Medicine Plus 1 (2024) 100030

Medicine Plus

Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant

Mingze Yuan[a,1], Peng Bao[a,1], Jiajia Yuan[b,1], Yunhao Shen[b], Zifan Chen[a], Yi Xie[b], Jie Zhao[c,g], Quanzheng Li[h,i], Yang Chen[b,*], Li Zhang[a,f,*], Lin Shen[b,*], Bin Dong[d,e,f,g,*]

[a] *Center for Data Science, Peking University, Beijing 100871, China*

[b] *Department of Gastrointestinal Oncology, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital and Institute, Beijing 100142, China*

[c] *National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871, China*

[d] *Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China*

[e] *Center for Machine Learning Research, Peking University, Beijing 100871, China*

[f] *National Biomedical Imaging Center, Peking University, Beijing 100871, China*

[g] *Peking University Changsha Institute for Computing and Digital Economy, Changsha 410205, China*

[h] *Massachusetts General Hospital, Boston, MA 02114–2696, USA*

[i] *Harvard Medical School, Boston, MA 02115, USA*

ARTICLE INFO

Keywords:
Large language models
Artificial intelligence
Medicine
Healthcare assistant
Prompt engineering
In-context learning

ABSTRACT

With the rapid development of artificial intelligence, large language models (LLMs) have shown promising capabilities in mimicking human-level language comprehension and reasoning. This has sparked significant interest in applying LLMs to enhance various aspects of healthcare, ranging from medical education to clinical decision support. However, medicine involves multifaceted data modalities and nuanced reasoning skills, presenting challenges for integrating LLMs. This review introduces the fundamental applications of general-purpose and specialized LLMs, demonstrating their utilities in knowledge retrieval, research support, clinical workflow automation, and diagnostic assistance. Recognizing the inherent multimodality of medicine, the review emphasizes the multimodal LLMs and discusses their ability to process diverse data types like medical imaging and electronic health records to augment diagnostic accuracy. To address LLMs' limitations regarding personalization and complex clinical reasoning, the review further explores the emerging development of LLM-powered autonomous agents for healthcare. Moreover, it summarizes the evaluation methodologies for assessing LLMs' reliability and safety in medical

* Corresponding authors.
E-mail addresses: yang_chen@bjcancer.org (Y. Chen), zhangli_pku@pku.edu.cn (L. Zhang), shenlin@bjmu.edu.cn (L. Shen), dongbin@math.pku.edu.cn (B. Dong).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.medp.2024.100030
Received 22 February 2024; Received in revised form 1 April 2024; Accepted 14 May 2024
Available online 17 May 2024

contexts. LLMs have transformative potential in medicine; however, there is a pivotal need for continuous optimizations and ethical oversight before these models can be effectively integrated into clinical practice.

1. Introduction

With the onset of the 21st century marked by a staggering growth in artificial intelligence (AI) capabilities, we have witnessed groundbreaking advancements and transformations across various industries.[1–5] Particularly in the medical field, such transformations are even more pronounced.[1,5] However, while AI has unveiled countless new opportunities and possibilities for us, it also sheds light on the profound complexity inherent in medical processes. When considering key stages like diagnosis, treatment, and prognosis, the real-world medical data we grapple with is incredibly diverse and intricate. Doctors, when dealing with this data, not only refer to a vast and complicated body of standard medical knowledge but also need to craft individualized treatment plans based on the unique circumstances of each patient. Furthermore, medical examinations are multimodal, encompassing domains like pathology, radiology, and genomics. Faced with such a scenario, integrating this plethora of data and information to form a coherent and comprehensive diagnosis and treatment strategy is undeniably challenging. Most of the current tools are isolated for single tasks, implying that clinical doctors must engage in more holistic analysis and judgment during decision-making. Hence, there's an urgent need for powerful intelligent assistance tools to aid these physicians. This is precisely what large language models (LLMs), like GPT-4,[6] offer. Not only can they help doctors consolidate and interpret intricate data, but they also provide insights grounded in extensive knowledge,[7] thus ensuring more efficient and precise assistance in pivotal stages like diagnosis, treatment, and prognosis.[1,5,8] With the aid of such intelligent tools, we aspire to delve deeper into a patient's genuine situation and make more apt and accurate medical decisions.

In this context, LLMs such as GPT-4,[6] ChatGPT,[9] and Claude[10] have gradually made their mark in the medical domain. Taking GPT-4 as an example, its exceptional performance in the United States Medical Licensing Examinations (USMLE) has far exceeded the expectations of many experts. Yet, this is only the tip of the iceberg. While the practical application of LLMs in healthcare is still in its early stages, preliminary research has already unveiled their tremendous potential in specialized medical research and potential clinical decision support.[11–17] Especially in tasks involving the integration of multimodal medical data from pathology, radiology, and genomics, LLMs have exhibited their unique ability for in-depth interpretation and linkage. Of course, their practical effects and values in real medical environments still require further study and validation. With the introduction of these advanced tools, we not only anticipate efficient consolidation of multisource medical data but also expect AI agents[18] to offer support in predictive analysis and patient management for physicians. For instance, AI agents could assist in analyzing patient histories, laboratory results, and radiological data, subsequently providing data-driven diagnostic suggestions.[19,20] Moreover, these tools can further help doctors in choosing the optimal treatment plan from a plethora of options, ensuring patients receive individualized and optimal therapeutic outcomes.[11,21] Through this approach, we can look forward to a medical decision-making process that is not only more scientific but also more systematic, ensuring patients receive the best medical care.

Given the outstanding potential of LLMs in medicine, this paper aims to conduct a systematic and progressive review of the recent advances achieved by LLMs in this field, as illustrated in Fig. 1. It highlights the use of both general and specialized medical LLMs, primarily focusing on text-based interactions. Traditional unimodal approaches often overlook the complex multimodal nature of the medical field, prompting the development of multimodal LLMs that enhance diagnostic accuracy and efficacy. Despite notable advancements, challenges such as achieving true personalization, maintaining ongoing model updates, and equipping AI with complex problem-solving abilities remain. In this context, LLM-driven autonomous agents emerge as promising tools with diverse applications in healthcare. Moreover, assessing the efficacy and safety of medical LLMs is crucial. This review's importance lies in shedding light on the current role of LLMs in healthcare, assessing their transformative effects on medical practices, and identifying obstacles to be addressed to fully leverage their potential in enhancing patient care and advancing medical science.

2. Overview of LLMs

2.1. Basic concepts

The emergence of transformer[22] has paved the way for the development of LLMs in the field of natural language processing (NLP), exemplified by two significant LLMs, namely GPT[23] and BERT.[24] These LLMs[6,9,23–26] consist of a vast number of learnable parameters, which can easily scale up to billions. They are pre-trained on a large volume of

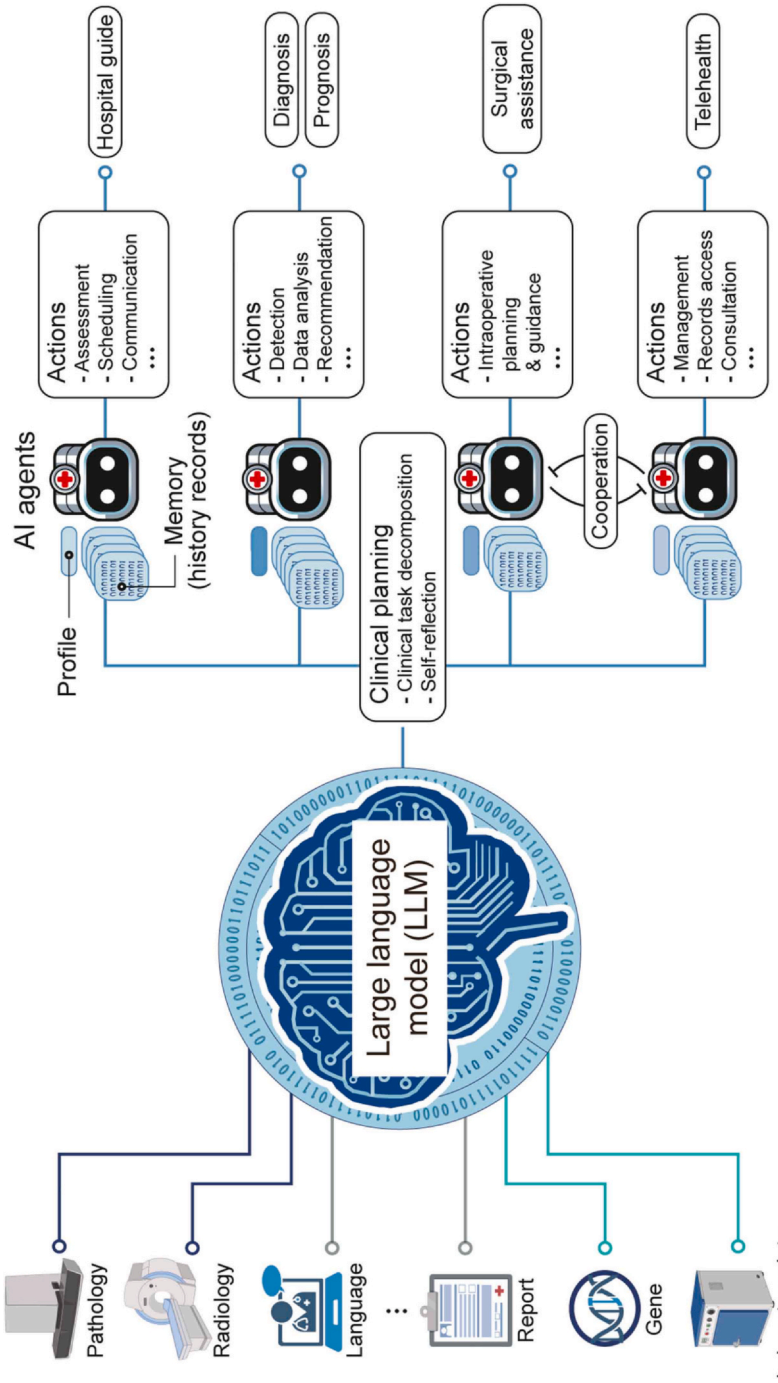

Fig. 1. Integration of an LLM into advanced healthcare support systems. Multimodal medical data, ranging from pathology, and radiology to laboratory sources, funnel into the LLM, symbolized by a digital brain. This LLM interacts seamlessly with AI agents, distinguished by components including profile, planning, memory, and actions. These agents facilitate a range of healthcare procedures, including diagnosis, prognosis, and surgical assistance, underscoring the pivotal role of the LLM in augmenting AI-empowered healthcare systems. AI: artificial intelligence. Part of the figure was created with BioRender.com.

unlabeled corpus using self-supervised learning techniques such as next token prediction[23] and masked language modeling.[24]

Recent advancements in LLMs, exemplified by models like ChatGPT[9] and GPT-4,[6] have showcased outstanding performance as zero-shot or few-shot learners, efficiently summarizing, extracting, and generating text with little to no prompting. The introduction of in-context learning[26] has further elevated their capabilities in this domain. Prompt strategies,[25,27–30] often used in conjunction with few-shot or zero-shot learning, enhance the performance of LLMs across diverse tasks by conditioning them on a limited set of examples. Standard prompting techniques[26] often involve presenting the LLM with a succinct prompt, typically a question or statement, steering the model toward the expected outcome. Techniques like chain-of-thought prompting,[29,31] which guides LLMs through a sequence of logical steps towards a solution, and least-to-most prompting,[28] which systematically breaks down intricate problems into more manageable sub-tasks, exemplify the sophisticated strategies employed to harness and optimize the reasoning prowess of these models.

To enable LLMs to understand natural language instructions and perform real-world tasks, researchers have been exploring methods for instruction-tuning of LLMs.[32] Among these methods, reinforcement learning from human feedback (RLHF)[9] has emerged as a crucial technique for training language models to align with human goals. RLHF has been extensively used to fine-tune state-of-the-art LLMs such as GPT-4,[6] Claude,[10] Bard,[33] and LLaMA-Chat.[34] It consists of three interconnected processes: feedback collection, reward modeling, and policy optimization. Specifically, feedback collection involves obtaining evaluations of model outputs from humans and reward modeling aims to train a reward model that mimics these evaluations via supervised learning. Finally, the policy optimization step fine-tunes the language model to produce outputs that garner positive evaluations from the reward model, ensuring alignment with human preferences. A significant challenge with scaling RLHF is the need for copious high-quality human annotations. However, recent research suggests an innovative solution—reinforcement learning from AI feedback (RLAIF).[35] This approach promises performance at par with human-level evaluations, potentially circumventing the scalability challenges inherent to RLHF.

