npj | artificial intelligence

Review

https://doi.org/10.1038/s44387-026-00076-4

AI agent in healthcare: applications, evaluations, and future directions

Lina Zhao[1], Shengrui Liu[2], Tangsiwei Xin[3], Jiawen Tan[4], Xiaoran Wang[5], Yafang Li[1], Zihao Bian[6], Yiyang Chen[6], Fanyi Kong[6], Jinwei Bian[7], Chen Qian[8] ✉ & Zongjiu Zhang[6] ✉

With the rapid advancement of large language model (LLM) technologies, AI agents have rapidly emerged in healthcare. This review traces the historical evolution and core characteristics of AI agents, and systematically examines their applications in assisted diagnosis, clinical decision support, medical report generation, patient-facing chatbots, healthcare system management, and medical education. We further analyze existing evaluation frameworks for AI agents in healthcare, focusing on key dimensions and performance metrics. Looking ahead, we propose seven critical directions for future development: integration with embodied systems, hybrid expert models, expanded evaluation paradigms, safety and controllability assurance, ethical governance and user trust, and guidance for evolving roles of healthcare staff. This review aims to offer a comprehensive perspective on the development and implementation of AI agents in healthcare, providing theoretical support for future research, practice, and governance.

Recent breakthroughs in large language models (LLMs) have led to their rapid adoption in the healthcare domain[1], with widespread applications in medical question answering, electronic health record (EHR) generation, and clinical decision support. Meanwhile, LLM-based agents are rapidly emerging. Actually, in clinical practice, healthcare professionals are often confronted with multimodal and highly heterogeneous data, heavy workloads, and the pressing demand for time-critical clinical decisions. Against this backdrop, AI agents offer a compelling solution: they not only understand and generate human language, but also autonomously orchestrate multi-step tasks through tool usage, thereby exhibiting goal-directed reasoning and decision-making capabilities[2]. Consequently, agents are increasingly recognized as a frontier in medical technology.

Initial research has explored the potential of AI agents in healthcare. For example, Qiu et al. examined their applications in diagnostic support and workflow optimization, while also highlighting challenges related to data privacy and over-reliance[3]. Other scholars have analyzed the core functionalities of agentic AI in diagnosis, clinical operations, drug development, and robotic-assisted interventions[4]. Still others have provided broad overviews of current application scenarios, emphasizing risks such as hallucination, limited generalizability, and ethical concerns[5]. Recent work by Moritz has further articulated the paradigm of coordinated multi-agent systems in healthcare (MASH), highlighting the potential of decentralized yet interoperable LLM-powered agents that collaborate to optimize clinical and operational workflows, while outlining core implementation challenges, including secure communication, interoperability, and validation in clinical settings[6]. However, compared to the rapidly expanding literature on LLMs in healthcare, research specifically focused on LLM-based AI agents remains scarce[7]. Moreover, existing reviews often fall short in terms of breadth, evaluation depth, and theoretical framing. This study aims to fill these gaps by providing a more comprehensive and structured review to support the development and deployment of AI agents in healthcare.

To address the growing interest and uncertainty around the deployment of LLM-based AI agents in healthcare, this review seeks to answer the overarching question: What is the current landscape of LLM-based AI agents in healthcare, and how can their implementation be effectively evaluated, and how can their deployment be guided to ensure safety, controllability, and reliability? In pursuit of this question, the objectives of this review are fourfold: (1) To synthesize the conceptual foundations, historical evolution, and defining features of LLM-based AI agents; (2) To map and analyze their application scenarios and representative systems in healthcare; (3) To construct a multi-dimensional evaluation framework encompassing clinical and humanistic dimensions; (4) To identify and discuss critical directions for future development.

[1]School of Biomedical Engineering, Tsinghua University, Beijing, China. [2]School of Government, Nanjing University, Nanjing, China. [3]Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Tsinghua, China. [4]Medical Engineering Department, the First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [5]Peking Union Medical College Hospital, Beijing, China. [6]Department of Hospital Management, Tsinghua University, ShenZhen, China. [7]School of Nursing, The University of Hong Kong, Hong Kong, China. [8]School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China. ✉e-mail: qianc62@gmail.com; zhangzongjiu@tsinghua.edu.cn

The remainder of this paper is organized as follows: Section 2 reviews the evolution and distinguishing features of AI agents, categorizes their applications in healthcare, and introduces a multi-dimensional evaluation framework; Section 3 discusses seven critical directions for future development, and concludes with a summary of findings, limitations, and practical implications; Section 4 presents the method we used.

Results

Historical evolution of agent

The conceptualization of the term "Agent" can be traced back to the depths of philosophical contemplation, traversing the realms of both theoretical musings and practical technological implementation. The quest to understand "Agent" has transcended disciplinary boundaries, marking a relentless journey across various academic domains. Ancient Greek philosophers showed an interest in intelligent machines. During this period, philosophers began to describe entities that possessed desires, beliefs, intentions, and the capacity to act, thereby illuminating the nascent concept of intelligent bodies[8]. The concept of "telos" (end), as articulated by Aristotle, subsequently provided a philosophical basis for the goal-directed characteristics that would come to define intelligent bodies in subsequent eras.

In the contemporary era, following the advancements witnessed in the domains of natural sciences and computer technology, the realm of research in artificial intelligence has undergone a transition, progressing from philosophical contemplations to practical applications. The 1950s witnessed the proposal of the renowned "Turing Test" by Alan Turing, which subsequently emerged as a pivotal benchmark for evaluating the intelligence of machines. The subsequent emergence of expert systems in the 1970s primarily entailed the utilization of human expert knowledge to facilitate reasoning and decision-making through the utilization of computer programmes[9]. The advent of machine learning techniques, which facilitate the acquisition of knowledge and skills from data, has led to a substantial enhancement in the intelligence of agents. In the 21st century, deep learning technology has achieved significant breakthroughs in perception, decision-making, execution capabilities, and the expansion of application scenarios, thereby bringing revolutionary progress to the development of agents[10]. Notably, the field of reinforcement learning (RL), particularly multi-agent Reinforcement Learning (MARL), has witnessed substantial advancements, successfully addressing numerous sequential decision-making problems in machine learning[11]. These advancements have enabled Agent to make more optimal decisions in complex environments.

After 2022, AI has penetrated various aspects of society, particularly with the proliferation of LLMs, which have opened up new avenues for the advancement of agents. AI agents built on the foundation of large AI models have led to a period of accelerated growth and development. AI agents based on LLM have a richer knowledge base, more natural human interaction capabilities, and better interpretability compared to reinforcement learning agents[12]. For instance, OpenAI has introduced the Custom GPT feature (GPTs), allowing users to create their own GPT by integrating knowledge, operations, and instructions. Google has launched an agent framework through the Gemini series of models, supporting multimodal task processing. The LLaMA series of models open-sourced by Meta has given rise to a large number of community-driven agent applications. Anthropic's Claude model sets new standards for agents in terms of safety and controllability through the Constitutional AI framework. DeepMind's Sparrow project demonstrates an innovative path of combining language models with reinforcement learning. These developments have paved the way for the widespread use of personalized AI assistants, marking the formation of a diverse ecosystem for AI agent technology.

The advent of LLMs has prompted numerous organizations to prioritize the development of LLM-based AI agents, particularly within the healthcare sector. For instance, IBM Watson Health utilizes natural language processing, machine learning, and big data technologies to furnish healthcare organizations with a range of intelligent services, encompassing assisted diagnosis, patient care, and drug development[13]. The whole progression of AI agent is visualized in Fig. 1 with key milestones.

Interpretation of agent

There is currently no universally accepted definition of AI agent in academic circles. In 1998, Cristiano Castelfranchi proposed the concept of an AI agent as an intelligent entity capable of goal-orientation, social intelligence, mind-reading, adaptability, and flexibility, and able to make decisions and take actions autonomously[14]. Weng defines an AI agent as an autonomous system with a LLM as its core controller, which handles complex tasks through the ability to plan, remember, and use tools, as "LLM+memory +task planning+tool use"[15]. Feifei Li's team characterizes an AI agent as an intelligent entity capable of sensing its environment, making decisions, and performing actions, with its core focus on using LLMs or visual language models (VLMs) to enhance the system's interactivity and adaptability, emphasizing its ability to plan for task execution and use and reason about large-scale knowledge. Parisi describes the AI agent as a system that enables using external API tools to extend the model's capabilities[16]. The language model explored by Schick et al. improves its performance by learning to use tools, suggesting a view of the AI agent as an intelligent system that can autonomously perceive its environment, understand the task requirements, and select and perform appropriate actions based on those requirements, including invoking external tools[17].

