly carried out on large-scale radiology datasets such as IU X-Ray, MIMIC-CXR, and CX-CHR. A series of language metrics such as BLEU, ROUGE, metric for evaluation of translation with explicit ordering (METEOR), and CIDEr are used to comprehensively examine the fluency, relevance, and readability of the generated reports. Liu et al [88] evaluated the radiology report generation performance of GPT-4V. The results showed that the model could generate descriptive reports with structured prompts, but there is still space for improvement in the accuracy of generating specific medical terms. Royer et al [16] systematically tested and demonstrated the ability of LLMs to generate diagnostic reports based on chest X-rays using the MIMIC-CXR dataset. Wang et al [89] proposed the R2GenGPT model and validated its performance in radiology report generation. Lee et al [90] constructed the LLM-CXR model and demonstrated its ability to accurately capture lesion characteristics, locations, and severity, closely aligning with the input text. Both methods are promising solutions for automating and improving radiology reports, but no human evaluation has been conducted. At the same time, some researchers have conducted comprehensive evaluations of human-computer interaction. Chen et al [91] introduced the ICGA-GPT model and assessed its ability to generate ophthalmic reports through automated and manual evaluations. The results showed that ICGA-GPT achieved satisfactory scores in BLEU1-4, CIDEr, ROUGE, and semantic propositional image caption evaluation metrics. Furthermore, subjective assessments conducted by ophthalmologists indicated high accuracy and completeness scores for the model.

4.3.3. Visual question answering

Medical VQA is a key task for evaluating the multimodal integration and reasoning capabilities of LLMs. In this task, given a medical image and related questions, LLMs need to comprehend the contextual information of the questions and images and then generate answers that conform to clinical facts and logic. In contrast to image classification and report generation, VQA requires stronger cross-modal understanding capabilities and higher demands on natural language expression. Additionally, the open-ended nature of VQA questions demands that the model possesses higher levels of creativity and flexibility, encompassing both image comprehension and extensive medical knowledge beyond the image. This puts forward high requirements for the model in terms of visual understanding, language analysis, knowledge reasoning, and other aspects, and is an important standard for examining its understanding ability in medical scenarios.

Due to the difficulty of medical VQA tasks, constructing large-scale, high-quality datasets is crucial for evaluating model performance. Lin et al [92] systematically reviewed representative datasets in this field. such as VQA-Med 2018, VQA-RAD, VQA-Med 2019, RadVisDial, PathVQA, VQA-Med 2020, SLAKE, and VQA-Med 2021. These datasets cover various medical domains, including radiology and pathology, providing a foundation for comprehensively evaluating the multimodal analysis capabilities of LLMs. Li et al [93] evaluated the capability of GPT-4V in medical VQA tasks. The results indicated that the model excelled in distinguishing question types but did not meet existing benchmarks in accuracy. The advantage lies in its strong language understanding and question classification abilities, while the limitation is the underutilization of medical image information. However, some medical specialties, such as ophthalmology, still lack large-scale VQA datasets. Mihalache et al [15] evaluated the performance of the AI chatbot ChatGPT-4 in interpreting a dataset of multimodal ophthalmic images. The study results demonstrated that ChatGPT-4 performed well in the task of identifying types of ophthalmic examinations, achieving an accuracy rate of 70%. However, when it came to lesion identification, the model's average accuracy was only 65%, indicating the limitations of ChatGPT-4 in accurately recognizing and describing lesions from ophthalmic images. Xu et al [7] tested the ability of the GPT-4V model on the VQA task on datasets of multiple ophthalmic imaging modalities. The datasets included slit-lamp, scanning laser ophthalmoscopy, fundus photography, OCT, FFA, and ocular ultrasonography images. The results showed that GPT-4V performed well in the task of identifying ophthalmic examination types, with an accuracy of 95.6%. However, in terms of lesion identification, the model's average accuracy was only 25.6%, indicating limitations of GPT-4V in accurately identifying and describing lesions from ophthalmic images. Therefore, constructing high-quality VQA datasets within each medical specialty is of great significance for comprehensively evaluating the performance of LLMs.

Evaluation of VQA tasks currently employs a comprehensive approach by combining both automated and manual assessments to thoroughly examine the quality of generated answers. Automated evaluation includes classification metrics such as accuracy, F1-score, and language metrics like BLEU, enabling large-scale assessment of answer accuracy and natural language generation capabilities. However, manual evaluation allows for a more detailed and personalized inspection of response expertise and appropriateness. Hallucination remains a critical challenge in medical VQA. Using the OmniMedVQA benchmark, Hu et al [94] found that even high-performing large-scale models often generate plausible-sounding but incorrect answers when faced with questions requiring detailed observation and specialized knowledge. To systematically evaluate this issue, Gu et al [95] proposed the MedVH assessment framework, which evaluates hallucinations in medical VQA across multiple dimensions, including factuality, consistency, and relevance. Wu et al [96] further developed a specialized benchmark dataset targeting hallucinations in medical VQA. Their findings revealed that models are particularly prone to hallucinations when handling negation-based questions or multi-step reasoning tasks, and they proposed targeted improvements to mitigate these issues. In terms of reasoning capabilities in medical VQA, Beşler et al [26] evaluated GPT-4o's performance in a radiology exam that includes a clinically oriented reasoning part. Their findings highlighted the potential of LLMs to assist radiologists in evaluating and managing cases, even in zero-shot scenarios.