The evolution of LLMs has given rise to the concept of foundation models, which are trained on expansive datasets and demonstrate versatility across diverse downstream applications.[1,36] Their influence is palpable across various domains, from linguistics[26] and vision[37] to other modalities.[38] Intrinsically linked to foundation models is the idea of a generalist model, wherein a consistent model structure, devoid of fine-tuning, performs commendably across myriad tasks.[1] The aspiration for a unified multitask model[39] that proficiently addresses an array of challenges has persisted over time.[40,41] This ambition is particularly evident in the medical sector,[1] where contemporary LLMs underscore the feasibility of crafting medical generalist frameworks.[42,43]

2.2. Related surveys

Alongside the evolution of LLMs, a plethora of in-depth reviews have surfaced. These offer profound understandings of varied facets, encompassing the background, leading-edge technologies, applications, and the inherent challenges in deploying LLMs.[44–46] Furthermore, pivotal topics such as harmonizing LLMs with human cognition,[47] their inherent reasoning capabilities,[48] instruction tuning approaches,[49] augmentation strategies,[50] and evaluation methodologies have been encapsulated in recent reviews.[51] Additionally, the burgeoning fields of multimodal LLMs[52] and LLM-based agents[18,53,54] have been extensively reviewed.

As we transition into the medical realm, a few studies stand out in their exploration of LLMs and their potential impact. Moor et al.[1] have proposed the concept of a generalist medical AI, though without concrete implementations and empirical evaluations. Rajpurkar and Lungren[55] have provided an insightful review of the evolution, impediments, and prospects of radiological AI models in clinical practice, emphasizing the integral role of LLMs. Qiu et al.[56] have examined the potential impact of expansive AI models, particularly LLMs, in health informatics, pinpointing seven key areas poised for transformation, including molecular biology and drug development. Liu et al.[57] have highlighted the capacity of artificial general intelligence (AGI) to enhance patient care in radiation oncology through the adept analysis of extensive multimodal clinical data. Moreover, a review by Thirunavukarasu et al.[58] evaluates LLMs' strengths and limitations, emphasizing their potential to enhance clinical, educational, and research activities. Other recent studies have comprehensively examined the potential applications and challenges of LLMs in healthcare.[59–64] Notably, He et al.[64] provided an in-depth overview of current specialized LLMs in the healthcare sector, detailing their training data, methodologies, and performance across three benchmarks.

Despite these invaluable insights, the pathway for forging advanced medical AI frameworks harnessing LLMs remains undefined. In the following sections of this survey, we will delve deeper into the manifold applications and implications of LLMs in the medical domain. Specifically, we will explore the broader applications of general-purpose and specialized LLMs in medicine, focus on the context of multimodal LLMs, and provide an in-depth look into LLM-driven autonomous agents in the medical field. Through this cohesive exploration, our intent is to methodically highlight the transformative potential of LLMs in medicine.

3. Unimodal LLMs in medicine

Within the medical field, there is a constant pursuit of AI-driven support systems to enhance healthcare delivery. Prior to the advent of LLMs, Watson for Oncology (WFO) emerged as an AI-assisted decision-making tool, collaboratively developed with leading oncologists.[65] This system underwent an extensive training period spanning over four years, drawing upon the National Comprehensive Cancer Network (NCCN) treatment guidelines and amassing over a century's worth of clinical cancer treatment expertise from the United States. WFO was engineered to propose suitable chemotherapy plans tailored to individual cancer patients. Nevertheless, this initial foray into AI-assisted medicine revealed a profound disconnect between traditional machine learning methodologies and the practical workflow of medical practitioners,[66] at times resulting in recommendations that were unsafe and incorrect.[67]

LLMs, on the other hand, have showcased an enhanced capability for logical reasoning and application of knowledge.[29,68,69] This section focuses on the application of basic LLMs in medicine that rely solely on textual information for interaction, which encompasses two main avenues: the first explores the direct application of general-purpose LLMs to medical contexts, whereas the second pursues the development of a specialized medical LLM.

3.1. Applying general-purpose LLMs to medicine

The ascendancy of general-purpose LLMs has sparked significant interest in the medical field.[6,9,10,33,34] Recent literature has provided a comprehensive review of ChatGPT's applications within healthcare and clinical practice.[59–61] To gauge the capability of these models, researchers frequently resort to benchmark question-answering datasets spanning various medical disciplines, utilizing metrics such as accuracy, recall, and F1 scores for assessment. OpenAI's pivotal study stands out, showcasing GPT-4's commendable performance on academic and professional tests tailored for an erudite audience.[6] The results pointed to GPT-4's distinguished aptitude in subjects like the Uniform Bar Exam and GRE. Furthermore, Microsoft's independent analysis placed GPT-4 above the USMLE,[70] an exhaustive medical residents' examination, marking a notable improvement from its predecessor, ChatGPT, which merely matched college-level performance on the USMLE.[71,72] This progression epitomizes the brisk evolution of LLMs in medical settings.

Subsequent research accentuates the adaptability of general-purpose LLMs across diverse medical subspecialties, ranging from oncology[11,12,73] to emergency medicine,[13] medical esthetics,[14] radiology,[15] ophthalmology,[74,75] surgery,[76] nursing[16] to complex medical cases.[77] These inquiries typically gauge an LLM's domain-specific expertise using carefully curated questions. For example, Hu et al.[75] evaluated GPT-4 by progressively introducing information on select ophthalmic conditions, simulating patient and physician interactions. Experienced ophthalmologists subsequently assessed the model's outputs, underscoring GPT-4's potential utility in both patient referrals and medical training. Additionally, Brin et al.[78] scrutinized both ChatGPT and GPT-4's capabilities in handling USMLE questions centered on communication nuances, ethics, and empathy, finding a noteworthy capacity for empathy and professionalism in AI.

In the realm of knowledge retrieval and dissemination, LLMs emerge as potent instruments, serving not only as invaluable repositories of medical information but also as sophisticated educators. These models afford healthcare professionals immediate access to contemporaneous medical data by meticulously analyzing a plethora of scientific journals, research articles, and clinical protocols. This analysis furnishes pertinent and timely details[60,79] pertaining to disease processes,[80,81] therapeutic approaches,[21,82,83] and drug interactions.[84,85] Such insights can be especially valuable in assisting with the diagnosis of rare diseases,[86] which often presents challenges for clinicians. Furthermore, LLMs have the dexterity to democratize medical knowledge through online medical consultation,[7,12,14,86–88] ensuring widespread availability while simultaneously offering customization to cater to individual prerequisites, potentially impacting telemedicine.[61,89]

The integration of LLMs in medical research and writing, as highlighted in various studies,[17,59,60,90–92] significantly enhances the efficiency, equity, and applicability of research endeavors. These models streamline experimental design, ensure the preservation of patient confidentiality through effective anonymization of medical records, and augment the available medical text data for training purposes. Notably, they facilitate the swift collection, processing, and sophisticated analysis of disease-specific data, fostering more comprehensive and insightful research initiatives. Clinical trials, an essential component of medical research, benefit immensely from LLMs, as they address challenges related to patient-trial matching and trial planning.[91,93–97] A detailed review by Ghim et al.[91] delves into the transformative potential of LLMs within clinical trials, identifying five key areas for imminent implementation: improved patient-trial matching, streamlined clinical planning, advanced free text narrative analysis for coding and classification, assistance in technical trial planning, and the facilitation of informed consent through LLM-powered chatbots. In particular, Jin et al.[95] demonstrated the capability of LLMs, through their proposed TrialGPT system, to aid patients

and referral physicians in selecting appropriate clinical trials from a vast array, validating the explanatory prowess and invaluable contribution of LLMs to medical research on three public cohorts encompassing 184 patients and 18,238 annotated clinical trials.

In the context of clinical workflow, LLMs can significantly mitigate the substantial burden shouldered by healthcare professionals by autonomizing the documentation of patient information, clinical observations, and test reports.[17] This automation does more than merely streamline the process; it enhances both the accuracy and the thoroughness of the clinical documentation. For example, LLMs have been efficaciously utilized to summarize radiology reports,[98] providing a prototype for analogous applications in various domains,[11] including using ChatGPT to write patient clinic letters.[99] Besides, the deployment of LLMs in clinical decision support is markedly beneficial, offering insightful recommendations pertaining to medication regimens, suggesting suitable imaging services grounded in clinical presentations, and enabling the astute diagnosis of diseases from comprehensive clinical data sets.[15] When synergistically integrated with other diagnostic instruments, such as medical imaging tools, LLMs proffer a more holistic perspective of patient health. Moreover, by analytically examining data from analogous cases, LLMs can prognosticate patient outcomes, thus assisting both healthcare professionals and patients in navigating toward enlightened treatment decisions.

Beyond such application-oriented evaluations, comparative studies have explored methodological optimizations in LLMs for medical applications, underscoring the critical role of task-adapted prompting for optimal performance. Wang et al.[100] assessed the prowess of state-of-the-art LLMs, including GPT-3.5,[9] GPT-4,[6] and Bard,[33] across an array of clinical linguistic tasks, leveraging diverse learning and prompting techniques.[26,29,31] Parallel to this, Yuan et al.[21] spotlighted GPT-4's proficiency in intricate clinical assignments, particularly in tumor treatment planning, by employing sophisticated prompting methods. Liu et al.[101] shed light on LLMs' efficiency in radiology, underlining the variance in results depending on specific shot configurations. Moreover, Tang et al.[102] critiqued zero-shot LLMs' abilities in condensing medical evidence, revealing occasional discrepancies in their summaries, thus highlighting the ongoing need for rigorous monitoring and perpetual enhancements.

The potential of general-purpose LLMs is undoubtedly transformative. Yet, when it comes to their integration into the medical sector, numerous challenges arise. Specific tasks and limited question domains can introduce selection bias.[75] Furthermore, the model's efficacy often hinges on the design of the prompt, which might not always be intuitive for users.[21] Notable commercial LLMs, such as GPT-4 and Claude-2, encounter difficulties when being integrated into clinical workflows. These difficulties range from ensuring Health Insurance Portability and Accountability Act (HIPAA) compliance and safeguarding patient privacy to obtaining necessary institutional review board (IRB) approvals, particularly when there's a possibility of patient data being transferred to external servers. From an ethical standpoint, it's worth noting that commercial LLMs like ChatGPT might be restricted from delivering medical diagnoses or drug recommendations.[103] Unlike physicians who delve deeper into patients' complaints, models like ChatGPT might provide more generic answers.[103] Moreover, they struggle to incorporate the latest health insights, leading to potentially outdated responses. A crucial observation is that some models exhibit deficiencies in specialized medical knowledge, a point emphasized by Antaki et al.[104] One reason for this might be that these LLMs primarily learn from clinical guidelines and research papers—sources that usually reflect controlled environments rather than the nuanced realities of everyday clinical practice. These challenges have intensified the urgent need for the development of LLMs tailored to medical contexts and specific diseases, in order to ensure more accurate, personalized, and effective patient care solutions.[105,106]

3.2. Developing specialized medical LLMs

Specialized medical LLMs are meticulously crafted to address the intricate needs of the healthcare sector, with endeavors centered on either developing entirely new LLMs pre-trained with a healthcare-centric focus or refining existing models to bolster their efficacy within medical contexts.[69,103,107–116] The advent of this category is rooted in research findings that highlight the limitations of general-domain LLMs, particularly when tasked with healthcare-specific challenges, as they tend to grapple with domain shifts.[70,75,95,104] These findings also suggest that merely depending on prompt engineering might not yield substantial enhancements in their performance regarding healthcare-specific applications.[21]

One prevalent approach to constructing these specialized medical LLMs involves fine-tuning a base LLM on medical dialog or datasets tailored to specific instructions. A primary concern with generic LLMs in the realm of medicine is the potential discord between their initial training goals and end-user expectations.[103] While LLMs traditionally aim to minimize prediction errors across diverse datasets, users desire models that deliver accurate and safe results. Instruction tuning emerges as a solution, refining LLMs with paired sets of user directives and expected outcomes to better align with users' anticipations.[32]

3.2.1. Modern medicine

A noteworthy breakthrough in the field is Google's Med-PaLM.[108] This system stands out as the first AI to excel in USMLE-style inquiries, consistently delivering insightful responses to general health queries. To ensure comprehensive evaluations, the researchers introduce MultiMedQA, an all-encompassing benchmark that amalgamates six previous medical QA datasets across different specialties. Furthermore, they introduce a novel dataset, HealthSearchQA. Building on this foundation, Flan-PaLM is derived through instruction tuning on PaLM[117] and subsequently evaluated using MultiMedQA. Impressively, Flan-PaLM outperforms the prior leading model by a margin of 17% in accuracy. This achievement underscores the value of leveraging real-world medical data. Additionally, human evaluators identify key areas for enhancement. In response to this feedback, researchers develop the "instruction prompt tuning" methodology. This innovation paves the way for the launch of Med-PaLM 2 in March 2023,[69] which boasts an admirable accuracy rate of 86.5% for USMLE-style questions.