In the preceding definitions, Weng's characterization, which foregrounds the LLM as the core controller and systematically integrates planning, memory, and tool use into a unified framework, is particularly well-suited for constructing autonomous systems capable of handling complex, multi-step tasks. Compared with Castelfranchi's earlier macro-level cognitive-architecture perspective, Weng's account is more operational and readily implementable. In contrast to the tool-calling and functional-extension emphases found in the work of Parisi, Schick, and others, it offers a more holistic and autonomous conception of the system. Therefore, this study adopts Weng's definition, conceptualizing an AI agent as an autonomous intelligent system whose central controller is a large language model, supplemented by four key modules: planning, memory, tool use, and self-reflection, to ensure efficient and reliable execution of domain-specific tasks in healthcare[15].

Characteristics of agent

Understanding and generating text. The AI agent, when combined with the LLMs, demonstrates strong proficiency in understanding and generating text. This proficiency is evidenced by its in-depth understanding of the contextual information of the text, as well as its powerful text generation capability, which is capable of generating natural and smooth text content[13]. This has resulted in a revolutionary change in the fields of dialog systems, content creation, and so on. This combination of comprehension and generation capabilities enables AI agents to interact more intelligently and personally with humans, providing more sophisticated and tailored services.

Tool Use and Interactivity. In addition to their powerful learning and processing capabilities, AI agents can self-learn how to use external tools[15]. They are able to select the most appropriate tools for a given situation and obtain the required information or perform specific operations through API calls and other means, thus further enhancing the efficiency and accuracy of task processing[18]. The introduction of this tool-using capability enhances the autonomy of the AI agent and provides more possibilities for its interaction with humans or other systems.

Task Processing and Generalizability. The integration capability of AI agents is of great significance to the development of the field of AI[15]. The ability of AI agents to seamlessly integrate with other information systems and devices facilitates collaboration and information sharing among them[17]. For instance, in diagnostic support scenarios, AI agents can integrate with Electronic Health Record (EHR) systems, Picture Archiving and Communication Systems (PACS), and Laboratory Information Systems (LIS) to automatically extract patients' multimodal data. This assists physicians in comprehensive decision-making, reduces human errors, and enhances diagnostic accuracy and efficiency[19,20]. The great versatility and flexibility that AI agents based on LLMs show in handling tasks is evidence that they are capable of handling many different tasks and problems, as well as freely

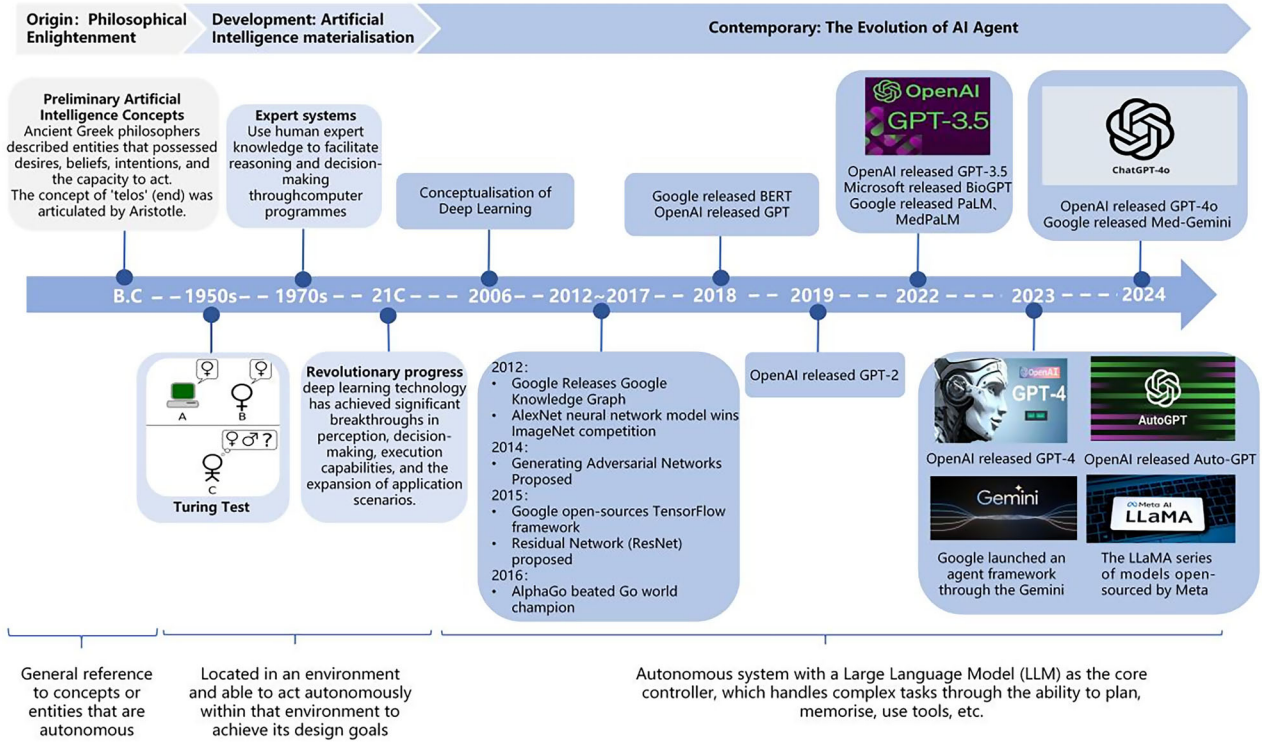

Fig. 1 | Evolution of AI agent

switching between multiple domains. This wide applicability and powerful task processing capability are important tools to promote AI agents as a means to solve complex problems and promote intelligent transformation.

Logical Reasoning and Task Decomposition. The employment of AI agents based on LLMs can be likened to the endowment of a more powerful "brain" to the agents in question. LLMs possess the faculty of logical reasoning, a capability that can be further augmented through the implementation of prompting strategies by agents. However, if the prompt is not sufficiently effective in stimulating the reasoning ability of the LLM itself, users may encounter difficulties in obtaining satisfactory answers. In contrast, the addition of auxiliary reasoning prompts can significantly improve the reasoning effectiveness of the LLM[12,21]. The ability of autonomous agents to generate bespoke prompts that align with specific objectives underscores their potential to more effectively stimulate and leverage reasoning capabilities in handling complex reasoning tasks.

Learning and Adaptation Capability. In comparison with traditional AI technologies, AI agents based on LLMs demonstrate exceptional learning and adaptation capabilities. These agents autonomously learn from large-scale data, extracting key information, and continuously optimizing their performance[22]. These agents are capable of self-learning from substantial amounts of data, extracting key information, and continuously optimizing their own performance. This process requires minimal reliance on a large number of manual annotations or preset rules. Furthermore, these agents possess the capacity to acquire knowledge from limited or even zero samples, rapidly adapting to new tasks or small data sets while demonstrating commendable performance[23]. Moreover, the system's highly scalable nature encourages continuous improvement in performance and self-driven evolution to meet the ever-growing demands of substantial applications[24]. Fig. 2 synthesizes these five core capabilities of AI agent.

Application of agent

The potential for AI agents to demonstrate significant application in a variety of fields, including education, industry, finance, transportation, logistics, and more, is attributable to their advanced flexibility and intelligent processing capabilities. For instance, in the domain of financial investment, Robo-advisors represent a prominent example of intelligent robotic investment advisor application, capable of creating and managing diversified investment portfolios through the utilization of technology, algorithms, and scientific portfolio theories[25]. FinRobot, a novel open-source AI agent platform, employs LLM to drive multiple AI agents. It specializes in finance, providing more effective financial advice, portfolio management, and risk prediction[26]. In the field of autonomous driving, the Agent-Driver researched by Jiageng Mao et al. empowers AI agents with intuitive common sense and powerful reasoning capabilities[27]. In the field of education, Khan Academy has launched the AI teaching assistant Khanmigo, which not only provides subject counseling for students, but also provides real-time tracking and intelligent evaluation, and writes lesson plans and plans courses in the role of a teacher[28].

AI agent applications in healthcare

The exploration of AI agent applications in healthcare focuses on assisted diagnosis, decision making, report generation, chatbots, healthcare management, and medical education. Figure 3 provides a detailed illustration of the applications of AI agents in these field.