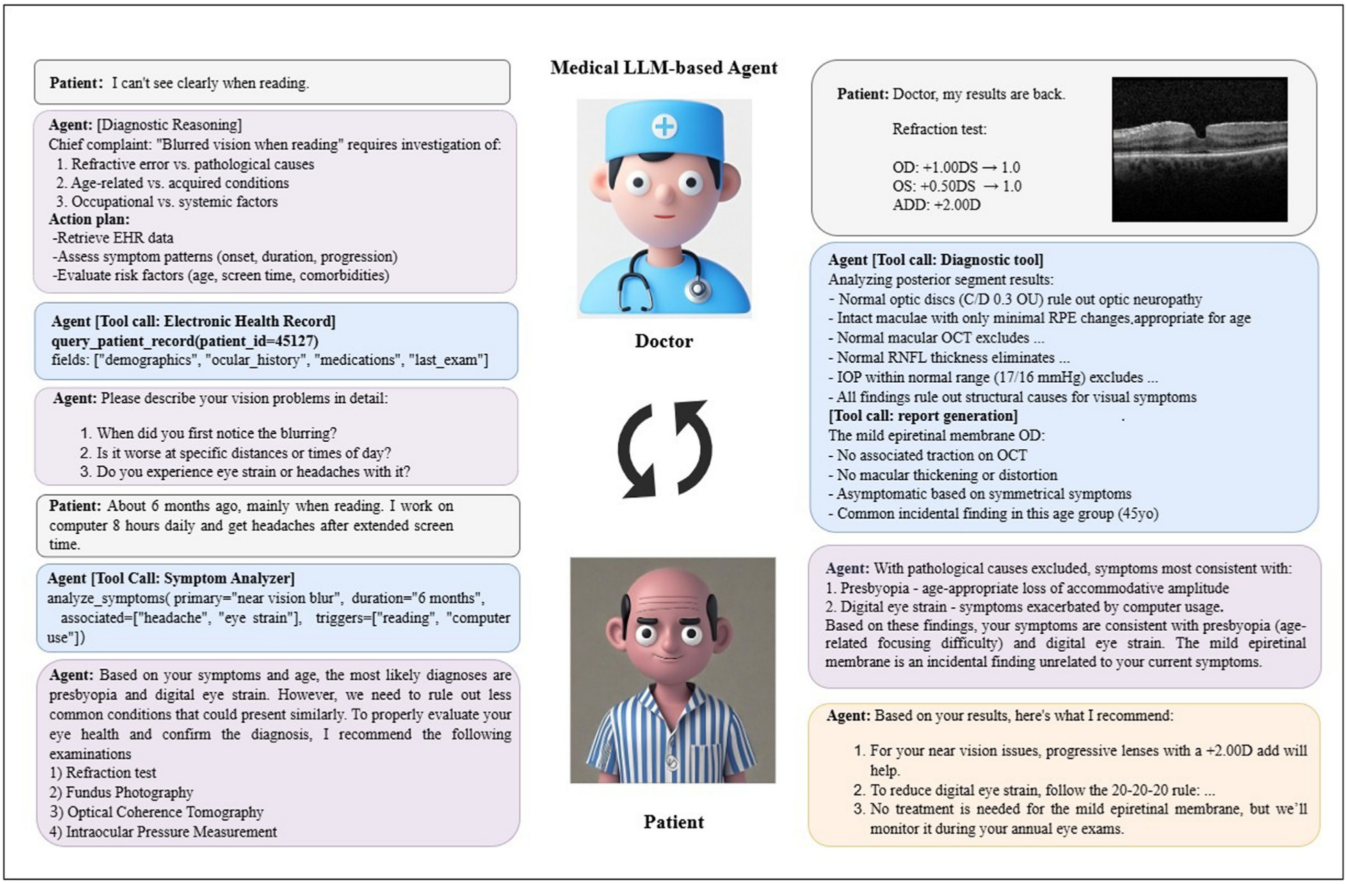
4.3.4. Others

In addition to the application in scenarios such as image classification, report generation, and VQA, LLMs also show good prospects in other medical image analysis tasks such as medical image segmentation and cross-modal retrieval. In medical image segmentation, LLMs can assist by interpreting key information from medical reports or prompts and integrating this knowledge to guide image segmentation algorithms, enhancing their accuracy and efficiency. Oh et al [97] proposed a multimodal AI system called LLMSeg, designed for target delineation in radiotherapy. They evaluated LLMSeg by calculating metrics such as the Dice coefficient, intersection over union, and the 95th percentile of Hausdorff distance (95-HD) and found that LLMSeg significantly outperformed image-only AI models on both internal and external validation datasets, maintaining consistent performance even with reduced training data. For cross-modal retrieval tasks, LLMs act as a bridge by translating natural language commands into actionable signals for image processing models. Lei et al [98] proposed GPT-CMR and evaluated its performance using the Chinese Medical Instructional Video Question Answering dataset. Metrics such as R@1, R@10, R@50, MRR, and overall value were employed. The results showed significant improvements over the baseline, highlighting the potential of LLMs in cross-modal retrieval tasks. Current evaluations of LLMs in these new tasks are inadequate, which shows a necessity for a deeper investigation into evaluation benchmarks. Future research should enhance LLM assessment by employing varied datasets and detailed metrics, and by focusing on improving model interpretability and explainability to facilitate clear decision-making.

4.4. Real-world multitask scenarios involving LLM agents

While the above 3 scenarios focus on evaluating LLMs in isolated tasks (e.g., single-turn diagnosis or structured report generation), real-world clinical workflows require seamless integration of multiple interdependent subtasks, including diagnostic reasoning, image lesion segmentation, report generation, and multimodal QA. In such scenarios, traditional LLM applications may fall short, whereas the latest LLM agents are more suitable [99,100] (Figure 5). Agent architecture can be can be categorized into 4 levels based on autonomy: Level 1 is the generator agent, which achieves precise responses with the help of techniques like retrieval-augmented generation (RAG); Level 2 builds on this by integrating expert model toolkits, allowing system function expansion through tool calls; Level 3 is the planning agent, capable of using tools while also employing reasoning and planning abilities to construct multi-step workflows based on user inquiries and adjusting execution according to outcomes; and Level 4 is characterized by high autonomy, incorporating tool calls, reasoning frameworks, and workflow planning capabilities [101–103]. Given the multi-step, multifunction nature of LLM agents, specialized evaluation frameworks are necessary. Several studies demonstrate promising directions. AgentBench, the first cross-domain benchmark, evaluates LLM agents across 8 simulated environments, revealing significant performance gaps in open-ended decision-making [104]. In healthcare, RadABench pioneered the evaluation of radiology-specific agents by simulating tool-rich workflows, encompassing 6 anatomies, 5 imaging modalities, 10 tool categories, and 11 radiology tasks [105].

In addition, the evaluation of LLMs and LLM agents in real-world medical scenarios should not be confined to static dataset benchmarks but should simulate the fluidity of real-world medical workflows. In this regard, Tu et al [106] innovatively assessed LLM performance in the style of a remote objective structured clinical examination (OSCE). OSCE is an objective and organized assessment framework that provides a simulation of real-world clinical scenarios to assess the clinical competency of medical students. Patient actors consulted with real physicians and LLMs in random order, and the conversations were assessed by patient actors and specialists, respectively. Johri et al [107] further shifted the benchmark from traditional methods reliant on structured medical



Patient: I can't see clearly when reading.

Agent: [Diagnostic Reasoning]
Chief complaint: "Blurred vision when reading" requires investigation of:
1. Refractive error vs. pathological causes
2. Age-related vs. acquired conditions
3. Occupational vs. systemic factors
Action plan:
-Retrieve EHR data
-Assess symptom patterns (onset, duration, progression)
-Evaluate risk factors (age, screen time, comorbidities)

Agent [Tool call: Electronic Health Record]
query_patient_record(patient_id=45127)
fields: ["demographics", "ocular_history", "medications", "last_exam"]

Agent: Please describe your vision problems in detail:

1. When did you first notice the blurring?
2. Is it worse at specific distances or times of day?
3. Do you experience eye strain or headaches with it?