However, high-performing models like GPT-4[6] and Med-PaLM 2[69] remain proprietary, hampering their private use in this sensitive field. To circumvent this, efforts have been directed at refining open-source LLMs,[101] exemplified by Meta's LLaMA.[34] ChatDoctor stands out in this respect, having fine-tuned LLaMA using extensive patient-doctor dialogs from a popular online medical platform.[109] It further enhances the model with a self-guided information retrieval feature, enabling access to real-time online resources and trusted offline medical databases. This approach markedly amplifies the model's capability to discern patient requirements and offer reliable counsel. Similarly, Radiology-GPT utilizes the Alpaca instruction-tuning framework[118] to create a radiology-centric LLM.[115] The model is tailored to generate diagnostic "Impression" narratives based on provided "Findings" data. While Radiology-GPT performs comparably to ChatGPT in understandability and even surpasses it in coherence, it lags slightly in relevance. Importantly, this underscores the potential for crafting domain-specific LLMs that cater to distinct medical niches, while upholding rigorous privacy standards.

When applied to unstructured textual data, such as electronic health records (EHR), a unique strength of specialized LLMs emerges. Specialized LLMs often outperform traditional structured predictive models due to their inherent flexibility. Jiang et al.[116] demonstrate that unstructured clinical notes from EHRs can facilitate the training of clinical language models. These models can then serve as versatile clinical predictive engines, allowing for streamlined development and deployment. Specifically, their approach, a specialized LLM named NYUTron, is pre-trained on a decade's worth of inpatient clinical notes. It is then fine-tuned for a broad spectrum of clinical and operational predictive tasks, delivering significant enhancements over conventional methods.

The evolution towards disease-specific LLMs stands as a crucial advancement for elevating patient care. Current medical LLMs, while beneficial, often fall short in delivering precise and reliable management recommendations for patients with specific conditions such as diabetes and cardiovascular diseases. Recognizing this gap, it becomes imperative to focus on the future development of LLMs tailored to these and other diseases, ensuring more accurate, personalized, and effective patient care solutions.

3.2.2. Traditional medicine

In addition to modern medicine, it is essential to recognize the importance of traditional medicine as a field worthy of attention. While the majority of LLMs are tailored for modern medical applications, there is a growing effort to adapt LLMs to the nuances of traditional medicine. Notable specialized LLMs have been developed in this area, such as ShenNong-TCM,[119] Huangdi,[120] Sunsimiao,[121] CMLM-Zhongjing,[122] MedChatZH,[123] TCM-GPT,[124] and Qihuangwendao. These models are generally developed by fine-tuning existing open-source LLMs, such as LLaMA,[34] with a carefully selected corpus encompassing classical texts, ancient manuscripts, and contemporary clinical case studies in traditional medicine.

The advent of these LLMs in the field of traditional medicine represents a significant milestone, enhancing the accessibility and understanding of its corpus of knowledge. They play a crucial role in facilitating more effective diagnostic procedures and treatment approaches that are consistent with traditional medicinal practices. Moreover, marrying the insights of traditional medical knowledge with the advances in modern medicine could lead to innovative breakthroughs in comprehensive patient care and holistic treatment approaches. Such progress addresses a critical void within the landscape of traditional medicine LLMs, where models designed primarily for modern medical contexts may fall short. However, the growth of traditional medicine LLMs is not without its challenges. A primary concern is the relatively limited size of datasets available for training these models. Additionally, most open-source LLMs used as a starting point are trained predominantly on English-language data, which can lead to a lack of essential insights and the complex logic inherent in traditional medicine knowledge systems.

3.3. Debate on general-purpose or specialized LLMs in medicine

The ongoing debate about whether to apply general-purpose or specialized LLMs in medicine warrants re-examination. This is particularly true for the common belief that models like GPT-4 cannot rival specialized models in terms of intensive domain-specific knowledge. Typically, research focusing on medical competency benchmarks has

prioritized domain-specific training, as illustrated by the Med-PaLM project.[69,108] However, this domain-specific fine-tuning might compromise a model's general reasoning abilities, potentially impacting its effectiveness across the entire medical process.[42] Recent developments, noted by researchers like Yuan et al.[21] and Nori et al.[125], suggest that leveraging general-purpose LLMs in medical decision-making through prompt engineering could be a promising approach. Systematic studies have explored the use of prompt engineering to enhance the performance of general-purpose models on medical benchmarks.[21,125] A notable finding from Medprompt reveals that creative prompting strategies can significantly enhance a model's specialist capabilities.[125] For instance, GPT-4, when guided by Medprompt, surpasses previous benchmarks in medical question-answering tasks and outperforms previous leading specialized models like Med-PaLM 2 by a substantial margin. However, the fine-tuning of foundation models remains an important area of research. It could offer synergistic benefits when combined with prompt engineering,[108] presenting an efficient and privacy-conscious alternative to commercial LLMs. Both fine-tuning and prompt engineering methodologies should be carefully explored to fully unlock the potential of foundation models in high-stake domains such as healthcare (Table 1).

4. Multimodal LLMs in medicine

Medicine inherently involves multiple data modalities, including text, images, genomics, and more. A recent review on multimodal biomedical AI has explored the opportunities for multimodal datasets in healthcare, and then discussed the key challenges and promising strategies for overcoming them.[126] However, unimodal LLMs still lack the ability to perceive visual modalities such as magnetic resonance imaging (MRI) and handle complex unstructured data (e.g., gene screening status), which limits their utilization for real-world medical scenarios.

This section emphasizes multimodal LLMs (MLLMs) in medicine,[42,43,127–140] which integrate various modalities into LLMs for diagnostic functions, as listed in Table 2. Here, we specifically examine the taxonomy of utilized data modality, primarily including imaging and intricate unstructured data,[137–140] and proceed with their methodologies for data collection and modality fusion.

Table 1
A glossary of LLM-related terms.

Terminology	Definition
LLMs	Language models with a large number of parameters, capable of performing a wide array of language tasks.
General-purpose LLMs	LLMs designed to handle a wide variety of tasks without task-specific optimization.
Specialized LLMs	LLMs optimized or fine-tuned for a specific task or domain.
Multimodal LLMs	LLMs capable of understanding and generating content across multiple modalities such as text and images.
AI agents	Advanced systems that act autonomously to carry out tasks or make decisions based on data or environment.
Reinforcement learning with human feedback (RLHF)	A training approach where human feedback is used to optimize the model's predictions or actions.
Prompt engineering	The practice of crafting and optimizing prompts to effectively instruct a language model.
Zero-shot learning	The ability of a model to generalize to unseen tasks or classes without needing explicit examples during training.
Few-shot learning	The ability of a model to adapt to new tasks or classes with very limited examples.
In-context learning (ICL)	Learning or adapting to new tasks by leveraging context or examples provided during inference.
Fine-tuning	A process of further training a pre-trained model on a specific task to improve its performance.
Instruction tuning	An approach where LLMs undergo additional training using a dataset of instruction-output pairs via supervised learning.
Chain-of-thought prompting	Crafting prompts in a way that guides the model through a multi-step reasoning process.
Reinforcement learning from AI feedback (RLAIF)	A reinforcement learning approach where feedback from another AI model is used to guide learning.
Hallucination	The phenomenon where LLMs may fabricate inconsistent or outright false information.

This glossary provides concise definitions of crucial terms related to LLMs and their applications. It encompasses various types of LLMs, training approaches, learning paradigms, as well as techniques used to optimize and instruct these models. AI: artificial intelligence; LLM: large language model.

Table 2
A summary of multimodal LLMs in medicine.

Model	Modality	Task	Base model	Sample size	Data source
PathAsst[127]	Pathology	Pathological diagnosis	PLIP,[141] Vicuna-13B[142]	1.42×10^5	Open-source (PubMed, books, websites), in-house
BiomedGPT[128]	Radiology, pathology	VQA, image captioning	OFA[143]	$\sim 1.84 \times 10^8$	Open-source datasets
PMC-VQA[129]	Radiology, pathology, microscopy, etc.	VQA	PMC-CLIP,[144] PMC-LLaMA[112]	2.27×10^5	Open-source (PubMed), self-instruction
LLaVA-Med[130]	Radiology, pathology	VQA	LLaVA,[145] CLIP[146]	6×10^5	Open-source (PubMed[147]), self-instruction
XrayGPT[131]	X-ray	VQA	MedCLIP,[148] Vicuna[142]	2.17×10^5	Open-source datasets (MIMIC-CXR[149], OpenI[150])
CephGPT-4[132]	Dental imaging	Orthodontic measurement & diagnostic	MiniGPT-4,[151] VisualGLM[152]	$\sim 6 \times 10^4$	Patient dialogs, real clinical case samples
Med-PaLM M[42]	Radiology, pathology, mammography, genomics, dermatology, etc.	QA, VQA, report summarization & generation, genomic variant calling, etc.	PaLM-E[153]	$>1 \times 10^6$	Open-source (12 de-identified datasets)
Med-Flamingo[133]	Radiology, pathology, etc.	VQA	OpenFlamingo[154]	1.3×10^6	Open-source (textbooks, PubMed[144])
RadFM[43]	Radiology	VQA, disease diagnosis, report generation	ViT,[155] PMC-LLaMA[112]	1.6×10^7	Open-source (a collection of existing datasets), self-instruction
BioMedGPT[137]	Molecule, protein	QA (medical, molecule, protein)	LLaMA2-7B-Chat[34]	$\sim 2.3 \times 10^6$	Open-source (literature[156], PubChem)
HeLM[138]	Individual-specific (e.g., lab values)	Disease risk estimation	PaLM-E[153]	$\sim 1.7 \times 10^4$	Open-source (UK Biobank)

This table does not include the multitude of open-source projects available on Github.[101] The "Sample size" column denotes the number of training samples, such as image-text pairs. QA: question answering; VQA: visual question answering; LLM: large language model.

4.1. A taxonomy of data modality usage

In the field of medicine, multimodal LLMs can be classified based on the data modality they utilize. They primarily fall into two main categories: imaging, which is the most prominent, and other complex unstructured data types, such as genomic sequences, time-series data, and audio recordings.

4.1.1. Imaging

Existing research on multimodal LLMs in medicine, as evidenced by numerous studies,[42,43,127–130,132,133,157] predominantly focuses on exploiting imaging data. The overarching goal is to devise a universal and adaptive model compatible with various imaging modalities and assignments. BiomedGPT exemplifies this by offering a versatile AI model for medical applications,[128] which integrates diverse modalities, from CT images to clinical notes. Uniquely, BiomedGPT encapsulates information from various input sources into a shared multimodal lexicon suitable for many tasks. It uniformly employs a sequence-to-sequence paradigm throughout both the pretraining and finetuning phases. Furthermore, task directives are seamlessly integrated into inputs as plain text, eliminating the need for supplemental parameters. After rigorous testing on multiple biomedical datasets and tasks, BiomedGPT not only demonstrates its ability to effectively disseminate knowledge across tasks but also matches or outperforms dedicated models optimized for specific datasets or modalities. Its strength is most apparent in vision-language assignments like image captioning and visual question answering, where it sets new benchmarks in performance.

The Med-PaLM Multimodal (Med-PaLM M) further exemplifies a cohesive model tailored to interpret a spectrum of biomedical data types, managing diverse tasks using a consistent set of model weights.[42] Addressing the lack of comprehensive multimodal medical benchmarks, it introduced MultiMedBench, an inclusive open-source multimodal medical benchmark. This benchmark covers language, medical imaging, and genomics, encompassing a wide range of tasks. These include question answering, visual question answering, medical image categorization, radiology report creation and summarization, and genomic variant identification. With this foundation, Med-PaLM M introduces a versatile multimodal sequence-to-sequence architecture capable of smoothly integrating diverse biomedical data. The model's universal language decoder provides inherent flexibility, enabling it to handle a variety of biomedical tasks within a unified generative framework. Impressively, even without task-specific fine-tuning, Med-PaLM M matches or surpasses dedicated models across several MultiMedBench tasks. Beyond just performance metrics, the model demonstrates intuitive medical reasoning, adaptability to novel concepts and responsibilities, and effective knowledge transfer. This underscores its vast potential, especially in areas with limited biomedical data.