Assisted Diagnosis: Assisted diagnosis represents one of the most common applications of AI agents in healthcare. From a technical perspective, some studies have shown that multi-intelligence interactions can improve diagnostic accuracy and correct errors in historical records[29,30]. Therefore, researchers often leverage expert simulations, patient interaction, and multi-agent collaboration to enhance diagnostic performance. For instance, Tsinghua University built an agent hospital by simulating the actual scenarios of medical staff and patients in healthcare institutions, thus improving the intelligence in the interaction between doctors and patients[31]. Similarly, the assistant-driven expert consulting (AMSC) model from Harbin Institute of Technology simulates expert seminars through multiple intelligent bodies with diverse knowledge backgrounds[32]. ClinicalAgent employs specialized LLMs to provide tailored departmental support, closely aligning simulations with real-world clinical environments[33]. From the perspective of target areas of assisted diagnosis, in addition to general diagnostic assistance systems such as Stanford University's MMedAgent,

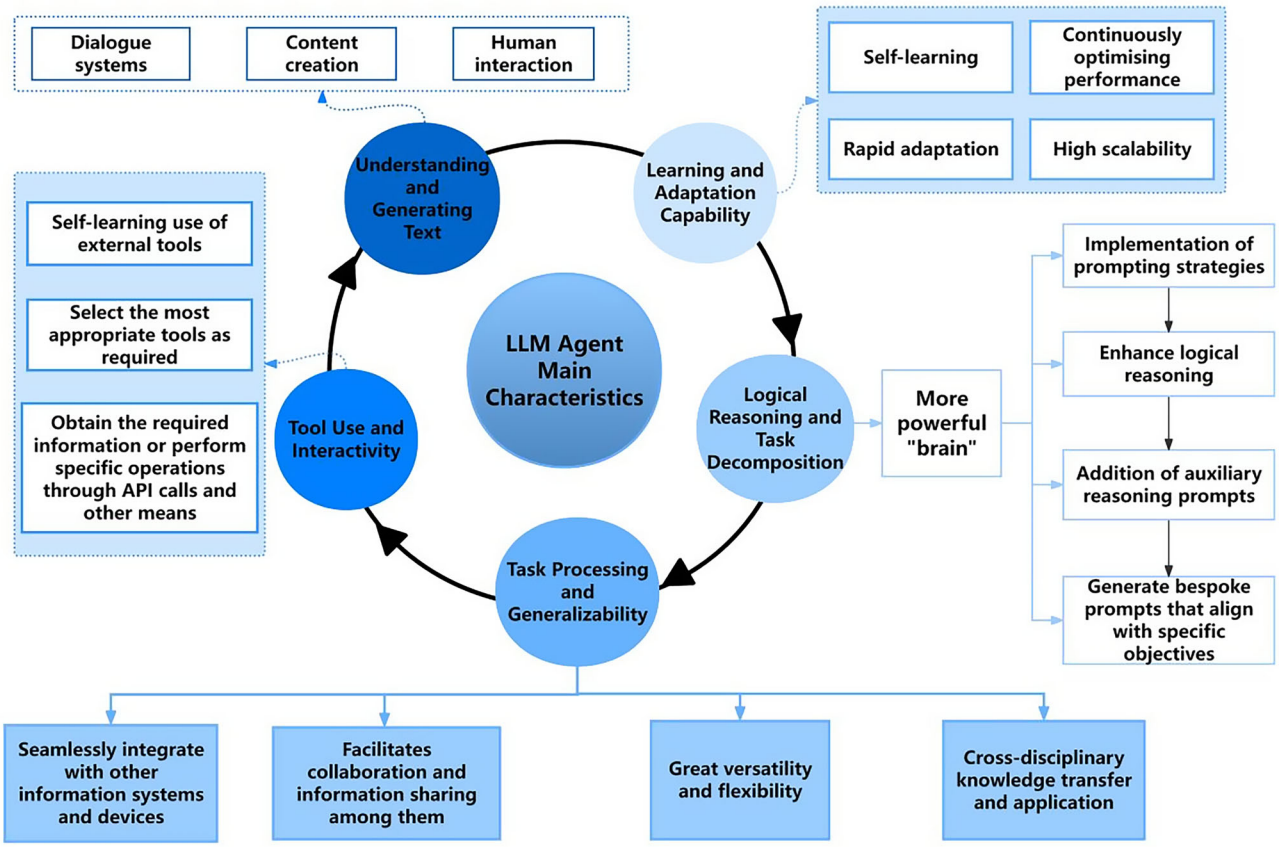

Fig. 2 | Characteristics of AI agent

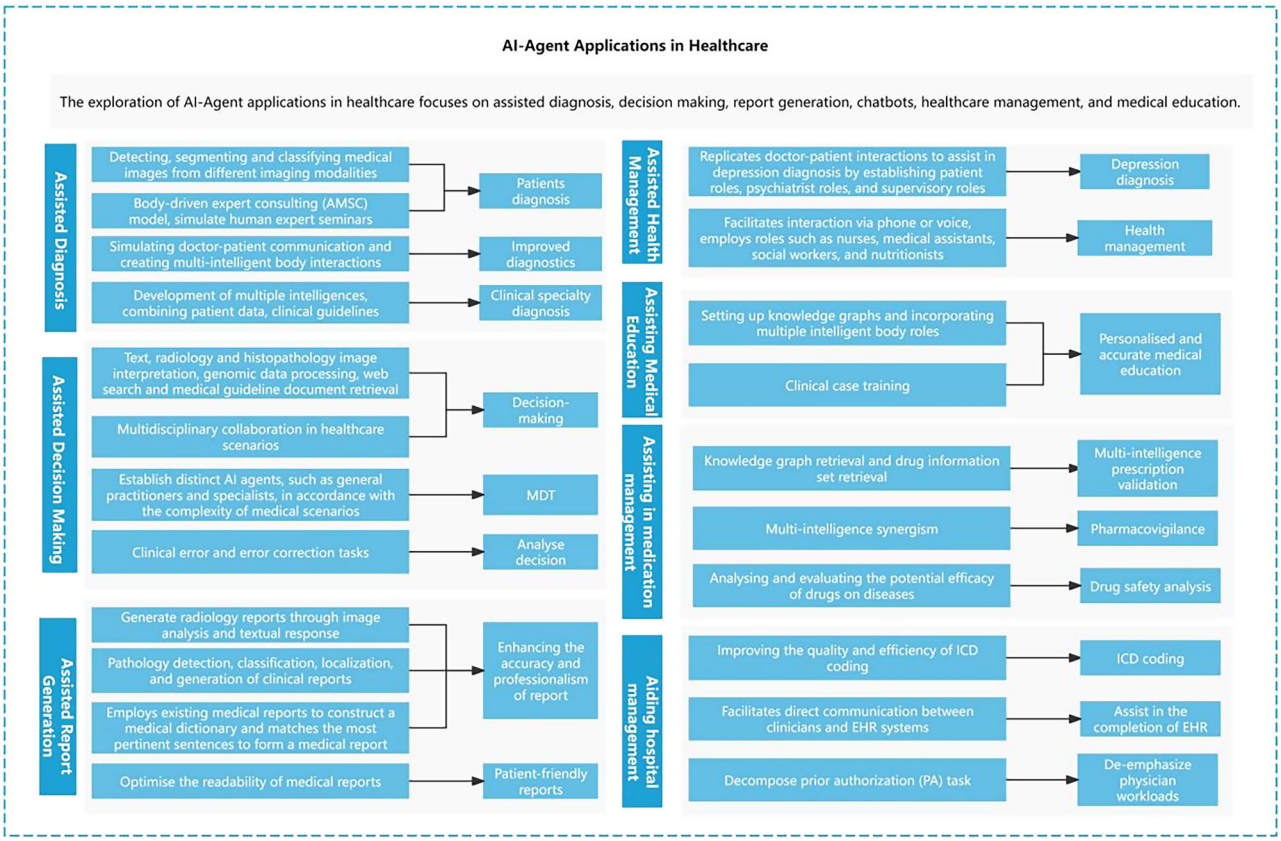

Fig. 3 | AI agent applications in healthcare

which utilizes multimodal imaging to detect, segment, and classify medical images[34], AI agents are increasingly applied to specialty domains. ZODIAC, developed for cardiology, extracts clinically relevant features, detects arrhythmias, and contributes to diagnostic decisions[35]. Baidu's AI agent could assist the ear deformities in newborns[36], and the MAGDA system integrates radiology images with clinical guidelines for enhanced reasoning[37].

Assisted decision-making: decision-making is another key area where AI agents have shown significant potential in healthcare. Similar to the field of assisted diagnosis, many medical scenarios involve multiple disciplines and roles, so researchers often integrate different data sources, establish distinct agents with complementary expertise, and enable role-based interactions, aiming to leverage multi-agent collaboration to enhance the quality, interpretability, and consensus of clinical decisions. For example, Yale's MedAgents employ role-playing and multidisciplinary discussion to iteratively improve credibility and interpretability, ultimately facilitating consensus decision-making[38]. MDAgents establish different agents, such as general practitioners and specialists, according to scenario complexity, supporting decision-making through structured Multidisciplinary Team (MDT) collaboration[39]. Similarly, MEDAIDE improves understanding of clinical intent via query rewriting, intent recognition, and multi-intelligence collaboration, enhancing decision-making effectiveness in complex situations[40]. From the perspective of application, assisted decision-making has also been applied in specialty domains. For example, in oncology treatment, the agent developed by the Heidelberg University Hospital for oncology scenarios are capable of text, radiology and histopathology image interpretation, genomic data processing, web search and medical guideline document retrieval[41]. In emergency care, a multi-intelligence system comprising emergency physicians, triage nurses, pharmacists, and dispatchers integrated the Emergency Triage Assessment Scale (ETAS) to improve quality, efficiency, and safety in decision-making[42]. Furthermore, there are multi-intelligentsia specialized in clinical error and error correction tasks that enable positive and negative analysis of medical decisions by breaking down the steps of observation, evaluation, reflection, and formatting[43].