Patient: About 6 months ago, mainly when reading. I work on computer 8 hours daily and get headaches after extended screen time.

Agent [Tool Call: Symptom Analyzer]
analyze_symptoms(primary="near vision blur", duration="6 months", associated=["headache", "eye strain"], triggers=["reading", "computer use"])

Agent: Based on your symptoms and age, the most likely diagnoses are presbyopia and digital eye strain. However, we need to rule out less common conditions that could present similarly. To properly evaluate your eye health and confirm the diagnosis, I recommend the following examinations
1) Refraction test
2) Fundus Photography
3) Optical Coherence Tomography
4) Intraocular Pressure Measurement

Patient: Doctor, my results are back.

Refraction test:

OD: +1.00DS → 1.0
OS: +0.50DS → 1.0
ADD: +2.00D

Agent [Tool call: Diagnostic tool]
Analyzing posterior segment results:
- Normal optic discs (C/D 0.3 OU) rule out optic neuropathy
- Intact maculae with only minimal RPE changes,appropriate for age
- Normal macular OCT excludes ...
- Normal RNFL thickness eliminates ...
- IOP within normal range (17/16 mmHg) excludes ...
- All findings rule out structural causes for visual symptoms
[Tool call: report generation]
The mild epiretinal membrane OD:
- No associated traction on OCT
- No macular thickening or distortion
- Asymptomatic based on symmetrical symptoms
- Common incidental finding in this age group (45yo)

Agent: With pathological causes excluded, symptoms most consistent with:
1. Presbyopia - age-appropriate loss of accommodative amplitude
2. Digital eye strain - symptoms exacerbated by computer usage.
Based on these findings, your symptoms are consistent with presbyopia (age-related focusing difficulty) and digital eye strain. The mild epiretinal membrane is an incidental finding unrelated to your current symptoms.

Agent: Based on your results, here's what I recommend:

1. For your near vision issues, progressive lenses with a +2.00D add will help.
2. To reduce digital eye strain, follow the 20-20-20 rule: ...
3. No treatment is needed for the mild epiretinal membrane, but we'll monitor it during your annual eye exams.

Figure 5. Examples of real-world multitask scenarios involving LLM Agents. The optical coherence tomography image used in the figure was originally published in the Retina Image Bank. John S. Pollack. Epiretinal Membrane. Retina Image Bank. 2012; 181. © the American Society of Retina Specialists. LLM: large language model; OCT: optical coherence tomography.

examinations to leveraging AI agents and LLMs for natural dialogue. The developed CRAFT-MD evaluation framework emphasizes realistic doctor-patient interactions, comprehensive medical history collection, open-ended questioning, and a combined approach using automation and expert evaluation. These frameworks represent the beginning of a shift toward a process-centered evaluation paradigm for LLMs. Future benchmarks may need to further refine multiple dimensions and conduct RCTs to compare whether LLM truly aids patients and assists in clinical practices.

5. Evaluation method

When evaluating LLMs in the field of medicine, it is necessary to consider both the performance of the model and its potential impact on patient health. This process involves not only automated assessments to quantify the model's task-specific capabilities but also manual evaluations to measure the quality, accuracy, and applicability of the model's outputs to real medical scenarios. These evaluation methods can prioritize different dimensions depending on the specific tasks.

5.1. Automatic evaluation

Automatic evaluation focuses on objectively assessing the performance of LLMs through automatic algorithms. In classification tasks, metrics such as accuracy, specificity, precision, sensitivity, and F1-score are used to quantify the performance of model predictions [108]. Duong et al [30] compared the accuracy of ChatGPT and human respondents in answering genetic questions. Cai et al [22] evaluated Bing Chat, ChatGPT 3.5, and 4.0 in answering 250 questions in basic and clinical science self-assessment projects, with the primary outcome being response accuracy.

For long-text tasks, several metrics are used to evaluate the quality of generated text. Metrics such as BLEU, ROUGE, CIDEr, and METEOR focus on text overlap and assess the literal accuracy of the generated text [29,91,109,110]. However, metrics like bidirectional encoder representations from transformers score (BERTScore) and moverscore measure semantic similarity, evaluating the semantic accuracy and consistency of expression [10,111]. In the context of evaluating medical text generation tasks in open-ended QA scenarios, specialized tools like QAEval and QAFactEval have also been employed [46]. To assess the fluency and readability of the text, several commonly used metrics are available, including the Flesch Reading Ease Score, Flesch-Kincaid grade level, Gunning Fog Index, Coleman-Liau Index, and Simple Measure of Gobbledygook [37]. These metrics are utilized to determine whether the medical content generated by the language model is informative, effective in conveying information, and user-friendly.

5.2. Human evaluation

Automated evaluation methods fall short in covering all essential aspects, particularly in sensitive domains like medicine which require advanced knowledge and ethical judgment, making manual evaluation crucial in such cases [29]. One relatively simple way to facilitate manual evaluation is to use qualitative methods such as case studies [112]. These allow manual evaluators to carefully compare the content of the

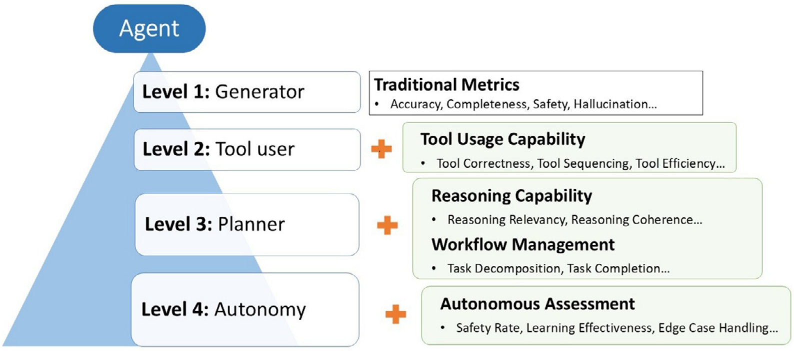

Agent
Level 1: Generator
Traditional Metrics
• Accuracy, Completeness, Safety, Hallucination...
Level 2: Tool user
Tool Usage Capability
• Tool Correctness, Tool Sequencing, Tool Efficiency...
Level 3: Planner
Reasoning Capability
• Reasoning Relevancy, Reasoning Coherence...
Workflow Management
• Task Decomposition, Task Completion...
Level 4: Autonomy
Autonomous Assessment
• Safety Rate, Learning Effectiveness, Edge Case Handling...

Figure 6. LLM Agent specific evaluation dimensions. LLM: large language model.

LLM with the ground truth (GT), thereby revealing subtle differences that automated evaluation methods cannot identify.