RadFM serves as a foundational model for radiology.[43] It curates an extensive multimodal dataset, MedMD, boasting roughly 1.6×10^7 medical scans. This dataset includes 1.55×10^7 two dimensional (2D) scans and 1.8×10^5 three dimensional (3D) radiological images, each accompanied by textual narratives, such as radiology reports, visual-language instructions, or vital diagnostic labels. Distinctively, RadFM operates as a text-generation model conditions on visual inputs, adeptly merging natural language with 2D or 3D medical imagery. Its output is primarily in the form of natural language, catering to a diverse set of medical tasks. Additionally, RadFM presents a comprehensive radiology benchmark, capturing a spectrum of clinical duties like disease identification, report drafting, and visual question answering across various radiological modalities and anatomical sectors. Also within the field of radiology, ELIXR employs a language-aligned image encoder and skillfully integrates it with a stable LLM,[157] specifically PaLM 2,[158] enabling it to handle a diverse set of tasks. This lightweight adapter architecture is trained on images paired with their corresponding free-text radiology reports, sourced from the MIMIC-CXR dataset.[149] This configuration underscores the potential of LLM-aligned multimodal models, demonstrating how the combination of chest X-rays with relevant radiology reports can address numerous medical tasks, including visual question answering and radiology report quality assessment.

The recent advancement in the GPT-4 series, GPT-4V,[159] has introduced support for multimodal inputs, garnering immediate attention due to its potential effectiveness.[135,160–163] Wu et al.[135] conduct an in-depth evaluation of GPT-4V's performance in multimodal medical diagnostics, encompassing 17 human body systems and employing images from 8 different modalities common in daily clinical practice. The researchers scrutinize GPT-4V's capacity to handle a variety of clinical tasks, assessing its proficiency both with and without patient history, and spanning activities such as imaging modality and anatomy recognition, disease diagnosis, report generation, and disease localization. While the model excels at distinguishing between medical modalities and identifying anatomical structures, it faces challenges in disease diagnosis and producing detailed medical reports. This study underscores that, despite considerable progress in computer vision and NLP within large multimodal models, there remains a considerable gap before these tools can be effectively integrated into real-world medical applications and clinical decision-making. However, it is crucial to acknowledge the limitations of this study, as real clinical settings primarily use 3D DICOM formatted radiological images, whereas GPT-4V can process only up to four 2D images simultaneously, necessitating the selection of 2D key slices or small patches for pathology.

4.1.2. Other modalities

In medical care, clinicians frequently analyze a variety of data types, not limited to medical imaging, but also including clinical notes, lab tests, vital signs, genomics, and other observational metrics. Therefore, effectively deciphering this vast, unstructured data is essential for the integration of multimodal LLMs in healthcare.[137,138,140] A recent perspective by Moor et al.[1] underscores the potential of foundational LLMs, which not only incorporate extensive medical knowledge but also adeptly handle intricate unstructured data.

4.1.2.1. Genomics. Inspired by the transfer learning paradigm of LLMs, Theodoris et al.[139] propose Geneformer. This model is pre-trained on a substantial corpus of approximately 30 million single-cell transcriptomes, allowing for context-specific predictions in network biology scenarios, especially when data is limited.

Audio In a recent review by Huang et al.,[140] the potential applications of multimodal LLMs in dentistry are explored. The authors delineate two primary deployment methodologies: automated dental diagnosis and cross-modal dental diagnosis, elaborating on their prospective utilities. Remarkably, an LLM equipped with a cross-modal encoder can process multi-source data and leverage advanced natural language reasoning for intricate clinical tasks. Beyond the realm of vision-language integration, they underscore the significance of a patient's voice in medical diagnoses, in conjunction with imaging and dialogs. They illustrate how waveforms and spectrograms from distinct patients could be ingested by pre-trained LLMs like GPT-4 to diagnose potential ailments and gauge their severity. Here, audio data serves a dual purpose: detecting voice anomalies and comprehending patient narratives. In voice anomaly detection, the system captures patient voice inputs, generates waveforms and spectrograms, and subsequently conducts amplitude and frequency analyses. For narrative understanding, patient accounts are transcribed into text via speech recognition technologies. Essential information, such as described symptoms, can then be distilled and organized into concise reports or bullet points for clinician reference.

4.1.2.2. Tabular data. Furthermore, HeLM demonstrates the value of multimodal LLMs in delivering personalized healthcare.[138] Designed specifically to process high-dimensional clinical data for disease risk assessment, HeLM employs specialized encoders to convert varied data into the LLM's token embedding space, while simpler tabular data is serialized into textual formats. Impressively, HeLM seamlessly merges both demographic and clinical data, including detailed time-series data, to predict disease risks. Moreover, its exceptional performance in zero-shot and few-shot learning for certain conditions reaffirms the immense foundational knowledge that LLMs can contribute to healthcare.

Notably, though the aforementioned multimodal LLMs show promise in handling multimodal data processing and personal user data, serval challenges remain to be addressed. Initially, a substantial volume of multimodal data, which is currently scarce in healthcare, is essential for training these models. There is also a pressing need for research on converting various modalities into aligned embeddings.

4.2. Core methodologies

The transition from language-exclusive LLMs to multimodal LLMs necessitates significant adaptations in how datasets are structured and how models are architecturally designed to handle multimodal inputs.[164] Therefore, techniques for multimodal data acquisition and modality fusion are important.

4.2.1. Multimodal data acquisition

Crafting robust multimodal medical datasets for training MLLMs is a meticulous task that has seen various concerted efforts.[127,130] In this context, we spotlight the dominant techniques for data collection, which mainly hinge on modifying established benchmark datasets and employing the innovative method of self-instruction.

4.2.1.1. Adapting established datasets. Although existing medical datasets and benchmarks offer a rich source of high-quality data,[165–169] they often require reconfiguration to be suitable for multimodal training. Consequently, many studies have undertaken the task of repurposing these datasets into instruction-oriented configurations.[42,127,128,131,134] One noteworthy example is PathAsst,[127] which introduces the PathCap dataset, comprised of image-caption pairs primarily curated from a range of reputable sources, including the PubMed database, medical textbooks, and pathology atlases. A pre-trained classifier sorts out pathological data from PubMed. Subsequently, sub-figures and sub-captions are isolated and refined using ChatGPT, leading to a dataset well-suited for multimodal instruction tuning. These image-caption pairs, along with designed instructions, naturally constitute multimodal inputs and responses, where instructions are selected from a pre-established pool. Concurrently, some research initiatives have opted to design a foundational set of instructions,[130] expanding them with the aid of GPT-4 for greater diversity and specificity. This practice ensures that the repurposed datasets have instructional value and can be effectively utilized for multimodal learning.

4.2.1.2. Self-instruction. Real-world complexities are sometimes inadequately represented in pre-established datasets. While established datasets are invaluable, they often misalign with real-world scenarios, especially in intricate contexts such as multi-turn conversations. To address this discrepancy, several studies[43,112,127,130] have adopted a self-instruction approach,[170] where LLMs autonomously create instructional text by extrapolating from an initial collection of manually annotated reference examples, termed "seed samples". The self-instruction method allows LLMs to autonomously generate data that mirrors more realistic, conversational interactions. For instance, LLaVA-Med leverages GPT-4 to curate instruction-following data with multi-turn conversations around biomedical images. Given an image caption, GPT-4 generates questions and answers, simulating a conversation as if the model could view the image. This interaction is enriched by integrating sentences from the related PubMed articles and by incorporating carefully curated seed examples to guide high-quality conversation generation based on the caption and its context.[170] Similarly, PathAsst prompts GPT-4 to produce conversation-based instruction-following data primarily from image captions.[127] This self-generated content proves highly beneficial, enabling MLLMs to train on diverse, contextual examples that mirror actual medical dialog.

4.2.2. Modality fusion

The integration of multimodal data into LLMs is a pivotal step in advancing their capabilities. Two principal strategies have been developed: translating non-textual modalities into natural language understanding using expert models, and directly infusing multimodal data into the model's embedding space.

4.2.2.1. Leveraging expert models. One common approach utilizes expert models, such as image captioning systems, to translate visual data or other non-textual data into textual content.[127,171] Rather than directly processing multimodal inputs, LLMs interpret the transformed textual representations. Visual Med-Alpaca[171] exemplifies this strategy, utilizing multiple medical visual expert systems, Med-GIT[172] and De-Plot,[173] in tandem with an LLM. A classifier discerns which specific captioning model should handle the image data, and the generated text is then integrated with the original textual query, enabling the LLM to craft a pertinent response. However, a notable downside is that such transformations could result in information loss. For instance, the granularity of spatial data in visual content might get oversimplified in the conversion to text, and HeLM has demonstrated that direct text serialization of tabular data does not yield a representation that fully captures the available signal for a complex set of features.[138]

4.2.2.2. Continuous injection. Various studies have aimed to seamlessly merge multimodal data by continuously injecting it into the embedding space of pre-trained LLMs.[42,131,133,137,138] For instance, Med-PaLM M[42] integrates visual information using the ViT encoder[155] with PaLM,[117] in a manner that sees the continuous integration of visual data. This creates multimodal sentences, where textual data is interspersed with visual embeddings. A representation of such a sentence might look like: Q: What happened between < img_1 > and < img_2 > ?, where < img_i > symbolizes an image's embedding. This approach bypasses the discrete token level, allowing for direct mapping of visual observations into the linguistic embedding space. HeLM employs a similar method, converting diverse data types into token embeddings.[138] Nonetheless, a notable challenge with HeLM is the degradation of conversational competence after fine-tuning, a phenomenon observed in other models as well.[174]

In more complex scenarios, like with RadFM that deals with 3D images such as MRI or computed tomography (CT) scans, the resulting token sequence from visual encoding can be quite lengthy.[43] Here, a perceiver module is used to compactly represent visual data.[175] Leveraging this architecture, diverse image sizes can be uniformly represented, facilitating easier fusion. Similarly, Med-Flamingo uses this perceiver module to efficiently bridge vision encoders and LLMs, translating varying numbers of visual features into a fixed set of outputs, thereby optimizing computational efficiency.[133] Contrasting with the initial technique that leverages expert models, this approach enables a more profound and seamless integration of multimodal content such as text, images, audio, and other data types. Consequently, it offers a greater potential for performance enhancement and has become the preferred method for cutting-edge multimodal LLMs.

5. LLM-powered autonomous agents in medicine

Although LLMs such as ChatGPT, GPT-4,[6] and Med-PaLM M[42] have made strides in the medical domain, they primarily focus on conversational elements and basic information retrieval. In addition, specialized multimodal LLMs demand a vast amount of multimodal data for training, which is scarce in healthcare. Consequently, these models tend to be task-specific, and their conversational capabilities are limited to their training topics. They are not yet fully equipped to serve as comprehensive healthcare agents due to hurdles in personalization, updating knowledge, and engaging in autonomous sequential thinking, strategic planning, and complex problem-solving, all of which are imperative for physicians in clinical practice.[19] The development of LLM-driven autonomous agents adept at navigating clinical complexities warrants exploration.

Table 3
A summary of LLM-powered autonomous agents in the field of medicine, detailing their specific tasks, planning mechanisms, memory components, and usage of external tools.

Model	Task	Planning	Memory	External tool usage
AD-AutoGPT[176]	Focus on Alzheimer's Disease in medical research	Utilizes GPT-4[6] for task decomposition	Stores environmental information and past actions	Employs tools for news searching, summarization, and result visualization
CHA[19]	Personalized healthcare conversations, including stress level assessment	Incorporates ReAct[177] for task planning	Retains environmental data and historical actions	Utilizes Google Search, translator, healthcare platforms, and stress estimation models
ImpressionGPT[178]	Summarization of radiology reports	Engages in self-reflection through iterative prompting optimization	Uses similar existing reports as contextual examples	N/A
PharmacyGPT[179]	Prescription of medication plans	Applies self-reflection through iterative prompting optimization	Employs patient clustering to provide relevant examples	N/A
ChatCAD + [20]	Assists in computer-aided diagnosis	Utilizes dynamic prompting for self-reflection	Gathers related information from professional sources	N/A
PathAsst[127]	Assistance in pathological diagnosis	N/A	Retrieves papers from a local knowledge database	Uses expert models for tasks like image segmentation and detection in pathology
Yuan et al.[21]	Gastrointestinal cancer management	Decomposes the task via templated prompt	Retrieves similar cases from storage of knowledge	N/A

LLM: large language model; N/A: not available.