Assisted Report Generation: Assisted report generation represents one of the earlier applications of AI agents in healthcare, with initial efforts primarily aimed at assisting radiologists in interpreting medical images and alleviating workforce shortages. For example, Stanford University's CheXagent focuses on the interpretation of chest X-rays and is able to generate radiology reports through image analysis and textual response, with its performance on visual tasks exceeding that of the generalized domain model by 97.5%[44]. Similarly, CXR-agent is another agent focusing on chest X-rays, capable of achieving pathology detection, classification, localization, and generation of clinical reports[45]. In later developments, research attention has expanded to improving report quality, accuracy, readability, and patient-centered communication. For example, MGA employs existing medical reports to construct a medical dictionary and matches the most pertinent sentences to form a medical report, thereby enhancing the accuracy and professionalism of report generation[46]. It should be noted that due to the specialized nature of report generation, previous studies have mostly focused on single-agent systems, which can ensure computational efficiency and semantic consistency. As more attention is paid to the doctor-patient experience, recent research has begun to incorporate multi-agent architectures to optimize the report generation process. This approach enables the production of patient-friendly reports, thereby reducing clinicians' workload while enhancing readability and improving the overall patient experience[47].

Assisted health management: assisted health management has emerged as a prominent direction for the application of AI agents in healthcare, with conversational agents being the predominant form. Conversational agents, referred to as chatbots, are able to interact with humans using natural language. As machine learning continues to evolve, conversational agents are beginning to emerge[48]. These are capable of processing more complex information, and thus are able to respond to health needs in a more personalized and precise manner[49]. Against this background, most studies in this area primarily focus on mental health, such as Agent Mental Clinic (AMC), is a conversational intelligent designed for depression diagnosis, which replicates doctor-patient interactions to assist in depression diagnosis by establishing patient roles, psychiatrist roles, and supervisory roles[50]. MISHA is targeted towards students, providing psychoeducation on stress management and relaxation techniques, among other topics, and facilitates alleviating students' perceivable stress[51]. The scope of research encompasses studies on topics such as alleviating suicidal thoughts[52] and reducing post-traumatic isolation[53]. Beyond the realm of mental health, research is also being conducted in the domain of weight loss counseling and skin management[54]. With regard to the modality of interaction, conversational agents are predominantly text-based, though Polaris[55], which facilitates interaction via phone or voice, employs roles such as nurses, medical assistants, social workers, and nutritionists to achieve health management functions, including medication adherence, appointment inquiries, and dietary adjustments.

Assisting medical education: medical education represents a further application scenario. Researchers often use multi-agent systems to simulate various roles, such as patients or teachers, based on real medical scenarios to create interactive scenarios to improve the abilities of medical students. For example, AI Patient, developed by the University of Michigan, is able to simulate patients, and by setting up knowledge graphs and incorporating multiple intelligent assistant roles such as retrieval, reasoning and generation, it can enhance the validity and credibility of the simulation and contribute to the education of medical students[56]. MEDCO, developed by the Chinese University of Hong Kong, emphasizes clinical case training to enhance the level of medical students by setting up patient simulations, feedback from senior doctors and experts, and multi-student interactions to provide more personalized and accurate medical education[57]. ChatCoach can help medical students improve their ability to communicate with patients by setting up roles such as doctors, patients, and coaches and simulating conversations about medical scenes[58]. In addition to the teaching of general medical knowledge, some studies have extended to specialized education, such as the LLM-based chatbots specifically designed for radiation oncology education have emerged as valuable tools for professional healthcare training, enhancing the accessibility, personalization, and interactivity of medical education[59].

Assisting in medication management: In the domain of drug management, researchers have explored applications in prescription management, adverse event prevention, and prediction of drug efficacy in clinical trials by simulating the processes of different stages. Correspondingly, Rx Strategist offers a multi-intelligence prescription validation concept, which facilitates indication and dosage validation through knowledge graph retrieval and drug information set retrieval[60]. MALADE prioritizes pharmacovigilance, enabling the identification of adverse drug reactions through the design of multi-intelligence synergism[61]. The ClinicalAgent system, a multi-intelligence system developed for clinical trials, enables analyzing and evaluating the potential efficacy of drugs on diseases, as well as carrying out drug safety analysis[62].

Aiding hospital management: In the domain of hospital management, it is important to reduce the burden on doctors, improve efficiency, and optimize processes. The intricate computer operations and management responsibilities of Electronic Health Records (EHRs, or Electronic Medical Records, EMRs) have been identified as contributing factors to the burden and burnout experienced by physicians[63]. Consequently, numerous researchers have directed their attention towards this area, endeavoring to devise solutions from the perspective of agents. EHRAgent facilitates direct communication between clinicians and EHR systems through autonomous code generation and execution, enhancing physician efficiency and experience[64]. Almanac Copilot can assist clinicians with EMR-related tasks by automating routine tasks and streamlining the documentation process[65]. ColaCare's approach centers on EHR modeling and clinical prediction, utilizing DoctorAgent and MetaAgent to emulate the collaborative decision-making process among doctors of diverse specialties. This facilitates enhanced clinical decision-making and the implementation of

personalized precision medicine[66]. In addition, there are researchers focusing on prior authorization (PA) to decompose this task by building a multi-intelligent assistant system to automate and de-emphasize physician workloads[67,68]. In terms of medical insurance, there is also study that has investigated the utilization of International Classification of Diseases (ICD) coding within the paradigm of multiple agents[61].

Furthermore, research is also being conducted in the field of biomedical knowledge, encompassing areas such as biological experiment design, cell biology, chemical biology, and genetics[69]. At the primary health care level, it involves the establishment of task-difficulty-assessment agents, expert agents, and response-simplification agents, as well as the incorporation of regional cultures and local languages to provide references for primary healthcare[70].

It is important to note that despite the broad potential of AI agents in the healthcare domain, their real-world implementation still faces several critical challenges: (1) Hallucinations. Diagnostic hallucinations may arise in the context of rare diseases or ambiguous clinical presentations, where the agent generates confident yet substantively incorrect conclusions, thereby posing clinical risks[71]. (2) Lack of interpretability. The decision-making processes of AI agents often lack transparency, making it difficult for clinicians to trace the underlying reasoning, which in turn undermines trust and limits adoption[72]. (3) Ambiguity in accountability. When AI agents generate diagnostic or therapeutic recommendations, the absence of clear definitions regarding legal and ethical responsibility in the event of erroneous outcomes remains a major challenge for clinical implementation and governance[73] (4) Data-related issues. On one hand, training datasets may exhibit imbalances across dimensions such as gender, ethnicity, and geography, resulting in performance degradation for specific populations and generating inequitable decisions that compromise health equity. On the other hand, the use of medical data involves highly sensitive personal information; in the absence of robust data governance frameworks and security safeguards, there is a heightened risk of privacy breaches and ethical violations[74,75].

In response to these challenges, the following sections will further explore a multi-dimensional evaluation framework designed to support the more scientific, robust, and trustworthy deployment of AI agents in healthcare.

Evaluation of AI agent in healthcare

As LLMs gain traction in healthcare, their potential to deliver clinical value depends critically on ensuring reliability, validity, and safety across every operational component. Without rigorous evaluation, AI agents may harbor latent flaws in medical reasoning that could translate into diagnostic inaccuracies or inappropriate treatment recommendations, thereby compromising patient safety. Even when designed for decision support, inadequately tested systems may generate ambiguous or inconsistent guidance, forcing clinicians to cross-check outputs and disrupting already burdened clinical workflows. Beyond these direct risks, insufficient evaluation also heightens concerns regarding bias, equity, and data privacy, all of which are crucial in sensitive healthcare environments. Against this backdrop, this section explores the evaluation subjects, comparison objects and dimensions and evaluation indicators of AI agents.

In the evaluation process of LLMs in the medical field, the evaluation subjects are typically divided into three categories. One such category is that of other LLMs, which are frequently employed, such as GPT–4/GPT-4o[76,33], Gemini-Pro[39]. These models enable analysing the intelligence of the medical LLM to be evaluated in terms of performance, functionality, and other relevant metrics. The second approach involves human evaluation[77], which involves inviting professionals from the relevant medical fields based on the type of intelligence required, including doctors of various disciplines, specialists[35], licensed nurses[55], clinical pharmacists[60], and radiology and imaging experts[78]. Clinical experts, drawing on their extensive professional knowledge and practical experience, assess the model's outputs, such as the rationality of diagnostic results and the viability of treatment plans, from the perspective of medical specialties. Their evaluation results embody authority and professionalism. Thirdly, the fair test sets[77], including MedQA, PubMedQA, MultiMedQA and other customized datasets for testing according to specific requirements[58,69,77]. The test set can provide a large number of standardized data samples, and the performance of the model on the test set allows for a more objective and quantitative evaluation of the model's level of competence in different tasks and knowledge domains.