To facilitate scientific review and statistical analysis of LLM outputs by manual evaluators, various scoring protocols have been adopted to assess the quality of the generated response. Standardized scales are applied, including the DISCERN scale [113], the *JAMA* benchmark criteria [114], the Global Quality scale [115], and others. Another approach is to identify and statistically analyze the occurrence probabilities of different types of predefined errors, such as factual errors and logical errors [12,109]. To diversify evaluation modes, some studies use custom grading rules or Likert-style rules to assess text quality across multiple levels. The Likert scale is widely used in social science and psychological research to evaluate people's views or attitudes toward specific viewpoints [116]. In the field of medicine, each evaluation dimension to be considered can be transformed into a series of statements, and corresponding answer options can be provided to investigate the degree of identification of respondents with different dimensions of performance. For example, Samaan et al [31] recruited board-certified bariatric surgeons and used a 4-point scale to evaluate both the accuracy and comprehensiveness of ChatGPT, where 1 represented comprehensive, 2 represented correct but insufficient, 3 represented partially correct and partially incorrect, and 4 represented completely incorrect. Chen et al [91] used a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree) to evaluate the completeness and correctness of reports generated by the model.

Collecting opinions from evaluators at various levels during manual assessment can enhance the comprehensiveness of the evaluation process. Currently, the assessment of most studies is conducted by professional physicians [53,117–119]. However, relying exclusively on doctors as evaluators may not align with the development of patient-centered medical LLMs. Therefore, some studies have also included non-professionals, such as patients and the general public, to participate in evaluating the LLMs [32,35]. For instance, Singhal et al [4] engaged 5 non-medical Indian raters to analyze the usefulness and practicality of LLM's responses to long-form questions. This approach enables the capture of details that may be overlooked when solely relying on expert perspectives.

6. Evaluation dimensions

6.1. Traditional evaluation dimensions

Accuracy is the most commonly used dimension, typically measured by the proportion of correct answers relative to the GT in MCQs [120], or by human scoring based on established guidelines in open-ended questions [10]. Many studies also use natural language metrics to assess the semantic consistency between the generated text and the GT text as an alternative to accuracy, such as BLEU, ROUGE, and others [91]. In addition to accuracy and semantic consistency, some studies also consider completeness, safety, communication, and user satisfaction (Figure 2). The growing capabilities of multifunctional LLMs underscore the need for more comprehensive evaluation dimensions. For instance, LLMs remain prone to generating inaccuracies, often referred to as "hallucinations." Currently, there is no standardized method for evaluating hallucinations in medicine. One study quantified hallucination using accuracy scores below 4 and classified errors into categories such as unrelated information, factual inaccuracies, incomplete responses, and faulty logic [9]. Another study evaluated "reference hallucinations" by determining whether the generated reference was real or fabricated [27]. Additionally, despite their importance, few models include assessments of bias, stability, or cost-effectiveness [47,121,122]. Regarding the integration of a multidimensional evaluation framework, Singhal et al [4] created a comprehensive evaluation framework involving 12 aspects, including scientific consensus, extent of possible harm, likelihood of possible harm, evidence of correct, comprehension, evidence of correct retrieval, evidence of correct reasoning, evidence of incorrect comprehension, evidence of incorrect retrieval, evidence of incorrect reasoning, inappropriate/incorrect content, missing content, possibility of bias, providing important reference for future research.

6.2. LLM agent evaluation dimensions

Traditional LLM evaluation dimensions remain applicable for assessing LLM agents, but they need to be expanded to include additional aspects of intermediate processes, such as tool usage capability, reasoning capability, workflow management, and autonomous assessment (Figure 6). Level 1 agents, which only involve RAG, can still be evaluated using traditional metrics like accuracy and safety. However, as we move to Level 2, where expert model toolkits are integrated, the evaluation must assess the agent's tool usage capabilities. This may include 3 key components: tool correctness (whether the correct tool was called), tool sequencing (whether the tools were optimally combined), and tool efficiency (whether there were redundant tool calls). For Level 3 evaluation, reasoning becomes crucial as it connects multiple steps in the workflow. A recent study proposed an evaluation system to assess the essential metacognition required for medical reasoning, incorporating confidence scores and metacognitive tasks into the benchmark, providing valuable insights for evaluating key reasoning aspects of LLM performance [123]. Another study categorized medical reasoning into diagnosis-related and management-oriented types: the former evaluates the logical chain from patient histories, signs, and differential diagnoses to the final diagnosis, while the latter examines the derivation process from specific medical entities to personalized treatment plans based on evidence-based medicine [124]. In addition, at this level, the agent's workflow management is a critical evaluation focus, emphasizing the rationale behind task decomposition for complex medical inquiries and the completion rate of planned actions. Level 4, which includes all the dimensions mentioned above, also focuses on autonomous assessment. This should involve evaluating the agent's safety rate during autonomous operations, as well as its ability to provide continuous feedback optimization through multiple

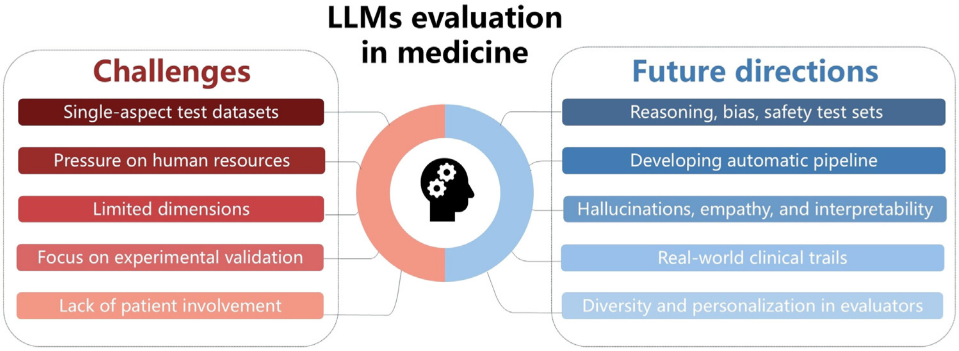

LLMs evaluation in medicine
Challenges
Single-aspect test datasets
Pressure on human resources
Limited dimensions
Focus on experimental validation
Lack of patient involvement
Future directions
Reasoning, bias, safety test sets
Developing automatic pipeline
Hallucinations, empathy, and interpretability
Real-world clinical trails
Diversity and personalization in evaluators

Figure 7. Overview of challenges and future directions of LLM evaluation in medicine. LLM: large language model.

rounds of responses, thus assessing its learning effectiveness. Moreover, this level must ensure the agent correctly identifies boundaries, avoiding speculative answers for unsolvable cases.

7. Discussion

To ensure the reliable application of LLMs in the medical field, constructing precise and effective evaluation frameworks is crucial. This section delves into the challenges currently faced by the evaluation of medical LLMs and proposes strategies for future development (Figure 7).