An overview of LLM-powered autonomous agents, outlining their key components, is presented in Table 3. Studies in this area are classified into two primary categories: those focused on developing holistic AI agents for medical applications and those aimed at enhancing individual functionalities of AI agents in healthcare.

5.1. Autonomous agents

Autonomous agents are advanced systems capable of independently accomplishing tasks through self-directed planning and instructions.[180] They have emerged as a promising solution for achieving AGI. Traditional approaches have focused on training agents with limited knowledge in isolated environments, failing to achieve human-level proficiency in open-domain settings.[181] However, recent breakthroughs in LLMs have exhibited remarkable reasoning capabilities,[29] leading to a growing trend in leveraging LLMs as central controllers to empower autonomous agents. This trend holds the potential to develop general problem solvers, as demonstrated by proof-of-concept demos like AutoGPT.[182] For a detailed analysis, readers are referred to the comprehensive survey on LLM-based autonomous agents.[18,53,54]

In this framework, the LLM acts as the cognitive core of the system, which is further reinforced with various essential capabilities to effectively execute diverse tasks.[54] These capabilities are fulfilled through multiple modules: profile, memory, planning, and action.[53] Specifically, the profile module aims to determine the role profiles of agents, such as radiologists or programmers, which are typically integrated into the prompt to influence and restrict the behaviors of the LLM.[183,184] The memory module stores environmental information and employs recorded memories to facilitate future actions.[185] This enables the agent to accumulate experiences, self-evolve, and exhibit more consistent, logical, and effective behavior.[186] The planning module empowers the agent to decompose complex tasks into simpler subtasks and solve them sequentially.[28–30] Additionally, it enables the agent to engage in self-criticism and self-reflection,[177,187] learning from past actions and refining themselves to enhance future performance. The action module translates the agent's decisions into specific outcomes by directly interacting with the environment. The action capability is enriched by the agent's skill in utilizing diverse external tools or knowledge sources,[50,188] such as application programming interfaces (APIs), knowledge bases, and specialized models. These modules are interconnected to establish an LLM-based autonomous agent, where the profiling module influences the memory and planning modules, which, together with the profile, collectively impact the action module.[18]

5.2. Developing comprehensive AI agents in medicine

Autonomous agents powered by LLM technology, equipped with advanced language comprehension and reasoning abilities, have had a revolutionary impact on various disciplines. They have been successfully employed as assistants in natural science experiments[189,190] and software engineering projects.[183,191] However, the potential of autonomous agents in the medical field remains largely unexplored,[19,176] due to the complex nature of clinical practice.

AD-AutoGPT has made the first attempt to develop comprehensive AI agents in the medical field, where a specialized AI agent is constructed to autonomously collect, process, and analyze complex health narratives related to Alzheimer's Disease based on textual prompts provided by the user.[176] This agent leverages ChatGPT[9] or GPT-4[6] for task decomposition and augments it with a library of instructions that includes customized tools such as news search, summarization, and result visualization. Moreover, it incorporates specific prompting mechanisms to enhance the efficiency of retrieving Alzheimer's Disease-related information and employs a tailored spatiotemporal information extraction functionality. This pipeline revolutionizes the conventional labor-intensive data analysis approach into a prompt-based automated framework, establishing a solid foundation for subsequent AI-assisted public health research.

Abbasian et al.[19] put forth an innovative system, the Conversational Health Agent (CHA), leveraging LLMs to revolutionize personal healthcare services through empathetic dialog and sophisticated processing of multimodal data. This comprehensive system navigates through a series of pivotal steps, initiating the extraction of user queries from conventional multimodal conversations and transforming them into a structured sequence of executable actions to craft the final response. It then demonstrates its problem-solving prowess by tapping into LLMs as a robust knowledge base for a variety of healthcare tasks. Concurrently, it meticulously retrieves the latest and most relevant healthcare information from reputable published sources, aligning it with the user's specific inquiries. The CHA extends its capabilities by forming connections with diverse external health platforms to acquire up-to-date personal user data. When necessary, it delves into multimodal data analysis, utilizing state-of-the-art external machine-learning healthcare tools. Culminating its processes, the system synthesizes all accumulated information to generate responses that are both tailored to the individual user and reflect the most current knowledge, ensuring transparent communication and providing elucidations on the reasoning and reliability of its approach upon user request. The framework proves its mettle by adeptly handling intricate, multi-step health tasks, such as assessing user stress levels, requiring a nuanced blend of personalization, multimodal data analysis, and extensive health knowledge retrieval.

5.3. Fulfilling individual functions of AI agents

While comprehensive AI agents tailored for the medical field remain relatively rare, previous research that addresses and fulfills individual modules or functions can be considered as initial iterations of AI agents in medicine, illuminating the potential of such agents.[20,21,127,178,179]

ImpressionGPT introduces the utilization of LLMs for the purpose of summarizing radiology reports.[178] It presents a dynamic prompt approach that employs similarity search techniques to incorporate existing reports that are semantically and clinically similar. These similar reports are then used as demonstrations to assist ChatGPT in learning the text descriptions and summarizations of comparable imaging manifestations within a dynamic context. This enables the model to acquire contextual knowledge from analogous instances in the available data, leveraging the in-context learning capabilities of LLMs in a manner that aligns with the memory module of autonomous agents. Furthermore, an iterative optimization algorithm is devised to automatically evaluate the generated results and formulate corresponding instruction prompts. This iterative process enhances the model by facilitating self-reflection, similar to the self-reflection capability observed in autonomous agents.[177,187] In addition, PharmacyGPT extends this framework to address various clinically significant challenges in the pharmacy domain, including patient outcome studies, the generation of AI-based medication prescriptions, and the interpretable clustering analysis of patients, showcasing the versatile potential of autonomous agents in medicine.[179]

Additionally, ChatCAD+, a comprehensive and dependable computer-assisted diagnosis system, is capable of analyzing medical images across a wide range of domains and utilizing up-to-date medical information from reputable sources to offer reliable advice.[20] Specifically, given the input medical image, the system incorporates contrastive language-image pretraining (CLIP) as a domain identifier to select an appropriate model to generate textual descriptions that sufficiently characterize image features. Rather than providing diagnostic advice directly, ChatCAD+ first retrieves pertinent knowledge from professional sources such as Mayo Clinic. This curated information bolsters the LLM's knowledge pool, amplifying the reliability of its diagnostic advice. Additionally, ChatCAD+ integrates a template retrieval mechanism, used in ImpressionGPT,[178] to further improve the report generation performance. Besides, PathAsst system is fine-tuned to have the capabilities of invoking external pathological models and retrieving relevant information from an extensive paper database, making the system capable of handling more complicated tasks and elevating the precision and thoroughness of the responses.[127]

6. Evaluation methods

Given the rapid advancements in the capabilities of LLMs, their integration into critical domains like medicine necessitates a stringent evaluation. These models, while technological marvels, have profound implications on medical decision-making, patient care, and the broader healthcare landscape. As we delve deeper into their technical prowess and applications, the need to ascertain their performance and reliability in medical settings is paramount.

Ensuring the efficacy and safety of LLMs in medicine is vital. An established evaluation framework serves two main purposes: it safeguards against inaccuracies or misjudgments that might have negative consequences in the high-stakes realm of healthcare and provides clear benchmarks and metrics that drive ongoing research and development. Within this framework, evaluating LLMs can be stratified into closed-set and open-set categories based on question types. This distinction is important as it illuminates the model's capabilities in predefined tasks and their adaptability to real-world, unpredictable medical inquiries. Recognizing the complex nature of medicine, which extends beyond mere question answering, this paper also reviews the current progress in assessing LLMs' effectiveness in clinical practice.

6.1. Closed-set evaluation

Closed-set questions come with predefined and limited answer options. Their evaluation often uses benchmark-adapted datasets, with performance metrics derived from these standards. For example, LLaVA-Med[130] measures accuracy for closed-set questions using datasets such as VQA-RAD[192] and SLAKE.[193] Evaluation settings typically utilize either a zero-shot approach or finetuning. The former takes a range of datasets encompassing various tasks, dividing them into "held-in" (used for training) and "held-out" sets (used for testing). After training on the "held-in" sets, performance on unseen datasets or tasks is measured. In contrast, finetuning is more common in domain-specific task evaluations, as demonstrated by LLaVA-Med's results on biomedical VQA.[192,193]

Despite their usefulness, these evaluations often cover only a limited set of tasks or datasets, lacking a broad quantitative comparison. Recent efforts have sought to bridge this gap, like Med-PaLM's introduction of MultiMedQA,[108] which consolidates six medical Q&A datasets, and the addition of another dataset from online medical queries. Another significant contribution is MultiMedBench by Med-PaLM M,[42] a comprehensive benchmark for biomedical tasks. RadFM's RadBench caters specifically to radiology.[43]

6.2. Open-set evaluation

Open-set questions allow for a wider range of responses, making LLMs function similarly to chatbots in this context. Given the diverse content, evaluating these responses is multifaceted. Metrics cover standard measures, expert reviews, model scores, and other unique aspects. The model should prioritize clinical relevance, ensuring its information directly influences patient care. Accuracy, safety, interpretability, ethical considerations, and scalability are also of paramount importance, ensuring the model's predictions are trustworthy and widely applicable.

6.2.1. Standard metrics

Standardized metrics established in the NLP community are often employed to evaluate LLM linguistic outputs. These include F1 score,[43] accuracy,[43] precision,[43] recall,[130] BLEU,[194] METEOR,[195] and ROUGE score.[196] For instance, BLEU evaluates word and phrase overlaps between a model's output and a reference, while METEOR measures lexical and semantic similarities between the generated summary and the reference. These metrics, which range from 0.0 to 1.0, reflect how closely generated outputs match reference answers.

6.2.2. Expert evaluation

In the healthcare domain, model evaluation goes beyond standard metrics like BLEU and ROUGE, given the evident discrepancies when human evaluations depart from automated benchmarks.[21,42,108] Med-PaLM's findings underscored that even top-performing models, such as Flan-PaLM,[108] might not always align with clinicians' preferences. The introduction of clinical radiology-tailored metrics, accompanied by expert assessments on aspects like clinical relevance, offers a more grounded evaluation. Both Yuan et al.[21] and Xu et al.[168] have developed metrics based on clinical evaluations to further refine model assessment. The robust evaluation process begins with pilot studies, followed by expert peer reviews, culminating in real-world clinical tests. This comprehensive framework ensures not only the model's accuracy but also its applicability and safety. Once thoroughly vetted, such models can gradually integrate into clinical workflows, aiding professionals in diagnostics, treatment suggestions, and more.

6.2.3. Model scoring

To address the resource-intensive nature of manual assessments, researchers[103,197–201] are exploring LLM-based scoring systems like GPT-Eval[197] and LLM-Mini-CEX[198]. These systems employ model-centric strategies, wherein one LLM, usually GPT-4, evaluates another one's medical dialogs. For instance, GPT-Eval provides a methodology where the task and criteria are fed into an LLM, leading to a series of evaluation steps that another LLM uses for assessment. LLM-Mini-CEX offers a unique LLM-tailored criterion, streamlining the evaluation of diagnostic abilities by automating interactions using a patient simulator and ChatGPT. However, these methods face challenges related to transparency, accuracy, and sometimes limited diagnostic performance, as noted by Shi et al.[198]

6.2.4. Other aspects

There are also evaluations focusing on unique LLM characteristics,[51] such as faithfulness,[202] hallucination,[203] safety,[204] and robustness against adversarial interventions.[205] For example, to address the challenges posed by hallucinations in LLMs, particularly in the context of the medical domain, Med-HALT provides a diverse multinational dataset derived from medical examinations across various countries and includes multiple innovative testing modalities, especially reasoning hallucination tests.[203]

6.3. Clinical evaluation

The efficacy of standard evaluation methods for general LLMs diminishes when applied to the medical field due to its unique challenges and critical nature. Traditional approaches may fail to capture the detailed concerns of healthcare professionals. Moreover, while quizzes and exams are useful for evaluating general models, they inadequately address the complexity of clinical practice. In response, we explore tailored evaluation strategies that align closely with the medical applications of LLMs.

To date, the development of an evaluation framework that aligns seamlessly with medical practice remains largely uncharted, possibly owing to the nascent stage of LLM development in medicine. A recent initiative by McDuff et al.[206] represents a significant step in this direction, focusing on evaluating LLMs for differential diagnosis (DDx). They introduced an LLM specifically optimized for diagnostic reasoning and assessed its capability to generate DDx independently or as a clinician's aid. In their study, 20 clinicians evaluated 302 challenging, real-world medical cases from the *New England Journal of Medicine* (NEJM) case reports. Each case was reviewed by two clinicians under randomized conditions: one with assistance from search engines and standard medical resources, and the other with additional LLM assistance. Clinicians provided an initial, unassisted DDx before employing any assistive tools. The study found that LLMs for DDx outperformed both unassisted clinicians and those assisted with search engines.