The main comparison objects are baseline models and expert behavioral results when assessing LLMs in healthcare. The baseline models cover industry-leading LLMs, such as GPT-4/GPT-4o[56,79], Gemini-Pro[33,35], LLaMA[55,60], Mixtral[79], as well as models specialized in healthcare, like BioGPT[35] 、 Meditron[40], Med–Flamingo[34], and BioMistral[40]. These baseline models provide a frame of reference for the evaluation of the model under investigation. By comparing the model being evaluated with the baseline model on various performance metrics, it is possible to obtain a clear picture of its position and level in relation to similar models. This comparison can also reveal the uniqueness of the model or the areas that require improvement. Conversely, the expert behavior results focus on comparing the performance of the big language model intelligences with the diagnostic results, treatment decisions, and question-answering scores of human clinical experts[55]. For instance, in disease diagnosis tasks, the diagnostic consistency of the model's diagnostic results is compared with that of clinical experts; in treatment plan recommendation, the rationality and effectiveness of the plan given by the model is compared with that formulated by experts. By measuring the discrepancy or similarity between the LLMs and human experts in medical professional judgment and decision-making, the practical application value and effectiveness of the model in the medical field

Table 1 | Evaluation dimensions and indicators of AI agent in the healthcare

Dimension	Primary indicator	Representative metrics	Operational example (typical agent)
Basic indicators	Objective correctness	Accuracy, Precision, Recall, F1-score, ROC-AUC – be used to measure the correctness of the model's prediction results	ClinicalAgent[62], MedAgents[38]
	Semantic correctness	BLEU, ROUGE, METEOR, BERTScore -- be utilized to assess the semantic correctness of a model	MedAide[40], MedReAct'N'MedReFlex[43]
	Task completion	Completion rate, success rate, (tool use) -- be used as indicators to examine how well the model achieves a specific medical task	Agent for oncology[41], MMedAgent[34]
Developmental indicators	Efficiency level	Response time, number of interaction rounds -- be placed on the response time and the number of interaction rounds	Diaggpt[81], MDAgents[39]
	Content & presentation quality	Richness, usefulness, safety, ethical compliance, readability, coherence – ensure output content meets requirements in terms of text quality and content value	Polaris[55], CheXagent[44],
	Humanistic care	Humanistic care, confidence, adherence, satisfaction –assess the appropriateness of humanistic considerations and user acceptability in the interaction.	AgentClinic[79], Chat Ella[96]

The "Operational example" column lists typical AI agents for each evaluation dimension, but these agents are not limited to that dimension alone.

can be determined with greater accuracy, thus providing a clear goal and direction for the optimization and improvement of the model.

Multifaceted indicator dimensions and corresponding evaluation indicators are covered in Table 1, which can be specifically divided into two categories: basic indicators and development indicators. Existing studies demonstrate that quantitative metrics such as accuracy and F1-score remain the most commonly used measures, offering precise statistical evaluations of model performance. However, recent research has increasingly emphasized additional aspects, including efficiency, ethical compliance, and the patient–clinician interaction experience[20,80]. Collectively, these indicators reflect an evolutionary progression from basic feasibility to comprehensive excellence. Basic feasibility corresponds to the basic indicators, representing the minimum standards required to ensure the safe and effective delivery of healthcare services, including objective correctness, semantic correctness, task completion. Comprehensive excellence corresponds to the developmental indicators, reflecting the pursuit of high-quality, human-centered, and sustainable performance in complex clinical contexts, including efficiency level, content and presentation level and humanistic care. For detailed explanations of the indicators, please see Supplementary Information A.

Objective correctness: includes indicators such as accuracy, precision, recall, F1-score, ROC, AUC, which are used to measure the correctness of the model's prediction results. These metrics evaluate the extent to which the outcomes generated by AI agents are objectively consistent with verified medical facts, benchmark datasets, or other reference standards, thereby reflecting the quantitative reliability of the model across diverse healthcare tasks. For example, Yale University's MEDAGENTS primarily uses accuracy to evaluate the performance of models[38]. ClinicalAgent[62], which focuses on clinical trials, also assessed its outcomes using accuracy, ROC-AUC, precision, recall, and F1-score.

Semantic correctness: there are metrics such as BLEU/GLEU, METEOR, BERTScore and ROUGE that can be utilized to assess the semantic correctness of a model. These metrics ascertain the model's capacity to comprehend and articulate semantics by evaluating the degree of similarity between the text generated by the model and the reference text with respect to vocabulary and semantic structure. For instance, Dingkang Yang[40] validated the multi-dimensional health risk assessment capability of the medical agent by comparing its pre-diagnosis results with those of benchmark models using metrics such as BLEU-1/2 (%), ROUGE-1/2/L (%), and GLEU (%).

Task completion: the completion rate and success rate are used as indicators to examine how well the model achieves a specific medical task. In more complex agentic settings, task completion may involve the ability to autonomously select, invoke, and coordinate external tools to achieve a given objective, reflecting the agent's procedural reasoning and execution capability. For instance, the clinical decision-making agent developed by Heidelberg University Hospital for oncology leverages the accuracy in identifying and using tools, as well as the accuracy and correctness of tool usage as key evaluation metrics[41] Similarly, tool utilization serves as a primary criterion for assessing task completion in Stanford University's MMedAgent[34]

Efficiency level: the emphasis is placed on the response time and the number of interaction rounds, in order to evaluate the model's operational speed and the ease with which it can be interacted with. A reduced response time signifies that the model can respond to user inputs with greater alacrity, which can enhance service efficiency in scenarios such as medical consultation. The number of interaction rounds is indicative of the frequency with which the model must engage effectively with the user. A reduced number of interaction rounds signifies that the model is capable of comprehending the user's needs and furnishing accurate responses or solutions with greater alacrity. For instance, Lang Cao[81] employed the "number of turns" metric, defined as the average number of turns required to complete a task-oriented dialogue, to test the agent's dialogue quality across 20 scenarios. A lower number of turns signifies higher efficiency.

Content and presentation level: the overall readability, clarity, coherence and practical application value of the text information provided by the model were analyzed through the following indicators of content richness, detail, usefulness, safety, and ethical compliance. These metrics examine whether outputs are clinically meaningful, understandable, and ethically appropriate. High-quality content and presentation should convey sufficient detail and depth for clinical tasks while remaining comprehensible to both professionals and patients. For example, CheXagent[44] is an agent system for generating radiology reports. The research team not only assessed the reports' completeness, correctness, and conciseness, but also invited radiologists to evaluate the text quality. Furthermore, the study evaluated the potential for bias related to gender, race, and age to ensure fairness.

Humanistic care: Includes indicators such as accuracy under hidden symptoms, humanistic care, confidence, compliance, counseling and satisfaction, which focuses on the extent to which the model pays attention to the patient's psychological and healthcare service needs in medical situations. These metrics reflect the humanistic concept that emphasizes respect for patients' emotions, autonomy, and social context by medical AI agent systems. For instance, Samuel Schmidgall[79] noted that implicit biases among doctors can influence diagnostic judgments and treatment planning, while patients' biases affect trust and adherence. Accordingly, when evaluating AgentClinic, the team incorporated metrics related to doctor–patient interaction and doctor empathy to capture human-centered dimensions of care.

It is important to note that the two-tiered framework proposed in this paper provides conceptual indicators for evaluating AI agents. However, its application in real-world scenarios remains to be further explored. In fact, developing official and actionable evaluation systems still faces numerous challenges, and no mature and widely adopted framework currently exists. Nevertheless, during this transitional phase, regulatory explorations for the evaluation of AI + healthcare have gradually emerged. For example, the UK's MHRA "AI Airlock" sandbox mechanism is a regulatory sandbox designed to provide a controlled testing environment for AI medical devices, with evaluations emphasizing the following indicators: Safety/quality, effectiveness, adoption, equity/robustness[82]. Meanwhile, the EU's CORE-MD project proposes an evaluation framework primarily consisting of the following indicators: Valid clinical association score, valid technical performance score and clinical performance score[83]. Additionally, China's National Medical Products Administration has issued guidelines for the clinical evaluation and registration review of AI-assisted detection medical devices (software), proposing the following key evaluation metrics: diagnostic accuracy indicators, such as sensitivity, specificity, and area under the ROC curve, and clinical reference standard construction. These evaluation frameworks provide valuable references for assessing AI agents, and the practical experiences and indicator systems offer important lessons for the development of future evaluation systems.