First, establishing suitable evaluation frameworks requires the preparation of appropriate datasets. Although datasets for categorization and report generation are relatively abundant, there is a lack of high-quality medical QA datasets. In particular, crafting high-quality VQA datasets poses a significant challenge; they require the selection of medically compliant images, including diverse types of questions and answers. Furthermore, the construction of evaluation datasets should not only encompass a rich understanding of professional medical knowledge but also take into consideration comprehensive tests for practical applications, such as clinical reasoning cases, bias test sets, and refusal safety test sets. This process demands not only data selection from real-world clinics but also a substantial involvement of medical professionals. However, given that medical professionals have limited time and energy, primarily focused on clinical work and research, the additional burden of participating in dataset construction is often impractical [125]. Therefore, future studies may explore the combination of human and LLM capabilities to collaboratively build and continuously optimize evaluation datasets, thereby enhancing the scalability of datasets and reducing the manual workload.

Second, setting comprehensive and effective assessment standards is also a key component in the current evaluation frameworks. Evaluation dimensions traditionally concentrate on accuracy, completeness, and safety, which are indeed crucial in the medical field as they relate to the model's ability to prevent misdiagnoses and missed diagnoses. However, the medical domain particularly needs to further refine other aspects, such as hallucination, empathy, and interpretability [126–128]. For instance, there is a need for a more precise and hierarchical classification system for hallucinations, taking into account the types and severity of hallucinations for scoring purposes. In addition, with the expansion of LLM application scenarios, evaluating them as intelligent agents has become an emerging direction. As previously mentioned, such evaluations need to integrate traditional LLM evaluation dimensions with agent-specific dimensions (e.g., tool usage capability, clinical reasoning) and build targeted evaluation frameworks based on the characteristics of specific tasks.

Regarding evaluation metrics, the focus remains predominantly on traditional classification metrics and NLP metrics. Some studies have attempted to utilize advanced LLMs like GPT4 for automated assessment [11,129]. For instance, Yan et al [130] have proposed an automatic evaluation model based on MedLLaMA, which aims to assess the correctness, expertise, and completeness of answers in open clinical scenarios. However, the stability and robustness of these automatic evaluation methods still require further validation. Appropriate response is not limited to correctness but also considers applicability in different contexts. Responses that deviate from the standard answer might be appropriate, while responses that are similar or consistent with the GT could still be misleading. Therefore, in the short term, we cannot entirely dispense with human evaluation. The future direction should involve further development and validation of automated evaluation systemsthat balance safeguarding rational assessments while alleviating pressure on human resources and facilitating large-scale evaluations.

As for the practical application of LLMs, most studies are currently at the preclinical validation stage. Li et al [131] have designed a clinical trial to validate the actual utility of LLM in enhancing primary diabetes care and diabetic retinopathy screening. Future trial designs need to be more aligned with real-world clinics and compare LLMs to existing practices—including other health care systems, traditional AI tools, and healthcare professionals of different levels—to truly assess their value in practical applications. Appropriate endpoints, such as reducing morbidity, improving work efficiency, and patient or physician satisfaction, are required to gauge success or failure. The design of LLM interventions in clinical trials can benefit from the application of non-LLM chatbots in RCTs [5]. When designing an assessment framework for medical LLMs, it is crucial to ensure diversity and personalization in the selection of evaluators. This includes not only physicians but also incorporating perspectives from patients, medical students, and other real users, based on specific application scenarios and functionalities. In fact, generalist LLMs have experimented with collecting user feedback for online assessments [132]. They analyze service logs and directly or indirectly assess user satisfaction, enabling close-to-real-world scenario efficacy assessment. This method not only garners valuable and actual user feedback but also provides an ideal path for continuous performance monitoring, which could also be expected to apply to LLMs in the medical field.

A precise and effective evaluation framework is indispensable for the assessment of medical LLMs. Recently, Abbasian et al [133] proposed a five-element evaluation framework including models, environments, interfaces, interactive users, and leaderboards, offering valuable references for research and applications in the sector.

In conclusion, we provide an overview of the recent advancements in evaluating LLMs in the medical field, with a special focus on key elements of the evaluation framework, including datasets, evaluation methods, dimensions, and scenarios. Future research should harness the interdisciplinary expertise of medical professionals and computer scientists to address the existing challenges in various domains and ultimately optimize the application of LLMs to enhance the quality and efficiency of medical services and patient experience.

Conflicts of interest statement

The authors declare that they have no competing interests.

Fundings

The study was supported by the Start-up Fund for RAPs under the Strategic Hiring Scheme (Grant No. P0048623) from HKSAR, the Global STEM Professorship Scheme (Grant No. P0046113) and Henry G. Leong Endowed Professorship in Elderly Vision Health.

Author contributions

Xiaolan Chen: Writing – original draft. **Jiayang Xiang:** Writing – original draft. **Shanfu Lu:** Writing – original draft. **Yexin Liu:** Writing – review & editing. **Mingguang He:** Funding acquisition. **Danli Shi:** Writing – review & editing, Conceptualization.

Acknowledgement

We thank the InnoHK HKSAR Government for providing valuable supports.The research work described in this paper was conducted in the JC STEM Lab of Innovative Light Therapy for Eye Diseases funded by The Hong Kong Jockey Club Charities Trust.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.imed.2025.03.002.

References

[1] Ozkaya I. Application of large language models to software engineering tasks: opportunities, risks, and implications. IEEE Softw 2023;40(3):4–8. doi:10.1109/ms.2023.3248401.

[2] Sarsa S, Denny P, Hellas A, et al. Automatic generation of programming exercises and code explanations using large language models. 2022. doi: 10.48550/ARXIV.2206.11861.

[3] Alqahtani T, Badreldin HA, Alrashed M, et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Res Social Adm Pharm 2023;19(8):1236–42. doi:10.1016/j.sapharm.2023.05.016.

[4] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80. doi:10.1038/s41586-023-06291-2.

[5] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med 2023;29(8):1930–40. doi:10.1038/s41591-023-02448-8.

[6] Demszky D, Yang D, Yeager DS, et al. Using large language models in psychology. Nat Rev Psychol 2023. doi:10.1038/s44159-023-00241-5.

[7] Xu P, Chen X, Zhao Z, et al. Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. Br J Ophthalmol 2024;108(10):1384–9. doi:10.1136/bjo-2023-325054.

[8] Zhang C, Liu S, Zhou X, et al. Examining the role of large language models in orthopedics: systematic review. J Med Internet Res 2024;26:e59607. doi:10.2196/59607.

[9] Chen X, Zhao Z, Zhang W, et al. EyeGPT for patient inquiries and medical education: development and validation of an ophthalmology large language model. J Med Internet Res 2024;26:e60063. doi:10.2196/60063.

