A Survey of LLM-based Agents in Medicine: How far are we from Baymax?

Wenxuan Wang[1*] **Zizhan Ma**[2*] **Zheng Wang**[2] **Chenghan Wu**[2]
Jiaming Ji[3] **Wenting Chen**[4†] **Xiang Li**[5] **Yixuan Yuan**[2]
[1]Renmin University of China [2]The Chinese University of Hong Kong
[3] Peking University [4] City University of Hong Kong
[5] Massachusetts General Hospital and Harvard Medical School
[1]wangwenxuan@ruc.edu.cn [2]zzma2@cse.cuhk.edu.hk [3]wentichen7-c@my.cityu.edu.hk

Abstract

Large Language Models (LLMs) are transforming healthcare through the development of LLM-based agents that can understand, reason about, and assist with medical tasks. This survey provides a comprehensive review of LLM-based agents in medicine, examining their architectures, applications, and challenges. We analyze the key components of medical agent systems, including system profiles, clinical planning mechanisms, medical reasoning frameworks, and external capacity enhancement. The survey covers major application scenarios such as clinical decision support, medical documentation, training simulations, and healthcare service optimization. We discuss evaluation frameworks and metrics used to assess these agents' performance in healthcare settings. While LLM-based agents show promise in enhancing healthcare delivery, several challenges remain, including hallucination management, multimodal integration, implementation barriers, and ethical considerations. The survey concludes by highlighting future research directions, including advances in medical reasoning inspired by recent developments in LLM architectures, integration with physical systems, and improvements in training simulations. This work provides researchers and practitioners with a structured overview of the current state and future prospects of LLM-based agents in medicine.

1 Introduction

Large Language Models (LLMs) are changing the field of artificial intelligence with their strong capabilities in text understanding, generation, and reasoning. The development of LLM-based agents has achieved notable success in many areas, from creative writing (Yuan et al., 2022) to complex decision-making (Chai et al., 2025; Wei et al., 2023), which opens new opportunities for automating and enhancing human expertise. These agents have been applied in various fields by using the ability of LLMs to process and analyze complex information (Wang et al., 2024a; Cheng et al., 2024; Xi et al., 2023).

In the medical domain, LLM-based agents have improved several clinical tasks. Recent work shows their use in diagnostic support (Kim et al., 2024a), patient communication (Mukherjee et al., 2024), and medical education (Yu et al., 2024). By combining LLMs with medical knowledge bases, clinical guidelines, and healthcare systems, these agents are designed to understand complex medical situations (Wei et al., 2024b), offer evidence-based recommendations (Tang et al., 2024), and support healthcare delivery (Mukherjee et al., 2024). While clinical tasks are our primary focus due to their prominence in current LLM-based agent research, we also survey applications in data analytics, training, service optimization, and emerging non-clinical domains to reflect their broader medical potential. Despite these advances, the field still faces several challenges, including implementation issues (SUN et al., 2024), safety concerns (Yuan et al., 2024), and ethical considerations (YAN et al., 2025). Addressing these challenges is essential for the safe and reliable integration of LLM-based agents into clinical practice. Therefore, a comprehensive review is needed to analyze the current status and future directions of LLM-based agents in medicine.

In this article, we provide a systematic review of LLM-based agents in medicine, examining important research questions and future directions. We first discuss the architectures and methods, including system profile, external capacity enhancement, clinical planning, and medical reasoning in Section 3. Section 4 covers the various clinical and administrative application scenarios in which these agents are used. Section 5 outlines the evaluation

[*] Wenxuan Wang and Zizhan Ma are equal contribute to this paper.
[†] Wenting Chen is the corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 10345–10359
July 27 - August 1, 2025

frameworks and metrics for assessing their performance in healthcare settings. Finally, Section 6 highlights key challenges and future research directions for improving the reliability, safety, and clinical integration of LLM-based agents.

This review analyzed 60 studies on LLM-based medical agents published between 2022-2024, selected from major databases using healthcare AI-related keywords. The initial search yielded 300 papers, narrowed to 80 after screening, with 60 meeting final inclusion criteria.

2 Background

This section outlines the core differences between **LLMs** and **LLM-based agents** and highlights the unique considerations required for deploying such agents in medicine.

2.1 LLM vs. LLM-based Agent

An agent, as defined in AI, perceives its environment and takes actions accordingly (Russell and Norvig, 2016). An LLM-based agent extends traditional LLMs by integrating external knowledge retrieval, task planning, and tool invocation, enabling structured decision-making in real-world applications (Xi et al., 2023). Unlike standard LLMs, which primarily process text, these agents operate autonomously and adapt dynamically to new information and tasks.

While LLMs primarily function as static models for text generation and analysis, LLM-based agents are interactive systems that actively engage with their environment through perception, memory, and action capabilities. Existing surveys (Zheng et al., 2024; Park et al., 2024) have focused on LLMs' technological applications and clinical validation. However, as LLM-based agents increasingly take on complex medical tasks by dynamically interacting with their surroundings, a systematic synthesis of their advancements remains lacking. This survey addresses this gap by providing a comprehensive perspective on medical agent technologies.

2.2 Unique Considerations for LLM-based Agents in Medicine

Deploying LLM-based agents in healthcare requires addressing several critical factors:

Multimodal Integration. Medical data spans text, imaging, and laboratory results. Agents must process and synthesize these inputs for accurate decision support (Zhang et al., 2024).

Clinical Collaboration. Healthcare relies on interdisciplinary work. Agents should facilitate information sharing and human-AI collaboration, ensuring physicians maintain oversight (Strong et al., 2024).

Accuracy and Reliability. Given the impact on patient outcomes, these agents must meet strict validation standards and minimize errors in diagnosis and treatment (Reddy, 2024).