Discussion

This section outlines seven future research directions:

First, integration with embodied robots. With the acceleration of global population aging and an increasing shortage of medical personnel, embodied robots (ERs), which possess a physical form and are capable of interacting directly with humans in real-world environments, have emerged as a promising solution to improve healthcare service delivery. Representative systems like the da Vinci Surgical System have demonstrated their core value in improving surgical precision and reducing medical risks, marking a shift in medical operations from purely manual procedures to human-robot collaboration paradigms[84]. Although ERs have achieved breakthroughs in specific medical fields, their large-scale adoption still faces challenges: most current systems rely heavily on pre-programmed routines, safety and fault-tolerance mechanisms remain inadequate for complex medical scenarios, and most critically, ER-patient interactions remain at a mechanical response level, falling short of achieving truly humanized doctor-patient communication. These bottlenecks have resulted in limited penetration rates of ERs

in healthcare applications. The ongoing development of LLMs projects to elevate the integration of AI agents and embodied robots to a new level. This integration will facilitate the provision of more direct medical services to patients through the utilization of physical robots, such as surgical assistance and healthcare[85]. This, in turn, will enable direct interaction between the AI agent and the patient in the real world, thereby achieving a more humanized and personalized healthcare service.

However, it should be noted that this integration raises important concerns related to safety, accountability, and patient privacy, and will require the healthcare system to adapt by establishing guidelines for ERs collaboration, strengthening oversight mechanisms, and ensuring compliance with ethical and regulatory frameworks.

Second, hybrid expert model combination. With the deepening application of AI technology in healthcare, intelligent agents increasingly process complex medical data, directly impacting diagnostic efficiency and patient safety. While large-scale pre-trained models offer strong generalization capabilities, they often exhibit limited precision, interpretability, and adaptability when confronted with specialized clinical tasks. Hybrid expert models, incorporating the Mixture of Experts (MoE) framework, have demonstrated substantial potential to address these limitations. By dynamically activating and integrating specialized sub-models for specific clinical tasks, MoE approaches enhance the reliability, interpretability, and domain relevance of medical AI systems[17].

Evidence of practical deployment has begun to emerge. For example, the early implementation of MoE-SLU in spoken medical consultation leveraged a mixture-of-experts model to perform weighted fusion of multiple ASR hypotheses, resulting in a 3.4–5.1 percentage point improvement in keyword-recognition accuracy across three benchmark datasets[86]. This demonstrates that MoE has progressed beyond theoretical conception and is being validated in professional medical AI systems. Overall, the incorporation of MoE offers a highly effective strategy to combine the strengths of each component, preserving efficiency and interpretability while improving decision-making reliability.

Third, expansion of evaluation metrics. A multi-dimensional evaluation tool helps comprehensively assess the true value of AI agents, enabling healthcare professionals, patients, and administrators to form realistic expectations. This is crucial for promoting the standardized application of AI technology in medical settings. However, the prevailing metrics predominantly emphasize conventional performance indicators such as accuracy, efficiency levels, and language expression proficiency, which fail to adequately reflect the comprehensive benefits of AI systems in real-world clinical settings—particularly critical factors such as economic costs and clinical safety. Moreover, they lack quantifiable assessment of user experience and human-centric considerations.

The prevailing evaluation metrics predominantly emphasize conventional performance indicators such as accuracy, efficiency levels, and language expression proficiency. It is imperative to acknowledge that in order to foster confidence among relevant user groups (including patients, medical professionals, and healthcare institutions) in the utilization of AI agents within the medical domain, there is a necessity for the development of a practical and comprehensive evaluation tool[87]. Consequently, there is an imminent necessity to extend the evaluation framework of intelligences in healthcare to encompass multidimensional factors, including economic indicators (e.g., cost-benefit analysis, return on investment, and long-term maintenance costs), safety indicators (e.g., incidence of adverse events), and patient satisfaction or other subjective metrics, in addition to technical performance.

Fourth, safety and risk management. As autonomy in the development of AI agents increases gradually, the application of artificial intelligence in the medical field is poised to undergo unparalleled breakthroughs. This technological breakthrough holds significant value for improving diagnostic and treatment efficiency and optimizing the allocation of medical resources. Nevertheless, this technological advancement is accompanied by numerous potential risks and challenges, which are primarily reflected in inadequate clinical trust and a lack of safety risk control mechanisms. A central concern pertains to the assurance of transparency and interpretability in the decision-making processes of these AI agent systems, and the mitigation of their potential impact on patient safety[88].

Consequently, there is an imminent necessity to formulate a comprehensive set of standardized and operational guidelines that encompass technical safety measures, whilst emphasizing the establishment of a comprehensive supervision framework that can perpetually monitor the performance of the AI agent system, address any potential deviations or errors, and ensure the system's capacity to respond to emergencies promptly.

Fifth, moral and ethical review. As LLM-based AI agents become increasingly embedded in clinical workflows, moral and ethical considerations will emerge as a crucial dimension of system design, deployment, and governance. Key among these issues are the protection of data privacy, the transparency of algorithms, and the attribution of responsibility[89]. These issues are directly related to the acceptability and sustainability of LLM-based AI agents[90]. For instance, previous collaborations between the UK NHS and Google DeepMind raised concerns over patient data privacy and consent, highlighting the potential ethical pitfalls when deploying AI in sensitive healthcare contexts.

At present, the corresponding regulatory frameworks and best practices are in a phase of active development and continuous iteration. The European Union's Artificial Intelligence Act classifies medical AI systems as "high-risk" and requires developers to provide detailed technical documentation and compliance proofs to ensure their safety and transparency. Meanwhile, legislative initiatives such as the U.S. Algorithmic Accountability Act are seeking to clarify the legal boundaries among developers, healthcare institutions, and regulatory bodies within the chain of responsibility. Against this backdrop, the introduction of an independent ethics committee mechanism helps review the ethical compliance of artificial intelligence in sensitive medical scenarios (such as end-of-life care), enhancing the acceptability of decisions and societal trust.

Nonetheless, ambiguity remains regarding the attribution of accountability among stakeholders, and LLM-based agents introduce unique challenges in explainability and autonomy, which must be appropriately addressed within these systems. To ensure the responsible evolution of LLM-based medical agents, it is imperative to integrate ethical review as a fundamental part of system development. This should include clear protocols for privacy protection, guidelines for transparent algorithmic decision-making, clinician oversight, mechanisms for assigning responsibility, patient privacy protection, and humanized care[85].

Sixth, user trust and feedback adoption. The wide application of AI agents not only relies on technological breakthroughs but also needs to address the trust and adoption issues of users, mainly patients, doctors and medical institutions. The establishment of a trust mechanism (Trustworthy AI agent) is a subject that merits attention in future research[91]. In practice, there are real-world challenges such as difficulty in balancing the demands of multiple stakeholders, the lack of an effective feedback loop, and user skepticism. These issues significantly constrain the clinical application value of AI technologies.

It is essential to comprehensively consider the needs and expectations of multiple stakeholders, including patients, doctors, and healthcare organizations, in order to enhance the acceptability of AI agents and the satisfaction of the user community[92]. Establishing a dynamic user feedback mechanism can help continuously optimize AI performance and ensure that its suggestions match real medical needs. Future research should place greater emphasis on multi-stakeholder collaboration, directly integrating user feedback into the AI development cycle to build an intelligent and trustworthy medical agency system.

Seventh, career development of medical staff. The question of whether robots can truly substitute for human medical staff in a career development context is a long-standing subject of considerable societal concern. The advent of sophisticated AI agent technology within the medical domain will likely wield a profound and pervasive influence on the psychological landscape and professional trajectories of medical personnel[93]. On the other hand, this will directly determine the quality and efficiency of future medical services.

However, concerns remain among some medical staff about being replaced by machines, which may lead to resistance or anxiety regarding AI adoption. In fact, from the perspective of socio-technical systems theory (STS)[94], the integration of AI agents represents not only a technological advancement but also a shift in how healthcare professionals interact with their working systems. Rather than replacing human roles, this transformation underscores the need for co-adaptation between technology and human practice. To facilitate this transition, healthcare systems should provide training opportunities, institutional support, and mechanisms for human–AI collaboration that ensure medical staff remain at the center of care delivery. Medical staff may encounter uncertainty as traditional skills are re-evaluated, but they also have opportunities to redefine their roles in ways that emphasize collaboration with intelligent systems, data-driven decision-making, and enhanced patient-centered care, in order to fully realize the benefits of AI-assisted healthcare[79,95].

To further strengthen the practical implications and real-world relevance of this review, we have included two supplementary components. Supplementary Information B presents a structured research agenda that identifies key research questions, theoretical lenses, and methodological directions to guide future inquiry into LLM-based AI agents in healthcare. Supplementary Information C provides a stakeholder value mapping that illustrates how specific AI agent functionalities generate tangible outcomes for key stakeholder groups, including developers, clinicians, hospital administrators, policymakers, and educators.