[10] Chen X, Zhang W, Xu P, et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. NPJ Digit Med 2024;7(1):111. doi:10.1038/s41746-024-01101-z.

[11] Zhang H, Chen J, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor. 2023. doi:10.48550/ARXIV.2305.15075.

[12] Tu T, Azizi S, Driess D, et al. Towards generalist Biomedical AI. NEJM AI 2024;1(3). doi:10.1056/aioa2300138.

[13] Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep 2023;13(1):16492. doi:10.1038/s41598-023-43436-9.

[14] Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025;333(4):319–28. doi:10.1001/jama.2024.21700.

[15] Mihalache A, Huang RS, Popovic MM, et al. Accuracy of an artificial intelligence Chatbot's interpretation of clinical ophthalmic images. JAMA Ophthalmol 2024;142(4):321–6. doi:10.1001/jamaophthalmol.2024.0017.

[16] Royer C, Sekuboyina A. MultiMedEval: A benchmark and a toolkit for evaluating medical vision-language models. Medical Imaging with Deep Learning.

[17] Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. doi:10.2196/45312.

[18] Wang H, Wu W, Dou Z, et al. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform 2023;177:105173. doi:10.1016/j.ijmedinf.2023.105173.

[19] Zong H, Li J, Wu E, et al. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ 2024;24(1):143. doi:10.1186/s12909-024-05125-7.

[20] Li KC, Bu ZJ, Shahjalal M, et al. Performance of ChatGPT on Chinese master's degree entrance examination in clinical medicine. PLoS One 2024;19(4):e0301702. doi:10.1371/journal.pone.0301702.

[21] Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023;3(4):100324. doi:10.1016/j.xops.2023.100324.

[22] Cai LZ, Shaheen A, Jin A, et al. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol 2023;254:141–9. doi:10.1016/j.ajo.2023.05.024.

[23] Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023;93(6):1353–65. doi:10.1227/neu.0000000000002632.

[24] Long C, Lowe K, Zhang J, et al. A novel evaluation model for assessing ChatGPT on otolaryngology-head and neck surgery certification examinations: performance study. JMIR Med Educ 2024;10:e49970. doi:10.2196/49970.

[25] Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary Care. JMIR Med Educ 2023;9:e46599. doi:10.2196/46599.

[26] Beşler MS, Oleaga L, Junquero V, et al. Evaluating GPT-4o's performance in the official European Board of Radiology exam: a Comprehensive Assessment. Acad Radiol 2024;31(11):4365–71. doi:10.1016/j.acra.2024.09.005.

[27] Gibson D, Jackson S, Shanmugasundaram R, et al. Evaluating the efficacy of ChatGPT as a patient education in prostate cancer: Multimetric assessment. J Med Internet Res 2024;26:e55939. doi:10.2196/55939.

[28] Mukherjee P, Hou B, Suri A, et al. Evaluation of GPT large language model performance on RSNA 2023 case of the day questions. Radiology 2024;313(1):e240609. doi:10.1148/radiol.240609.

[29] Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med 2023;6(1):158. doi:10.1038/s41746-023-00896-7.

[30] Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. Eur J Hum Genet 2024;32(4):466–8. doi:10.1038/s41431-023-01396-8.

[31] Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023;33(6):1790–6. doi:10.1007/s11695-023-06603-5.

[32] Ye C, Zweck E, Ma Z, et al. Doctor versus artificial intelligence: patient and physician evaluation of large language model responses to rheumatology patient questions in a cross-sectional study. Arthritis Rheuratol 2024;76(3):479–84. doi:10.1002/art.42737.

[33] Kim H, Kim P, Joo I, et al. ChatGPT vision for radiological interpretation: an investigation using medical school radiology examinations. Korean J Radiol 2024;25(4):403–6. doi:10.3348/kjr.2024.0017.

[34] Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digit Health 2021;3(1):e7. doi:10.1016/S2589-7500(20)30290-9.

[35] Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 2023;13:1219326. doi:10.3389/fonc.2023.1219326.

[36] Hurley ET, Crook BS, Lorentz SG, et al. Evaluation of high-quality information from ChatGPT (artificial intelligence-large language model) artificial intelligence on shoulder stabilization surgery. Arthroscopy 2024;40(3):726–31. doi:10.1016/j.arthro.2023.07.048.

[37] Onder CE, Koc G, Gokbulut P, et al. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep 2024;14(1):243. doi:10.1038/s41598-023-50884-w.

[38] Marshall RF, Mallem K, Xu H, et al. Investigating the accuracy and completeness of an artificial intelligence large language model about uveitis: an evaluation of ChatGPT. Ocul Immunol Inflamm 2024;32(9):2052–5. doi:10.1080/09273948.2024.2317417.

[39] Zakka C, Shad R, Chaurasia A, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI 2024;1(2). doi:10.1056/aioa2300068.

[40] Liu J, Zhou P, Hua Y, et al. Benchmarking large language models on CMExam-a comprehensive Chinese medical exam dataset. Adv Neural Inf Process Syst 2024;36.

[41] Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American college of gastroenterology self-assessment test. Am J Gastroenterol 2023;118(12):2280–2. doi:10.14309/ajg.0000000000002320.

[42] Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307(5):e230582. doi:10.1148/radiol.230582.

[43] Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J 2023;43(12):NP1078–82. doi:10.1093/asj/sjad128.

[44] Li W, Li L, Xiang T, et al. Can multiple-choice questions really be useful in detecting the abilities of LLMs? 2024. doi:10.48550/ARXIV.2403.17752.

[45] Cheng SL, Tsai SJ, Bai YM, et al. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. J Med Internet Res 2023;25:e51229. doi:10.2196/51229.

[46] Tang X, Cohan A, Gerstein M. Aligning factual consistency for clinical studies summarization through reinforcement learning. In: Proceedings of the 5th Clinical Natural Language Processing Workshop; 2023.

[47] Hake J, Crowley M, Coy A, et al. Quality, accuracy, and bias in ChatGPT-based summarization of medical abstracts. Ann Fam Med 2024;22(2):113–20. doi:10.1370/afm.3075.

[48] Caterson J, Ambler O, Cereceda-Monteoliva N, et al. Application of generative language models to orthopaedic practice. BMJ Open 2024;14(3):e076484. doi:10.1136/bmjopen-2023-076484.

[49] Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5(3):e107–8. doi:10.1016/S2589-7500(23)00021-3.

[50] Decker H, Trang K, Ramirez J, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. JAMA Netw Open 2023;6(10):e2336997. doi:10.1001/jamanetworkopen.2023.36997.

[51] Dubinski D, Won SY, Trnovec S, et al. Leveraging artificial intelligence in neurosurgery-unveiling ChatGPT for neurosurgical discharge summaries and operative reports. Acta Neurochir (Wien) 2024;166(1):38. doi:10.1007/s00701-024-05908-3.