While this study serves as a preliminary model for rigorously evaluating LLMs in clinical settings, there are still areas of concern. The use of NEJM cases, which might have been part of the LLMs' training data, raises questions about the fairness and validity of the evaluation. Moving forward, a more robust and convincing framework should involve designing prospective studies that integrate LLMs into actual clinical practice, rather than relying solely on retrospective studies. In summary, as LLMs show promising advancements in the medical domain, rigorous and comprehensive evaluations are crucial. A combination of automated metrics, expert evaluations, and real-world testing ensures the models' efficacy and safety. As technology and medicine further intertwine, evaluation frameworks must evolve accordingly to ensure the best patient outcomes. Once a model passes this framework, its gradual integration into clinical workflows can commence, starting with tasks like summarizing medical records or aiding in diagnostics, but always under medical professionals' supervision.

7. Discussion

In this review, we have meticulously navigated through the multifaceted landscape of LLMs in the medical domain, illuminating their promising potential. We initiated our exploration by delving into the foundational applications of LLMs in medicine, emphasizing text-based interactions and distinguishing between general-purpose and specialized medical LLMs. Recognizing the inherent multimodality of the medical field, our discussion transitioned to multimodal LLMs, highlighting their capability to integrate diverse data types and thereby augment diagnostic accuracy. Despite these advancements, we acknowledged the persisting challenges such as the need for personalized responses, maintaining currency with the latest medical knowledge, and navigating complex problem-solving scenarios—skills that are indispensable for clinicians in clinical settings. In response to these challenges, we scrutinized the emerging role of LLM-powered autonomous agents in medicine, categorizing their applications and summarizing prevailing evaluation methodologies.

In comparing our findings to the existing body of research on LLMs in healthcare, several studies have paved the way for our understanding of the impact of LLMs. For instance, Moor et al.[1] introduced the concept of a general medical AI, albeit without practical application and data-backed findings. Qiu et al.[56] explored the broad effects that advanced AI models, particularly LLMs, may have on health informatics, identifying seven areas ripe for innovation, such as molecular biology and pharmaceutical research. Further analysis by Thirunavukarasu et al.[58] assesses the capabilities and constraints of LLMs, highlighting their promise in clinical practice, education, and medical research. A body of recent literature[59–64] has also undertaken a comprehensive exploration of the prospective uses and obstacles LLMs face in the healthcare sector. Although these studies offer valuable perspectives, there is still no clear-cut approach for developing advanced medical AI frameworks utilizing LLMs. Therefore, the focus of our paper is to build upon these foundational insights by concentrating on the extensive applications of LLMs in medicine, with a special emphasis on the burgeoning field of multimodal LLMs and LLM-driven autonomous agents in healthcare. Our systematic examination seeks to elucidate LLMs' revolutionary role in advancing medical practice.

Through this extensive analysis, we aimed to provide a balanced and nuanced perspective on the current state of LLMs in medicine. In the contemporary landscape of LLM development, there is a discernible trend toward harnessing LLMs specifically for the medical domain. While general-purpose LLMs exhibit remarkable proficiency, our observations suggest a strategic advantage in not directly fine-tuning them on specialized, long-tailed medical data. Instead, employing highly specialized expert models to handle such nuanced data, followed by storing the processed information in vector databases,[20,127] emerges as a promising paradigm. In practice, querying this database can offer accurate and domain-specific insights. This approach not only presents a potential solution to the "hallucination" phenomenon, where the models may fabricate inconsistent or outright false information, but also paves the way for integration within an autonomous agent-based system. By harnessing these technological strides, it becomes conceivable to develop a cutting-edge, AI-assisted digital healthcare ecosystem.[207] Amidst these advancements, we emphasize the potential of LLMs to revolutionize medical practice while underscoring the imperative for ethical vigilance and continuous scrutiny.

The integration of LLMs in the medical field necessitates a nuanced evaluation from both technological and medical standpoints. From a technical perspective, the proficiency of LLMs in parsing and generating complex, nuanced language is crucial, particularly in understanding and formulating medical terminology, patient narratives, and intricate case details. However, the efficacy of these models hinges on their ability to handle sensitive information ethically, maintain patient confidentiality, and navigate the consequences of misinformation, requiring strict accuracy benchmarks. For medical professionals, the assessment revolves around the practical applicability of LLMs: do they enhance diagnostic accuracy, improve patient communication, and aid in advanced research and treatment methodologies without compromising professional responsibilities or patient trust? Ultimately, the symbiotic evaluation seeks to ensure that LLMs not only exhibit technical excellence but also adhere to the rigorous ethical and professional standards indispensable in healthcare.

Incorporating LLMs into clinical settings hinges on their inference efficiency, a key factor for practical integration. Current LLMs are well-equipped for this task, thanks to advancements in inference acceleration technologies. For private deployments, model compression techniques effectively reduce model size while preserving satisfactory performance.[208,209] Hardware and system redesigns further enhance processing speed.[210,211] Software innovations, such as speculative decoding and KV-cache optimization,[212–215] streamline inference, improving agility and reducing resource demands. Furthermore, physicians have the option to use either commercial models or cloud-deployed LLMs. Cloud deployment leverages substantial computational resources, significantly boosting inference speed. This capability ensures that LLMs can provide real-time assistance and support decision-making in medical environments. The deployment of these models in clinical practice highlights the need to balance complexity with usability, making LLMs' advanced features both accessible and practical in the critical context of healthcare.

Although LLMs possess remarkable intellectual abilities, they inherently face certain constraints and are prone to producing inaccurate or possibly damaging content. Their understanding of context is limited, which is particularly problematic in medical settings where precision is crucial. The models' potential to misinterpret or oversimplify complex medical information necessitates rigorous human oversight. Moreover, LLMs' knowledge is constrained to their training data,[16,17,60,90] potentially leading to outdated or biased responses, and their inability to ask follow-up questions can be a significant drawback in patient care.

Ethical dilemmas significantly pervade the deployment of LLMs in the medical sphere. These models, by inheriting biases from their vast swathes of training data, are at risk of producing responses that are prejudiced or discriminatory, thereby challenging the principles of fairness and equity in healthcare. The "hallucination" phenomenon poses a substantial risk, particularly in medical settings where accuracy is paramount.[15] Moreover, the legal landscape surrounding the use of LLMs in medicine is intricate and fraught with potential pitfalls,[90] necessitating meticulous attention. Issues such as copyright infringement, plagiarism,[60] defamation, and breaches of privacy are prominent concerns that must be proactively addressed to safeguard against legal repercussions and uphold ethical standards. Complicating these issues is the models' inherent complexity and the opaqueness of their internal mechanisms. This

lack of transparency hinders the ability to decipher how specific outputs are generated, leading to potential trust and accountability issues.[60,216]

Despite these challenges, LLMs hold the potential to transform healthcare by enhancing research, improving operational efficiency, and aiding in decision-making. For successful integration and adoption in healthcare, it is imperative to address these limitations and ensure ethical, reliable, and safe applications. Future directions should focus on developing frameworks that recognize and mitigate these constraints, promoting a responsible and informed use of LLMs in medicine. To harness the transformative potential of LLMs in medicine, future developments must prioritize enhancing the models' contextual understanding and ensuring ethical applications. Key efforts should focus on integrating sophisticated algorithms that enable LLMs to grasp the subtleties of medical language and patient information, reducing the risk of misinterpretation. Additionally, real-time learning capabilities are essential for keeping models up-to-date with the latest medical knowledge and practices. In pursuit of this goal, utilizing expert models with high specialization to manage intricate data sets, coupled with the subsequent storage of processed information within vector databases, represents a promising avenue to explore.

Ethical considerations are paramount, necessitating the advancement of bias detection mechanisms and the establishment of robust frameworks for human oversight and model transparency. These measures will contribute to the responsible deployment of LLMs, fostering trust among healthcare professionals and patients. Moreover, the refinement of model compression and inference acceleration will facilitate the use of LLMs across diverse medical environments, ensuring equitable access to AI-powered healthcare solutions. Addressing these focal areas will be instrumental in realizing the full potential of LLMs to enhance medical research, operational efficiency, and patient care while adhering to the high ethical standards required in the healthcare domain.

8. Conclusion

In conclusion, this review offers a comprehensive analysis of the transformative potential of LLMs in modern medicine (an accompanying GitHub repository containing the latest papers is available at https://github.com/mingze-yuan/Awesome-LLM-Healthcare). It demonstrates the fundamental applications of general-purpose and specialized LLMs in areas like knowledge retrieval, research support, workflow automation, and diagnostic assistance. Recognizing the multimodal nature of medicine, the review explores multimodal LLMs and their ability to process diverse data types, including medical imaging and EHRs to enhance diagnostic accuracy. To address the limitations of LLMs regarding personalization and complex clinical reasoning, the emerging role of LLM-powered autonomous agents in medicine is discussed. The review also summarizes methodologies for evaluating the reliability and safety of LLMs in medical contexts.

Overall, LLMs hold remarkable promise in medicine but require continuous optimization and ethical oversight before effective integration into clinical practice. Key challenges highlighted include data limitations, reasoning gaps, potential biases, and transparency issues. Future priorities should focus on developing frameworks to identify and mitigate LLM limitations, guiding responsible and informed applications in healthcare. With prudent progress, LLMs can transform modern medicine through enhanced knowledge consolidation, personalized care, accelerated research, and augmented clinical decision-making. But ultimately, human expertise, ethics, and oversight will remain indispensable in delivering compassionate, high-quality, and equitable healthcare.

CRediT authorship contribution statement

Mingze Yuan: Writing – review & editing, Writing – original draft. Peng Bao: Writing – review & editing, Writing – original draft. Jiajia Yuan: Writing – review & editing, Writing – original draft. Yunhao Shen: Writing – original draft. Zifan Chen: Visualization, Data curation. Yi Xie: Data curation. Jie Zhao: Data curation. Quanzheng Li: Supervision. Yang Chen: Writing – review & editing, Project administration. Li Zhang: Writing – review & editing, Project administration. Lin Shen: Writing – review & editing, Supervision. Bin Dong: Writing – review & editing, Supervision.

Declaration of competing interest

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

Acknowledgements

This work was supported by the National Natural Science Foundation of China (91959205, U22A20327, 82203881, 12090022, 11831002, and 81801778), Beijing Natural Science Foundation (7222021), Beijing Hospitals Authority Youth Programme (QML20231115), Clinical Medicine Plus X-Young Scholars Project of Peking University (PKU2023LCXQ041), and Guangdong Provincial Key Laboratory of Precision Medicine for Gastrointestinal Cancer (2020B121201004).