This review provides a systematic overview of the current state of research on AI agents based on LLM within the healthcare domain. We revisit their conceptual evolution and core characteristics, and summarize their major application areas, including diagnostic assistance, clinical decision-making, medical report generation, health management, medical education, pharmaceutical services, and hospital operations. We further propose a multidimensional evaluation framework that incorporates both technical performance and human-centered dimensions, and outline seven critical directions for future development. Positioned at the intersection of artificial intelligence and healthcare, this work contributes to clarifying the evolving role of AI agents in medical settings and offers theoretical grounding and practical guidance for their design, implementation, and governance. It should be noted that this review has certain limitations. Although we endeavored to comprehensively cover relevant literature, the rapidly evolving nature of the field may have led to omissions due to the time-lag in publication and indexing. Future studies may benefit from case-based validation and empirical investigations to further support the development of AI agents that are safe, controllable, and trustworthy in real-world healthcare environments.

Methods

Search strategy and selection criteria

To review the applications of AI agents in healthcare, and to ensure broad coverage of both medical domains and the frontier of agent technologies, we conducted a literature search across Web of Science, PubMed, and arXiv[80], focusing on studies published over the past five years, particularly from the emergence of LLMs in 2022 through February 2025. The search combined technical and domain-specific keywords. Technical terms included "agent", "agentic", "large language model", "foundation model", "natural language processing", "vision-language models", "multimodal large language models", "generative AI" and "GPT". Domain-specific terms included "healthcare", "medicine", "clinical", and "medical".

A total of 510 articles were retrieved across the three databases. Each article was independently screened by three research team members, with discrepancies resolved by a fourth member. Studies were excluded if they involved: (1) research about traditional agent; (2) non-English and incomplete publications; or (3) editorials, commentaries, letters, viewpoints, newspaper pieces, published errata or reviews. Following this process, 81 representative studies were included in the final synthesis. A visual summary of this process is provided in Supplementary Information D.

Data analysis

During the analysis phase, we conducted thematic categorization of the included literature to systematically present AI agent applications. The categories were determined based on the primary applications and functionalities reported in the literature, including assisted diagnosis, decision-making, report generation, health management, medical education, drug management, and hospital management. We note that this classification is inductive and content-driven, intended to summarize the current research focus and trends, rather than following a fixed, standardized framework.

Data availability

The analyzed data are included in this article. Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Received: 17 May 2025; Accepted: 23 January 2026;
Published online: 05 March 2026

References

1. Wang, D. & Zhang, S. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artif. Intell. Rev.* **57**, 299 (2024).
2. NewWhitepaper_Agents2.pdf. Google Docs. https://drive.google.com/file/d/1oEjiRCTbd54aSdB_eEe3UShxLBWK9xkt/view?pli=1 (n.d).
3. Qiu, J. et al. LLM-based agentic systems in medicine and healthcare. *Nat. Mach. Intell.* **6**, 1418–1420 (2024).
4. Karunanayake, N. Next-generation agentic AI for transforming healthcare. *Inform. Health* **2**, 73–83 (2025).
5. Wang, W. et al. A survey of LLM-based agents in medicine: how far are we from Baymax? Findings of the association for computational linguistics. ACL **2022**, 10345–10359 (2025).
6. Moritz, M., Topol, E. & Rajpurkar, P. Coordinated AI agents for advancing healthcare. *Nat. Biomed. Eng.* **9**, 432–438 (2025).
7. Yuan, H. Agentic large language models for healthcare: current progress and future opportunities. *Med. Adv.* **3**, 37–41 (2025).
8. ADIDI, D. T. Aristotle concept of telos and artificial intelligence (AI): exploring the relevance of classical philosophy to contemporary AI development. *Adv. J. Sci. Eng. Technol.* **9**, 5–14 (2024).
9. Tan, C. F. et al. The application of expert system: a review of research and applications. *ARPN J. Eng. Appl. Sci.* **11**, 2448–2453 (2016).
10. Jędrzejowicz, P. Machine learning and agents. In *Proc. KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, 2-15 (Springer Nature 2011).
11. Zhang, K., Yang, Z. & Başar, T. Multi-agent reinforcement learning: a selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control*. 321–384 (2021).
12. Wang, L. et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345 (2024).
13. Strickland, E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectr* **56**, 24–31 (2019).
14. Castelfranchi, C. Modelling social action for AI agents. *Artif. Intell.* **103**, 157–182 (1998).
15. Weng, L. LLM-powered autonomous agents. lilianweng. github. io, Jun 2023. URL https://lilianweng.github.io/posts/2023-06-23-agent (2023).
16. Parisi, A., Zhao, Y. & Fiedel, N. Talm: Tool augmented language models. Preprint at arXiv https://doi.org/10.48550/arXiv.2205.12255 (2022).
17. Schick, T. et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems* **36**, 68539–68551 (2023).
18. Qin, Y. et al. ToolLLM: facilitating large language models to master 16000+ real-world APIs. In *Proceedings of the 12th International Conference on Learning Representations*, 18267 (2024).

19. Huang, K. AI agents in healthcare. in *Agentic AI: Theories and Practices*, 303–321 (Springer Nature, 2025).
20. Chen, X. et al. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine* **5**, 151–163 (2025).
21. Kojima, T. et al. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **35**, 22199–22213 (2022).
22. Chen, J., Chen, P. & Wu, X. Generating Chinese event extraction method based on ChatGPT and prompt learning. *Appl. Sci.* **13**, 9500 (2023).
23. Lin, J. et al. AgentSims: an open-source sandbox for large language model evaluation. Preprint at arXiv https://doi.org/10.48550/arXiv.2308.04026 (2023).
24. Colas, C. et al. Augmenting autotelic agents with large language models. In *Proc. Conference on Lifelong Learning Agents*, 205–226 (PMLR, 2023).
25. Belanche, D., Casaló, L. V. & Flavián, C. Artificial Intelligence in FinTech: understanding robo-advisors adoption among customers. *Ind. Manag. Data Syst.* **119**, 1411–1430 (2019).
26. Yang, H. et al. FinRobot: an open-source AI agent platform for financial applications using large language models. Preprint at arXiv https://doi.org/10.48550/arXiv.2405.14767 (2024).
27. Mao, J. et al. A language agent for autonomous driving. In *Proceedings of the 2nd Conference on Language Modeling*, (2024).
28. Pang, S., Nol, E. & Heng, K. Generative AI as a personal tutor for English language learning: a review of benefits and concerns. International Journal of Changes in Education 2025, 1–10 (2025).
29. Ke, Y. et al. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *J. Med. Internet Res.* **26**, e59439 (2024).
30. Chang, J. J. & Chang, E. Y. SocraHealth: enhancing medical diagnosis and correcting historical records. In *Proc. 2023 International Conference on Computational Science and Computational Intelligence*, 1400–1405 (2023).
31. Li, J. et al. Agent hospital: a simulacrum of hospital with evolvable medical agents. Preprint at arXiv https://doi.org/10.48550/arXiv.2405.02957 (2024).
32. Wang, H. et al. Beyond direct diagnosis: LLM-based multi-specialist agent consultation for automatic diagnosis. Preprint at arXiv https://doi.org/10.48550/arXiv.2401.16107 (2024).
33. Yan, W. et al. ClinicalLab: aligning agents for multi-departmental clinical diagnostics in the real world. Preprint at arXiv https://doi.org/10.48550/arXiv.2406.13890 (2024).
34. Lin, Z. & Wang, Y. MMedAgent: learning to use medical tools with multi-modal agent. In *Proc. Findings of the Association for Computational Linguistics: EMNLP 2024*, 8745–8760 (Association for Computational Linguistics, 2024).
35. Zhou, Y. et al. Zodiac: a cardiologist-level LLM framework for multi-agent diagnostics. Preprint at arXiv https://doi.org/10.48550/arXiv.2410.02026 (2024).
36. Wang, S. et al. Large language model-enhanced interactive agent for public education on newborn auricular deformities. Preprint at arXiv https://doi.org/10.48550/arXiv.2409.12984 (2024).
37. Bani-Harouni, D., Navab, N. & Keicher, M. MAGDA: multi-agent guideline-driven diagnostic assistance. In *Proc. International Workshop on Foundation Models for General Medical AI*, 163–172 (Springer, 2024).
38. Tang, X. et al. MedAgents: large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 599–621 (2024).
39. Kim, Y. et al. MDAgents: an adaptive collaboration of LLMs for medical decision-making. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems* Vol, **37** 79410–79452 (2024).
40. Yang, D. et al. MedAide: information fusion and anatomy of medical intents via LLM-based agent collaboration. *Inf. Fus.* **127**, 103743 (2026).
41. Ferber, D. et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer* **6**, 1337–1349 (2025).
42. Han, S. & Choi, W. Development of a large language model-based multi-agent clinical decision support system for Korean Triage and Acuity Scale (KTAS)-based triage and treatment planning in emergency departments. Preprint at arXiv https://doi.org/10.48550/arXiv.2408.07531 (2024).
43. Corbeil, J.-P. IryoNLP at MEDIQA-CORR 2024: tackling the medical error detection & correction task on the shoulders of medical agents. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 570–580 (2024).
44. Chen, Z. et al. Chexagent: towards a foundation model for chest X-ray interpretation. Preprint at arXiv https://doi.org/10.48550/arXiv.2401.12208 (2024).
45. Sharma, N. CXR-Agent: vision-language models for chest X-ray interpretation with uncertainty aware radiology reporting. Preprint at arXiv https://doi.org/10.48550/arXiv.2407.08811 (2024).
46. Huang, W. et al. MGA: medical generalist agent through text-guided knowledge transformation. Preprint at arXiv https://doi.org/10.48550/arXiv.2303.08562 (2023).
47. Sudarshan, M. et al. Agentic LLM workflows for generating patient-friendly medical reports. Preprint at arXiv https://doi.org/10.48550/arXiv.2408.01112 (2024).
48. Ma, Z., Mei, Y. & Su, Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings 2023*, 1105–1114 (2024).
49. Li, H. et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit. Med.* **6**, 23 (2023).
50. Lan, K.et al. Depression diagnosis dialogue simulation: self-improving psychiatrist with tertiary memory. Preprint at arXiv https://doi.org/10.48550/arXiv.2409.15084 (2024).
51. Ulrich, S. et al. A chatbot-delivered stress management coaching for students (MISHA app): pilot randomized controlled trial. *JMIR mHealth uHealth* **12**, e54945 (2024).
52. Maples, B. et al. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Ment. Health Res* **3**, 4 (2024).
53. Chou, Y. et al. User-friendly chatbot to mitigate the psychological stress of older adults during the COVID-19 pandemic: development and usability study. *JMIR Form. Res.* **8**, e49462 (2024).
54. Kotov, A., Carcone, A. I. & Towner, E. Neural conversational agent for weight loss counseling: protocol for an implementation and feasibility study. *JMIR Res. Protoc.* **13**, e60361 (2024).
55. Mukherjee, S. et al. Polaris: a safety-focused llm constellation architecture for healthcare. Preprint at arXiv https://doi.org/10.48550/arXiv.2403.13313 (2024).
56. Yu, H. et al. Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Communications Medicine* **6**, 27 (2026).
57. Wei, H. et al. MEDCO: medical education copilots based on a multi-agent framework. In *Computer Vision – ECCV 2024 Workshops* Vol. 15630, 125–142 (2025).
58. Huang, H. et al. Benchmarking large language models on communicative medical coaching: a dataset and a novel system. In *Proc. Findings of the Association for Computational Linguistics ACL 2024*, 1624–1637 (ACL, 2024).
59. Chow, J. C. L. & Li, K. Developing effective frameworks for large language model–based Medical Chatbots: Insights from Radiotherapy Education with ChatGPT. *JMIR Cancer* **11**, e66633 (2025).