[52] Zaretsky J, Kim JM, Baskharoun S, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024;7(3):e240357. doi:10.1001/jamanetworkopen.2024.0357.

[53] Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6(1):9. doi:10.1186/s42492-023-00136-5.

[54] Chung EM, Zhang SC, Nguyen AT, et al. Feasibility and acceptability of ChatGPT generated radiology report summaries for cancer patients. Digital Health 2023;9. doi:10.1177/20552076231221620.

[55] Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134–42. doi:10.1038/s41591-024-02855-5.

[56] Chen Q, Sun H, Liu H, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. Bioinformatics 2023;39(9):btad557. doi:10.1093/bioinformatics/btad557.

[57] Jahan I, Laskar MTR, Peng C, et al. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. Comput Biol Med 2024;171:108189. doi:10.1016/j.compbiomed.2024.108189.

[58] Gu Z, He X, Yu P, et al. Automatic quantitative stroke severity assessment based on Chinese clinical named entity recognition with domain-adaptive pre-trained large language model. Artif Intell Med 2024;150:102822. doi:10.1016/j.artmed.2024.102822.

[59] Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med 2024;7(1):6. doi:10.1038/s41746-023-00970-0.

[60] Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023;6(1):210. doi:10.1038/s41746-023-00958-w.

[61] Cinquin O. ChIP-GPT: a managed large language model for robust data extraction from biomedical database records. Brief Bioinform 2024;25(2):bbad535. doi:10.1093/bib/bbad535.

[62] Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023;329(10):842–4. doi:10.1001/jama.2023.1044.

[63] Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology 2024;80(5):1158–68. doi:10.1097/HEP.0000000000000834.

[64] Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023;21(1):269. doi:10.1186/s12967-023-04123-5.

[65] Xie Y, Seth I, DJ Hunter-Smith, et al. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. Aesthetic Plast Surg 2023;47(5):1985–93. doi:10.1007/s00266-023-03338-7.

[66] Huang AS, Hirabayashi K, Barna L, et al. Assessment of a large language model's responses to questions and cases about glaucoma and retina management. JAMA Ophthalmol 2024;142(4):393. doi:10.1001/jamaophthalmol.2024.1158.

[67] Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589–96. doi:10.1001/jamainternmed.2023.1838.

[68] Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29(3):721–32. doi:10.3350/cmh.2023.0089.

[69] Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. J Med Internet Res 2023;25:e48659. doi:10.2196/48659.

[70] Civettini I, Zappaterra A, Granelli BM, et al. Evaluating the performance of large language models in haematopoietic stem cell transplantation decision-making. Br J Haematol 2024;204(4):1523–8. doi:10.1111/bjh.19200.

[71] Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023;9(1):44. doi:10.1038/s41523-023-00557-8.

[72] Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. Ophthalmic Plast Reconstr Surg 2023;39(3):221–5. doi:10.1097/IOP.0000000000002418.

[73] Potapenko I, Boberg-Ans LC, Stormly Hansen M, et al. Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. Acta Ophthalmol 2023;101(7):829–31. doi:10.1111/aos.15661.

[74] Rasmussen MLR, Larsen AC, Subhi Y, et al. Artificial intelligence-based ChatGPT chatbot responses for patient and parent questions on vernal keratoconjunctivitis. Graefes Arch Clin Exp Ophthalmol 2023;261(10):3041–3. doi:10.1007/s00417-023-06078-1.

[75] Wilhelm TI, Roos J, Kaczmarczyk R. Large Language models for therapy recommendations across 3 clinical specialties: comparative study. J Med Internet Res 2023;25:e49324. doi:10.2196/49324.

[76] Uz C, Umay E. Dr ChatGPT": is it a reliable and useful source for common rheumatic diseases? Int J Rheum Dis 2023;26(7):1343–9. doi:10.1111/1756-185X.14749.

[77] Lee TC, Staller K, Botoman V, et al. ChatGPT answers common patient questions about colonoscopy. Gastroenterology 2023;165(2):509–11. doi:10.1053/j.gastro.2023.04.033.

[78] Cao JJ, Kwon DH, Ghaziani TT, et al. Accuracy of information provided by ChatGPT regarding liver cancer surveillance and diagnosis. AJR Am J Roentgenol 2023;221(4):556–9. doi:10.2214/AJR.23.29493.

[79] Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023;124(5):101471. doi:10.1016/j.jormas.2023.101471.

[80] Munoz-Zuluaga C, Zhao Z, Wang F, et al. Assessing the accuracy and clinical utility of ChatGPT in laboratory medicine. Clin Chem 2023;69(8):939–40. doi:10.1093/clinchem/hvad058.

[81] Johnson SB, King AJ, Warner EL, et al. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. JNCI Cancer Spectr 2023;7(2):kad015. doi:10.1093/jncics/pkad015.

[82] Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis 2023;23(4):405–6. doi:10.1016/S1473-3099(23)00113-5.

[83] Young JN, O'Hagan Ross, Poplausky D, et al. The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. J Am Acad Dermatol 2023;89(3):602–4. doi:10.1016/j.jaad.2023.05.024.

[84] Cadamuro J, Cabitza F, Debeljak Z, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. Clin Chem Lab Med 2023;61(7):1158–66. doi:10.1515/cclm-2023-0355.

[85] Menz BD, Kuderer NM, Bacchi S, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ 2024;384:e078538. doi:10.1136/bmj-2023-078538.

[86] Van MH, Verma P, Wu X. On large visual language models for medical imaging analysis: an empirical study. IEEE; 2024.

[87] Wan Z, Liu C, Zhang M, et al. Med-unic: unifying cross-lingual medical vision-language pre-training by diminishing bias. Adv Neural Inf Process Syst 2024;36.

[88] Liu Y, Li Y, Wang Z, et al. A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis. Meta-Radiology 2024;2(4):100099. doi:10.1016/j.metrad.2024.100099.

[89] Wang Z, Liu L, Wang L, et al. R2gengpt: radiology report generation with frozen llms. Meta-Radiology 2023;1(3):100033. doi:10.1016/j.metrad.2023.100033.

[90] Lee S, Kim WJ, Chang J, et al. LLM-CXR: instruction-finetuned LLM for CXR image understanding and generation The Twelfth International Conference on Learning Representations; 2023.

[91] Chen X, Zhang W, Zhao Z, et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. Br J Ophthalmol 2024;108(10):1450–6. doi:10.1136/bjo-2023-324446.

[92] Lin Z, Zhang D, Tao Q, et al. Medical visual question answering: a survey. Artif Intell Med 2023;143:102611. doi:10.1016/j.artmed.2023.102611.

[93] Li Y, Liu Y, Wang Z, et al. A comprehensive sstudy of GPT-4V's multimodal capabilities in medical imaging. medRxiv 2023;2023:11:03.23298067. doi:10.1101/2023.11.03.23298067v1.