References

1. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature.* 2023;616(7956):259–265.
2. Ahmed I, Jeon G, Piccialli F. From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where. *IEEE Trans Ind Inf.* 2022;18(8):5031–5042.
3. Wingström R, Hautala J, Lundman R. Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. *Creat Res J.* 2024;36(2):177–193.
4. Lu P, Qiu L, Yu W, et al. A survey of deep learning for mathematical reasoning. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 2023:14605–14631.
5. Lee P, Goldberg C, Kohane I. *The AI Revolution in Medicine: GPT-4 and Beyond.* London: Pearson; 2023.
6. OpenAI. GPT-4 technical report. *arXiv:230308774*. 2023.
7. Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. *JAMA.* 2023;329(16):1349–1350.
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med.* 2023;388(13):1233–1239.
9. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst.* 2022;35:27730–27744.
10. Bai Y., Kadavath S., Kundu S., et al. Constit utional AI: Harmlessness from AI feedback. *arXiv:221208073*. 2022.
11. Haver HL, Ambinder EB, Bahl M, et al. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology.* 2023;307(4):e230424.
12. 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):1–4.
13. Bushuven S, Bentele M, Bentele S, et al. ChatGPT, can you help me save my child's life?"-diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital basic life support and paediatric advanced life support cases–An in-silico analysis. *J Med Syst.* 2023;47(1):123.
14. Xie Y, Seth I, Hunter-Smith DJ, et al. Aesthetic surgery advice and counseling from artificial intelligence: A rhinoplasty consultation with ChatGPT. *Aesthet Plast Surg.* 2023;47(5):1985–1993.
15. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology.* 2023;307(2):e230163.
16. Kothari A. ChatGPT, large language models, and generative AI as future augments of surgical cancer care. *Ann Surg Oncol.* 2023;30(6):3174–3176.
17. Arora A, Arora A. The promise of large language models in health care. *Lancet.* 2023;401(10377):641.
18. Xi Z., Chen W., Guo X., et al. The rise and potential of large language model based agents: A survey. *arXiv:230907864*. 2023.
19. Abbasian M., Azimi I., Rahmani A.M., et al. Conversational health agents: A personalized LLM-powered agent framework. *arXiv:231002374*. 2023.
20. Zhao Z., Wang S., Gu J., et al. ChatCAD+: Towards a universal and reliable interactive CAD using LLMs. *arXiv:230515964*. 2023.
21. Yuan J, Bao P, Chen Z, et al. Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers. *Innov Med.* 2023;1(2):100019.
22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst.* 2017;31:6000–6010.
23. Radford A., Narasimhan K., Salimans T., et al. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed May 9, 2024.
24. Devlin J., Chang M.W., Lee K., et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019. 4171–4186.
25. Radford A., Wu J., Child R., et al. Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed May 9, 2024.
26. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst.* 2020;33:1877–1901.
27. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst.* 2020;33:9459–9474.
28. Zhou D., Schärli N., Hou L., et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv:220510625*. 2022.
29. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst.* 2022;35:24824–24837.
30. Yao S., Yu D., Zhao J., et al. Tree of thoughts: Deliberate problem solving with large language models. *arXiv:230510601*. 2023.
31. Kojima T, Gu SS, Reid M, et al. Large language models are zero-shot reasoners. *Adv Neural Inf Process Syst.* 2022;35:22199–22213.
32. Peng B., Li C., He P., et al. Instruction tuning with GPT-4. *arXiv:230403277*. 2023.
33. Google. Try Bard and share your feedback. https://blog.google/technology/ai/try-bard/. Accessed April 26, 2024.
34. Touvron H., Martin L., Stone K., et al. LlaMA 2: Open foundation and fine-tuned chat models. *arXiv:230709288*. 2023.
35. Lee H., Phatale S., Mansoor H., et al. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv:230900267*. 2023.
36. Bommasani R., Hudson D.A., Adeli E., et al. On the opportunities and risks of foundation models. *arXiv:210807258*. 2021.
37. Dehghani M, Djolonga J, Mustafa B, et al. Scaling vision transformers to 22 billion parameters. In: *International Conference on Machine Learning*. 2023:7480–7512.
38. Borsos Z, Marinier R, Vincent D, et al. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Trans Audio, Speech, Lang Process.* 2023;31:2523–2533.
39. Caruana R. Multitask learning. *Mach Learn.* 1997;28:41–75.
40. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: *International Conference on Machine Learning*. 2008:160–167.

41. Ruder S. An overview of multi-task learning in deep neural networks. *arXiv:170605098*. 2017.
42. Tu T., Azizi S., Driess D., et al. Towards generalist biomedical AI. *arXiv:230714334*. 2023.
43. Wu C., Zhang X., Zhang Y., et al. Towards generalist foundation model for radiology. *arXiv:230802463*. 2023.
44. Zhao W.X., Zhou K., Li J., et al. A survey of large language models. *arXiv:230318223*. 2023.
45. Yang J., Jin H., Tang R., et al. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *arXiv:230413712*. 2023.
46. Chang T.A., Bergen B.K. Language model behavior: A comprehensive survey. *arXiv:230311504*. 2023.
47. Wang Y., Zhong W., Li L., et al. Aligning large language models with human: A survey. *arXiv:230712966*. 2023.
48. Huang J., Chang K.C.C. Towards reasoning in large language models: A survey. *arXiv:221210403*. 2022.
49. Zhang S., Dong L., Li X., et al. Instruction tuning for large language models: A survey. *arXiv:230810792*. 2023.
50. Mialon G., Dessì R., Lomeli M., et al. Augmented language models: A survey. *arXiv:230207842*. 2023.
51. Chang Y., Wang X., Wang J., et al. A survey on evaluation of large language models. *arXiv:230703109*. 2023.
52. Yin S., Fu C., Zhao S., et al. A survey on multimodal large language models. *arXiv:230613549*. 2023.
53. Wang L., Ma C., Feng X., et al. A survey on large language model based autonomous agents. *arXiv:230811432*. 2023.
54. Weng L. LLM-powered autonomous agents. https://lilianweng.github.io/posts/2023–06-23-agent/. Accessed April 25, 2024.
55. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. *N Engl J Med.* 2023;388(21):1981–1990.
56. Qiu J., Li L., Sun J., et al. Large AI models in health informatics: Applications, challenges, and the future. *arXiv:230311568*. 2023.
57. Liu C., Liu Z., Holmes J., et al. Artificial general intelligence for radiation oncology. *arXiv:230902590*. 2023.
58. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med.* 2023;29(8):1930–1940.
59. Li J, Dada A, Kleesiek J, et al. ChatGPT in healthcare: A taxonomy and systematic review. *Comput Methods Prog Biomed.* 2024;245:108013.
60. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare.* 2023;11(6):887.
61. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res.* 2023;25:e48568.
62. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med.* 2023;3(1):141.
63. Omiye J.A., Gui H., Rezaei S.J., et al. Large language models in medicine: The potentials and pitfalls. *arXiv:230900087*. 2023.
64. He K., Mao R., Lin Q., et al. A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics. *arXiv:231005694*. 2023.
65. Jie Z, Zhiying Z, Li L. A meta-analysis of Watson for oncology in clinical application. *Sci Rep.* 2021;11(1):5792.
66. Strickland E. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectr.* 2019;56(4):24–31.
67. Ross C., Swetlitz I. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/. Accessed May 9, 2024.
68. Ott S., Hebenstreit K., Liévin V., et al. ThoughtSource: A central hub for large language model reasoning data. *arXiv:230111596*. 2023.
69. Singhal K., Tu T., Gottweis J., et al. Towards expert-level medical question answering with large language models. *arXiv:230509617*. 2023.
70. Nori H., King N., McKinney S.M., et al. Capabilities of GPT-4 on medical challenge problems. *arXiv:230313375*. 2023.
71. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ.* 2023;9(1):e45312.
72. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit Health.* 2023;2(2):e0000198.
73. 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.
74. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol.* 2023;141(6):589–597.
75. Hu X, Ran AR, Nguyen TX, et al. What can GPT-4 do for diagnosing rare eye diseases? A pilot study. *Ophthalmol Ther.* 2023;12(6):3395–3402.
76. Humar P, Asaad M, Bengur FB, et al. ChatGPT is equivalent to first year plastic surgery residents: Evaluation of ChatGPT on the plastic surgery in-service exam. *Aesthetic Surg J.* 2023;43(12):NP1085–NP1089.
77. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI.* 2023;1(1):AIp2300031.
78. 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.
79. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: How will ChatGPT impact information seeking from the medical literature? *J Am Soc Nephrol.* 2023;34(8):1302–1304.
80. Biswas SS. Role of chat GPT in public health. *Ann Biomed Eng.* 2023;51(5):868–869.
81. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. *Radiology.* 2023;307(5):e230922.
82. Cheng K, Wu H, Li C. ChatGPT/GPT-4: Enabling a new era of surgical oncology. *Int J Surg.* 2023;109(8):2549–2550.
83. Carlbring P, Hadjistavropoulos H, Kleiboer A, et al. A new era in internet interventions: The advent of Chat-GPT and AI-assisted therapist guidance. *Internet Interv.* 2023;32:100621.
84. He N, Yan Y, Wu Z, et al. Chat GPT-4 significantly surpasses GPT-3.5 in drug information queries. *J Telemed Telecare.* 2023 1357633X231181922.
85. Blanco-Gonzalez A, Cabezon A, Seco-Gonzalez A, et al. The role of AI in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals.* 2023;16(6):891.
86. Sun YX, Li ZM, Huang JZ, et al. GPT-4: The future of cosmetic procedure consultation? *Aesthetic Surg J.* 2023;43(8):NP670–NP672.
87. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: The end of the consulting infection doctor? *Lancet Infect Dis.* 2023;23(4):405–406.
88. 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):721–732.
89. Shea YF, Lee CMY, Ip WCT, et al. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. *JAMA Netw Open.* 2023;6(8):e2325000.
90. Biswas S. ChatGPT and the future of medical writing. *Radiology.* 2023;307:e223312.
91. Ghim JL, Ahn S. Transforming clinical trials: The emerging roles of large language models. *Transl Clin Pharm.* 2023;31(3):131–138.
92. Peng C., Yang X., Chen A., et al. A study of generative large language model for medical research and healthcare. *arXiv:230513523*. 2023.
93. Woo M. An AI boost for clinical trials. *Nature.* 2019;573(7775):S100–S102.
94. Hamer D.M. den, Schoor P., Polak T.B., et al. Improving patient pre-screening for clinical trials: Assisting physicians with large language models. *arXiv:230407396*. 2023.
95. Jin Q., Wang Z., Floudas C.S., et al. Matching patients to clinical trials with large language models. *arXiv:230715051*. 2023.

96. White R, Peng T, Sripitak P, et al. CliniDigest: A case study in large language model based large-scale summarization of clinical trial descriptions. In: *ACM Conference on Information Technology for Social Good*. 2023:396–402.
97. Wang Z., Xiao C., Sun J. AutoTrial: Prompting language models for clinical trial design. *arXiv:230511366*. 2023.
98. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology*. 2023;308(3):e231362.
99. Ali SR, Dobbs TD, Hutchings HA, et al. Using ChatGPT to write patient clinic letters. *Lancet Digit Health*. 2023;5(4):179–181.
100. Wang Y., Zhao Y., Petzold L. Are large language models ready for healthcare? A comparative study on clinical language understanding. *arXiv:230405368*. 2023.
101. Liu Z., Zhong T., Li Y., et al. Evaluating large language models for radiology natural language processing. *arXiv:230713693*. 2023.
102. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. *NPJ Digit Med*. 2023;6(1):158.
103. Zhang H., Chen J., Jiang F., et al. HuatuoGPT, towards taming language model to be a doctor. *arXiv:230515075*. 2023.
104. 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.
105. Mao R., Chen G., Zhang X., et al. GPTEval: A survey on assessments of ChatGPT and GPT-4. *arXiv:230812488*. 2023.
106. Sheng B, Guan Z, Lim LL, et al. Large language models for diabetes care: Potentials and prospects. *Sci Bull*. 2024;69(5):583–588.
107. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194.
108. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180.
109. 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.
110. Wang H., Liu C., Xi N., et al. Huatuo: Tuning LLaMA model with Chinese medical knowledge. *arXiv:230406975*. 2023.
111. Xiong H., Wang S., Zhu Y., et al. DoctorGLM: Fine-tuning your Chinese doctor is not a herculean task. *arXiv:230401097*. 2023.
112. Wu C., Zhang X., Zhang Y., et al. PMC-LLaMA: Further finetuning LLaMA on medical papers. *arXiv:230414454*. 2023.
113. Chen Y., Wang Z., Xing X., et al. BianQue: Balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT. *arXiv:2310.15896*. 2023.
114. Wang G., Yang G., Du Z., et al. ClinicalGPT: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv:230609968*. 2023.
115. Liu Z., Zhong A., Li Y., et al. Radiology-GPT: A large language model for radiology. *arXiv:230608666*. 2023.
116. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357–362.
117. Chowdhery A., Narang S., Devlin J., et al. PaLM: Scaling language modeling with pathways. *arXiv:220402311*. 2022.
118. Taori R., Gulrajani I., Zhang T., et al. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca. Accessed April 25, 2024.
119. Wei Zhu W.Y., Wang X. ShenNong-TCM: A traditional Chinese medicine large language model. https://github.com/michael-wzhu/ShenNong-TCM-LLM. Accessed April 25, 2024.
120. Zhang J., Yang S., Huang Di. https://github.com/Zlasejd/HuangDi. Accessed May 9, 2024.
121. Wang M., Yan X., Xue D. Sunsimiao: Chinese medicine LLM. https://github.com/thomas-yanxin/Sunsimiao. Accessed April 25, 2024.
122. Kang Y., Chang Y., Fu J., et al. CMLM-ZhongJing: Large language model is good story listener. https://github.com/pariskang/CMLM-ZhongJing. Accessed April 25, 2024.
123. Zhang Z, Tan Y, Li M, et al. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. *Comput Biol Med*. 2024;172:108290.
124. Yang G., Shi J., Wang Z., et al. TCM-GPT: Efficient pre-training of large language models for domain adaptation in traditional Chinese medicine. *arXiv:231101786*. 2023.
125. Nori H., Lee Y.T., Zhang S., et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv:231116452*. 2023.
126. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28(9):1773–1784.
127. Sun Y., Zhu C., Zheng S., et al. PathAsst: Redefining pathology through generative foundation AI assistant for pathology. *arXiv:230515072*. 2023.
128. Zhang K., Yu J., Yan Z., et al. BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv:230517100*. 2023.
129. Zhang X., Wu C., Zhao Z., et al. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv:230510415*. 2023.
130. Li C., Wong C., Zhang S., et al. LlaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv:230600890*. 2023.
131. Thawakar O., Shaker A.M., Mullappilly S.S., et al. XrayGPT: Chest radiographs summarization using medical vision-language models. *arXiv:230607971*. 2023.
132. Ma L., Han J., Wang Z., et al. CephGPT-4: An interactive multimodal cephalometric measurement and diagnostic system with visual large language model. *arXiv:230707518*. 2023.
133. Moor M., Huang Q., Wu S., et al. Med-flamingo: A multimodal medical few-shot learner. *arXiv:230715189*. 2023.
134. Wang R., Duan Y., Li J., et al. XrayGLM: The first Chinese medical multimodal model that chest radiographs summarization. https://github.com/WangRongsheng/XrayGLM. Accessed April 25, 2024.
135. Wu C., Lei J., Zheng Q., et al. Can GPT-4V (ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. *arXiv:231009909*. 2023.
136. Zhou J., Chen X., Gao X. Path to medical AGI: Unify domain-specific medical LLMs with the lowest cost. *arXiv:230610765*. 2023.
137. Luo Y., Zhang J., Fan S., et al. BioMedGPT: Open multimodal generative pre-trained transformer for biomedicine. *arXiv:230809442*. 2023.
138. Belyaeva A., Cosentino J., Hormozdiari F., et al. Multimodal LLMs for health grounded in individual-specific data. *arXiv:230709018*. 2023.
139. Theodoris CV, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. *Nature*. 2023;618(7965):616–624.
140. Huang H, Zheng O, Wang D, et al. ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model. *Int J Oral Sci*. 2023;15(1):29.
141. Huang Z, Bianchi F, Yuksekgonul M, et al. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat Med*. 2023;29(9):2307–2316.
142. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-judge with mt-bench and chatbot arena. *Adv Neural Inf Process Syst*. 2024;36:46595–46623.
143. Wang P, Yang A, Men R, et al. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: *International Conference on Machine Learning*. 2022:23318–23340.
144. Lin W., Zhao Z., Zhang X., et al. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. *arXiv:230307240*. 2023.
145. Liu H., Li C., Wu Q., et al. Visual instruction tuning. *arXiv:230408485*. 2023.

146. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. 2021:8748–8763.
147. Zhang S., Xu Y., Usuyama N., et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv:230300915*. 2023.
148. Wang Z., Wu Z., Agarwal D., et al. MedCLIP: Contrastive learning from unpaired medical images and text. *arXiv:221010163*. 2022.
149. Johnson AE, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data.* 2019;6(1):317.
150. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inf Assoc.* 2016;23(2):304–310.
151. Zhu D., Chen J., Shen X., et al. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv:230410592*. 2023.
152. Du Z, Qian Y, Liu X, et al. GLM: General language model pretraining with autoregressive blank infilling. In: *Annual Meeting of the Association for Computational Linguistics*. 2022:320–335.
153. Driess, Xia D, Sajjadi MSM F, et al. PaLM-E: An embodied multimodal language model. In: *International Conference on Machine Learning*. 2023:8469–8488.
154. Awadalla A., Gao I., Gardner J., et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv:230801390*. 2023.
155. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*. 2021:1–21.
156. Lo K., Wang L.L., Neumann M., et al. S2ORC: The semantic scholar open research corpus. *arXiv:191102782*. 2019.
157. Xu S., Yang L., Kelly C., et al. ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. *arXiv:230801317*. 2023.
158. Anil R., Dai A.M., Firat O., et al. Palm 2 technical report. *arXiv:230510403*. 2023.
159. Yang Z., Li L., Lin K., et al. The dawn of LMMs: Preliminary explorations with GPT-4V (ision). *arXiv:230917421*. 2023.
160. Yang Z., Yao Z., Tasmin M., et al. Performance of multimodal GPT-4V on USMLE with image: Potential for imaging diagnostic support with explanations. *medRxiv*. 2023. https://doi.org/10.1101/2023.10.26.23297629.
161. Sorin V., Glicksberg B.S., Barash Y., et al. Diagnostic accuracy of GPT multimodal analysis on USMLE questions including text and visuals. *medRxiv*. 2023. https://doi.org/10.1101/2023.10.29.23297733.
162. Yan Z., Zhang K., Zhou R., et al. Multimodal ChatGPT for medical applications: An experimental study of GPT-4V. *arXiv:231019061*. 2023.
163. Li Y., Liu Y., Wang Z., et al. A systematic evaluation of GPT-4V's multimodal capability for medical image analysis. *arXiv:231020381*. 2023.
164. Wei J., Bosma M., Zhao V., et al. Finetuned language models are zero-shot learners. *arXiv:2109.01652*. 2023.
165. Chen W, Li Z, Fang H, et al. A benchmark for automatic medical consultation system: Frameworks, tasks and datasets. *Bioinformatics.* 2022;39(1):btac817.
166. Karargyris A, Umeton R, Sheller MJ, et al. Federated benchmarking of medical artificial intelligence with MedPerf. *Nat Mach Intell.* 2023;5(7):799–810.
167. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine.* 2023;95:104770.
168. Xu J., Lu L., Yang S., et al. MedGPTEval: A dataset and benchmark to evaluate responses of large language models in medicine. *arXiv:230507340*. 2023.
169. Qian B, Chen H, Wang X, et al. DRAC 2022: A public benchmark for diabetic retinopathy analysis on ultra-wide optical coherence tomography angiography images. *Patterns.* 2024;5(3):100929.
170. Wang Y., Kordi Y., Mishra S., et al. Self-instruct: Aligning language model with self generated instructions. *arXiv:221210560*. 2022.
171. Han T., Adams L.C., Papaioannou J.M., et al. MedAlpaca–an open-source collection of medical conversational AI models and training data. *arXiv:230408247*. 2023.
172. Wang J., Yang Z., Hu X., et al. GIT: A generative image-to-text transformer for vision and language. *arXiv:220514100*. 2022.
173. Liu F., Eisenschlos J.M., Piccinno F., et al. DePlot: One-shot visual language reasoning by plot-to-table translation. *arXiv:221210505*. 2022.
174. Wang Y., Si S., Li D., et al. Preserving in-context learning ability in large language model fine-tuning. *arXiv:221100635*. 2022.
175. Jaegle A, Gimeno F, Brock A, Vinyals O, Zisserman A, Carreira J. General perception with iterative attention. In: *International Conference on Machine Learning*. 2021:4651–4664.
176. Dai H., Li Y., Liu Z., et al. AD-AutoGPT: An autonomous GPT for Alzheimer's disease infodemiology. *arXiv:230610095*. 2023.
177. Yao S., Zhao J., Yu D., et al. ReAct: Synergizing reasoning and acting in language models. *arXiv:221003629*. 2022.
178. Ma C., Wu Z., Wang J., et al. ImpressionGPT: An iterative optimizing framework for radiology report summarization with ChatGPT. *arXiv:230408448*. 2023.
179. Liu Z., Wu Z., Hu M., et al. PharmacyGPT: The AI pharmacist. *arXiv:230710432*. 2023.
180. Franklin S, Graesser A. Is it an agent, or just a program?: A taxonomy for autonomous agents. In: *International Workshop on Agent Theories, Architectures, and Languages*. 1996:21–35.
181. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature.* 2015;518(7540):529–533.
182. Significant G. AutoGPT. https://github.com/Significant-Gravitas/AutoGPT. Accessed May 13, 2024.
183. Hong S., Zheng X., Chen J., et al. MetaGPT: Meta programming for multi-agent collaborative framework. *arXiv:230800352*. 2023.
184. Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D. Out of one, many: Using language models to simulate human samples. *Polit Anal.* 2023;31(3):337–351.
185. Hu C., Fu J., Du C., Luo S., Zhao J., Zhao H. ChatDB: Augmenting LLMs with databases as their symbolic memory. *arXiv:230603901*. 2023.
186. Zhong W., Guo L., Gao Q., et al. MemoryBank: Enhancing large language models with long-term memory. *arXiv:230510250*. 2023.
187. Shinn N., Cassano F., Labash B., et al. Reflexion: Language agents with verbal reinforcement learning. *arXiv:230311366*. 2023.
188. Schick T., Dwivedi-Yu J., Dessì R., et al. Toolformer: Language models can teach themselves to use tools. *arXiv:230204761*. 2023.
189. Boiko D.A., MacKnight R., Gomes G. Emergent autonomous scientific research capabilities of large language models. *arXiv:230405332*. 2023.
190. Bran A.M., Cox S., White A.D., et al. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv:230405376*. 2023.
191. Qian C., Cong X., Yang C., et al. Communicative agents for software development. *arXiv:230707924*. 2023.
192. Lau JJ, Gayen S, Ben Abacha A, et al. A dataset of clinically generated visual questions and answers about radiology images. *Sci Data.* 2018;5(1):180251.
193. Liu B, Zhan LM, Xu L, et al. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: *International Symposium on Biomedical Imaging*. 2021:1650–1654.

194. Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation. In: *Annual Meeting of the Association for Computational Linguistics*. 2002:311–318.
195. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 2005:65–72.
196. Lin CY. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. 2004:74–81.
197. Liu Y., Iter D., Xu Y., et al. G-Eval: NLG evaluation using GPT-4 with better human alignment. *arXiv:230316634*. 2023.
198. Shi X., Xu J., Ding J., et al. LLM-mini-CEX: Automatic evaluation of large language model for diagnostic conversation. *arXiv:230807635*. 2023.
199. Fu J., Ng S.K., Jiang Z., Liu P. GPTScore: Evaluate as you desire. *arXiv:230204166*. 2023.
200. Chen Y., Wang R., Jiang H., et al. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv:230400723*. 2023.
201. Chiang C.H., Lee H. Can large language models be an alternative to human evaluations? *arXiv:230501937*. 2023.
202. Xie Q., Schenck E.J., Yang H.S., et al. Faithful AI in medicine: A systematic review with large language models and beyond. *medRxiv*. 2023. https://doi.org/10.1101/2023.04.18.23288752.
203. Umapathi L.K., Pal A., Sankarasubbu M. Med-HALT: Medical domain hallucination test for large language models. *arXiv:230715343*. 2023.
204. Zhang Z., Lei L., Wu L., et al. SafetyBench: Evaluating the safety of large language models with multiple choice questions. *arXiv:230907045*. 2023.
205. Wang B., Xu C., Wang S., et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv:211102840*. 2021.
206. McDuff D., Schaekermann M., Tu T., et al. Towards accurate differential diagnosis with large language models. *arXiv:231200164*. 2023.
207. Guan Z, Li H, Liu R, et al. Artificial intelligence in diabetes management: Advancements, opportunities, and challenges. *Cell Rep Med.* 2023;4(10):101213.
208. Frantar E, Ashkboos S, Hoefler T, et al. OPTQ. Accurate quantization for generative pre-trained transformers. In: *International Conference on Learning Representations*. 2022:1–16.
209. Ahmadian A., Dash S., Chen H., et al. Intriguing properties of quantization at scale. *arXiv:230519268*. 2023.
210. Tian R, Zhao Z, Liu W, et al. SAMP: A model inference toolkit of post-training quantization for text processing via self-adaptive mix-edprecision. In: *Conference on Empirical Methods in Natural Language Processing: Industry Track*. 2023:123–130.
211. Sheng Y, Zheng L, Yuan B, et al. Flexgen: High-throughput generative inference of large language models with a single GPU. In: *International Conference on Machine Learning*. 2023:31094–31116.
212. Kim S, Mangalam K, Moon S, et al. Speculative decoding with big little decoder. *Adv Neural Inf Process Syst.* 2023;36:39236–39256.
213. Leviathan Y, Kalman M, Matias Y. Fast inference from transformers via speculative decoding. In: *International Conference on Machine Learning*. 2023:19274–19286.
214. Zhang Z., Sheng Y., Zhou T., et al. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. *arXiv:230614048*. 2023.
215. Liu Z., Desai A., Liao F., et al. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. *arXiv:230517118*. 2023.
216. Ufuk F. The role and limitations of large language models such as ChatGPT in clinical settings and medical journalism. *Radiology.* 2023;307(3):e230276.