60. Phan Van, P. et al. Rx Strategist: prescription verification using LLM Agents System. Preprint at arXiv https://doi.org/10.48550/arXiv.2409.14924 (2024).
61. Choi, J. et al. MALADE: orchestration of LLM-powered agents with retrieval augmented generation for pharmacovigilance. In *Proceedings of the Machine Learning for Healthcare 2024* Vol. 252, 1–51 (2024).
62. Yue, L. et al. Clinicalagent: clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–10 (2024).
63. Gebreab, S. A. et al. LLM-Based Framework for Administrative Task Automation in Healthcare. In *Proc. 12th International Symposium on Digital Forensics and Security (ISDFS)* 1–7 (IEEE, 2024).
64. Shi, W. et al. EHRAgent: code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22315–22339 (2024).
65. Zakka, C. et al. Almanac Copilot: towards autonomous electronic health record navigation. https://doi.org/10.21203/rs.3.rs-6102516/v1.
66. Wang, Z. et al. ColaCare: enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, 2250–2261 (2025).
67. Pandey, H. G., Amod, A. & Kumar, S. Advancing healthcare automation: multi-agent system for medical necessity justification. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 39–49 (2024).
68. Roohani, Y. et al. BioDiscoveryAgent: an AI agent for designing genetic perturbation experiments. In *Proceedings of the 13th International Conference on Learning Representations*, Vol. 30240 (2025).
69. Gangavarapu, A. & Gangavarapu, A. IMAS: a comprehensive agentic approach to rural healthcare delivery. Preprint at arXiv https://doi.org/10.48550/arXiv.2410.12868 (2024).
70. Farquhar, S. et al. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
71. Liao, Q. V. & Vaughan, J. W. *AI transparency in the age of LLMs: a human-centered research roadmap* (Harvard Data Science Review, 2024).
72. Jung, K. H. Large language models in medicine: clinical applications, technical challenges, and ethical considerations. *Healthc. Inform. Res.* **31**, 114–124 (2025).
73. Mirzaei, T., Amini, L. & Esmaeilzadeh, P. Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. *BMC Med. Inform. Decis. Mak.* **24**, 250 (2024).
74. Ong, J. C. L. et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit. Health* **6**, e428–e432 (2024).
75. Pahune, S. et al. The importance of AI data governance in large language models. *Big Data Cogn. Comput* **9**, 147 (2025).
76. Li, B. et al. MMedAgent: learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP* Vol. 2024, 8745–8760 (2024).
77. Mehandru, N. et al. Evaluating large language models as agents in the clinic. *NPJ Digit. Med.* **7**, 84 (2024).
78. Calisto, F. M. et al. BreastScreening-AI: evaluating medical intelligent agents for human-AI interactions. *Artif. Intell. Med.* **127**, 102285 (2022).
79. Schmidgall S. et al. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments[J]. Preprint at ArXiv https://doi.org/10.48550/arXiv.2405.07960.
80. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit. Med.* **7**, 258 (2024).
81. Cao, L. Diaggpt: an llm-based chatbot with automatic topic management for task-oriented dialogue. Preprint at arXiv https://doi.org/10.48550/arXiv.2308.08043 (2023).
82. Medicines and Healthcare products Regulatory Agency. (2025, August 8). AI Airlock: the regulatory sandbox for AIaMD. GOV.UK. https://www.gov.uk/government/collections/ai-airlock-the-regulatory-sandbox-for-aiamd.
83. Cancela-Outeda, C. The EU's AI act: a framework for collaborative governance. *Internet of Things* **27**, 101291, https://doi.org/10.1016/j.iot.2024.101291 (2024).
84. Wanjari, M. et al. The role of robotics in modern neurosurgery: current trends and future prospects. *Neurosurg. Rev.* **47**, 619 (2024).
85. Pashangpour, S. & Nejat, G. The future of intelligent healthcare: a systematic analysis and discussion on the integration and impact of robots using large language models for healthcare. *Robotics* **13**, 112 (2024).
86. Mu, S & Sen L. "A comprehensive survey of mixture-of-experts: algorithms, theory, and applications." Preprint at https://doi.org/10.48550/arXiv.2503.07137 (2025).
87. Reddy, S. et al. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health & Care Informatics* **28**, e100444 (2021).
88. Gao, S. et al. Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151 (2024).
89. Fiske, A., Henningsen, P. & Buyx, A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *J. Med. Internet Res.* **21**, e13216 (2019).
90. W.H.O. *Ethics and governance of artificial intelligence for health* (World Health Organization, 2021).
91. Marcus, G. & Davis, E. *Rebooting AI: building artificial intelligence we can trust* (Vintage, 2019).
92. Kaur, D. et al. Trustworthy artificial intelligence: a review. *ACM Comput. Surv.* **55**, 1–38 (2022).
93. Shuaib, A., Arian, H. & Shuaib, A. The increasing role of artificial intelligence in health care: will robots replace doctors in the future? *International Journal of General Medicine* **13**, 891–896 (2020).
94. Carayon, P. et al. Socio-technical systems analysis in health care: a research agenda. *IIE Trans. Healthc. Syst. Eng.* **1**, 145–160 (2011).
95. Ahuja, A. S. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ* **7**, e7702 (2019).
96. Zhang, S. & Song, J. A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model. *Sci. Rep.* **14**, 17118 (2024).

Acknowledgements

This study received no funding. We acknowledge the use of AI (specifically DeepL and Grammarly) for linguistic polishing and editing of this manuscript. Its use was exclusively limited to refining language, grammar, and clarity. The core ideas, data analysis, results, and intellectual content remain entirely our own.

Author contributions

L.Z., S.L., T.X., and J.T. drafted the manuscript and contributed to the conceptualization and writing of the review. X.W. prepared Figs 1–3 and assisted with visualization. Y.L., Z.B., Y.C., and F.K. contributed to literature collection, data cross-checking, and provided critical comments during manuscript revision. J.B. assisted with methodological refinement and technical review. C.Q. supervised the study design and provided guidance throughout the development of the manuscript. Z.Z. oversaw the overall research framework, ensured the coherence of the review, and provided senior academic supervision. All authors discussed the results, contributed to the final revision, and approved the final version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44387-026-00076-4.

Correspondence and requests for materials should be addressed to Chen Qian or Zongjiu Zhang.