[94] Hu Y, Li T, Lu Q, et al. OmniMedVQA: a new large-scale comprehensive evaluation benchmark for medical LVLM. In: Proceedings of 2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2024. p. 22170–83. doi:10.1109/cvpr52733.2024.02093.

[95] Gu Z, Yin C, Liu F, et al. MedVH: towards systematic evaluation of hallucination for large vision language models in the medical context. 2024. doi:10.48550/ARXIV.2407.02730.

[96] Wu J, Kim Y, Wu H. Hallucination benchmark in medical visual question answering 2024 In The Second Tiny Papers Track at ICLR.

[97] Oh Y, Park S, Byun HK, et al. LLM-driven multimodal target volume contouring in radiation oncology. Nat Commun 2025;16(1):718. doi:10.1038/s41467-025-55963-2.

[98] Lei N, Cai J, Qian Y, et al. A two-stage Chinese medical video retrieval framework with LLM. In: Proceedings of CCF International Conference on Natural Language Processing and Chinese Computing. Springer; 2023.

[99] Gao S, Fang A, Huang Y, et al. Empowering biomedical discovery with AI agents. Cell 2024;187(22):6125–51. doi:10.1016/j.cell.2024.09.022.

[100] Harrer S, Rane RV, Speight RE. Generative AI agents are transforming biology research: high resolution functional genome annotation for multiscale understanding of life. EBioMedicine 2024;109:105446. doi:10.1016/j.ebiom.2024.105446.

[101] Li B, Yan T, Pan Y, et al. MMedAgent: learning to use medical tools with multimodal agent. In: Proceedings of the Association for Computational Linguistics: EMNLP 2024; 2024. p. 8745–60. doi:10.18653/v1/2024.findings-emnlp.510.

[102] Wang W, Ma Z, Wang Z, et al. A survey of LLM-based agents in medicine: how far are we from Baymax? 2025 arXiv:2502.11211.

[103] Mehandru N, Miao BY, Almaraz ER, et al. Evaluating large language models as agents in the clinic. NPJ Digit Med 2024;7(1):84. doi:10.1038/s41746-024-01083-y.

[104] Liu X, Yu H, Zhang H, et al. AgentBench: evaluating LLMs as agents 2024 ICLR..

[105] Zheng Q, Wu C, Qui P, et al. Can modern LLMs act as agent cores in radiology environments? 2024 arXiv:2412.09529.

[106] Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025:1–9. doi:10.1038/s41586-025-08866-7.

[107] Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med 2025;31(1):77–86. doi:10.1038/s41591-024-03328-5.

[108] Rjoob K, Bond R, Finlay D, et al. Machine learning and the electrocardiogram over two decades: Time series and meta-analysis of the algorithms, evaluation metrics and applications. Artif Intell Med 2022;132:102381. doi:10.1016/j.artmed.2022.102381.

[109] Chen X, Xu P, Li Y, et al. ChatFFA: an ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. iScience 2024;27(7):110021. doi:10.1016/j.isci.2024.110021.

[110] Zhao Z, Zhang W, Chen X, et al. Slit lamp report generation and question answering: development and validation of a multimodal transformer model with large language model integration. J Med Internet Res 2024;26:e54047. doi:10.2196/54047.

[111] Nguyen V, Karimi S, Rybinski M, et al. MedRedQA for medical consumer question answering: dataset, tasks, and neural baselines. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, 1; 2023. Long Papers.

[112] Li Y, Li Z, Zhang K, 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):e40895. doi:10.7759/cureus.40895.

[113] Charnock D, Shepperd S, Needham G, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999;53(2):105–11. doi:10.1136/jech.53.2.105.

[114] Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor–Let the reader and viewer beware. JAMA 1997;277(15):1244–5.

[115] Ozduran E, Büyükçoban S. Evaluating the readability, quality and reliability of online patient education materials on post-covid pain. PeerJ 2022;10:e13686. doi:10.7717/peerj.13686.

[116] Rensis L. A technique for the measurement of attitudes. Arch Psychol 1932;22(140):5–55.

[117] Lahat A, Shachar E, Avidan B, et al. Evaluating the use of large language model in identifying top research questions in gastroenterology. Sci Rep 2023;13(1):4164. doi:10.1038/s41598-023-31412-2.

[118] Chervenak J, Lieman H, Blanco-Breindel M, et al. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. Fertil Steril 2023;120(3 Pt 2):575–83. doi:10.1016/j.fertnstert.2023.05.151.

[119] Wang D, Liang J, Ye J, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: comparative study. J Med Internet Res 2024 Nov 8;26:e58041. doi:10.2196/58041.

[120] Community D. LLM evaluation metrics: ensuring optimal performance & relevance. Deepchecks; 2024.

[121] Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology 2023;308(1):e230970. doi:10.1148/radiol.230970.

[122] Zhu L, Mou W, Hong C, et al. The evaluation of generative AI should include repetition to assess stability. JMIR Mhealth Uhealth 2024;12:e57978. doi:10.2196/57978.

[123] Griot M, Hemptinne C, Vanderdonckt J, et al. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun 2025;16(1):642. doi:10.1038/s41467-024-55628-6.

[124] Xu P, Wu Y, Jin K, et al. DeepSeek-R1 Outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning 2025 arXiv:2502.14739.

[125] Resnikoff S, Lansingh VC, Washburn L, et al. Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? Br J Ophthalmol 2020;104(4):588–92. doi:10.1136/bjophthalmol-2019-314336.

[126] Wan P, Huang Z, Tang W, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. Nat Med 2024;30(10):2878–85. doi:10.1038/s41591-024-03148-7.

[127] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst 2025;43(2):1–55. doi:10.1145/3703155.

[128] Stan GBM, Aflalo E, Rohekar RY, et al. LVLM-Interpret: an interpretability tool for large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024.

[129] Yang SH, Zhao HJ, Zhu GY, et al. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. 2024AAAI 2024;38(17):19368-19376. doi: 10.1609/aaai.v38i17.29907.

[130] Cai Y, Wang LL, Wang Y, et al. MedEvalHub: a large-scale Chinese benchmark for evaluating medical large language models. AAAI 2024;38(16):17709–17. doi:10.1609/aaai.v38i16.29723.

[131] Li J, Guan Z, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care. Nat Med 2024;30(10):2886–96. doi:10.1038/s41591-024-03139-8.

[132] Chiang WL, Zheng L, Sheng Y, et al. Chatbot Arena: an open platform for evaluating LLMs by human preference. In: Proceedings of Forty-first International Conference on Machine Learning; 2024.

[133] Abbasian M, Khatibi E, Azimi I, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med 2024;7(1):82. doi:10.1038/s41746-024-01074-z.