Transparency and Traceability. Clinical decisions must be auditable and explainable to align with medical ethics and regulatory requirements (Kiseleva et al., 2022).

3 LLM-based Medical Agent Architecture

LLM-based agents in medicine require well-defined architectures to integrate complex clinical knowledge, facilitate medical decision-making, and ensure safe and effective deployment. This section presents a systematic overview of their architectural components, focusing on how these agents structure their operations to enhance clinical performance.

3.1 Profile

The profile of an agent plays a key role in defining and managing its role attributes, behavioral patterns, and operational competencies within medical systems, which traditionally involve information dissemination, resource distribution, and quality assurance. In medical applications, agent profiles follow three prototypes:

Functional Modularization. This approach structures the agent system into specialized functional modules, each responsible for distinct tasks such as clinical data analysis or diagnostic reasoning. Systems like *MEGDA* (Bani-Harouni et al., 2024) implement function-driven profiles where task assignments and workflows are explicitly defined to improve efficiency and adaptability.

Role Specialization By mirroring real-world medical roles, this paradigm assigns agents to specific clinical functions, including diagnosis, medical imaging, treatment planning, and surgical assistance. These agents incorporate domain-specific knowledge and interact with healthcare systems for tasks such as imaging analysis and interdisciplinary coordination. In agent-driven operating room simulations (Wu et al., 2024), LLM-based agents take on distinct medical roles to support clinical decision-making.

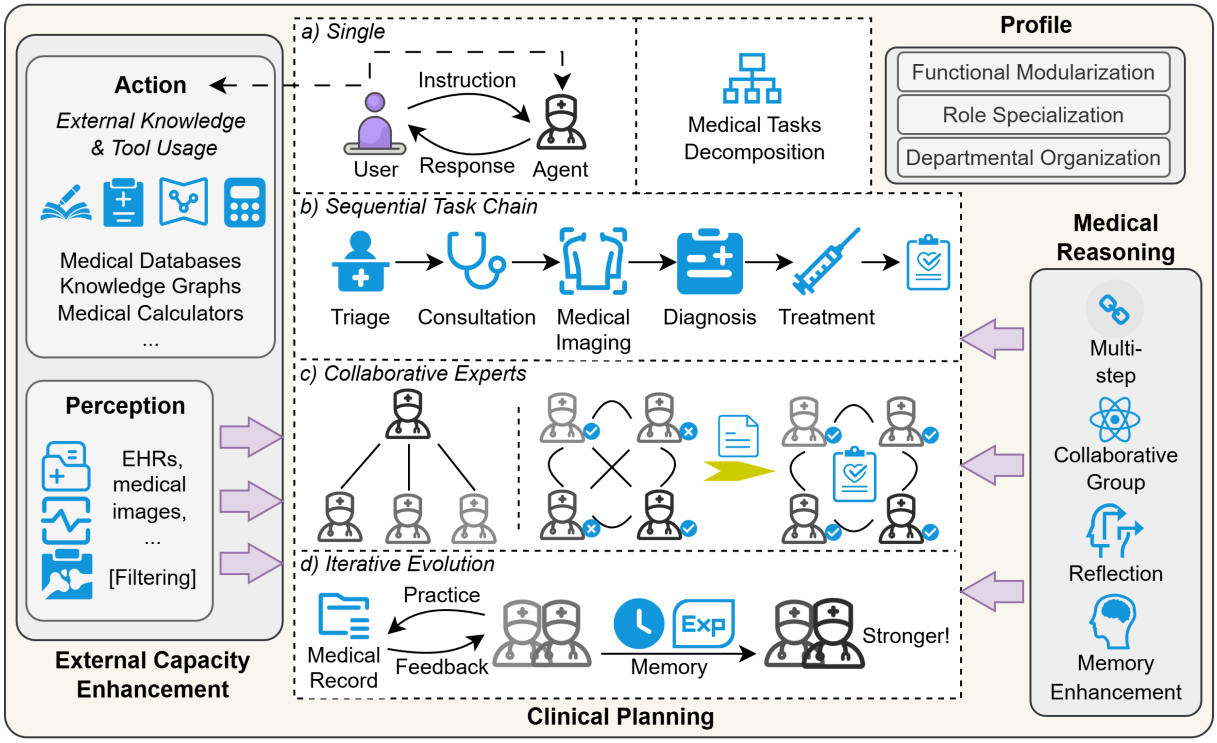

Figure 1: Conceptual framework of LLM-based medical agents. This figure depicts the architecture of the proposed LLM-based Medical Agent, consisting of system profile, external capacity enhancement, clinical palnning and medical reasoning. It supports four agent paradigms: **a) Single Agent**, **b) Sequential Task Chain**, **c) Collaborative Experts**, and **d) Iterative Evolution**. The framework integrates external tools and reasoning mechanisms to enable applications in medicine.

Departmental Organization This framework structures agents based on medical disciplines, such as cardiology or hematology, establishing domain-specific knowledge boundaries. Agents rely on specialized disease knowledge graphs and dynamic collaboration mechanisms to facilitate interdisciplinary consultations. In multi-agent medical applications (Tang et al., 2024), profiles are defined to reflect departmental expertise, improving coordination and decision-making in complex medical scenarios.

3.2 Clinical Planning

Effective clinical planning is at the core of LLM-based medical agents. The planning process breaks down complex medical tasks into smaller subtasks so that the system can interact with tools and databases specific to each clinical area (Mehandru et al., 2024). This division of tasks improves operational efficiency and aids in locating and correcting errors.

The Single Agent paradigm offers simplicity and lower computational overhead but limits capacity for complex tasks. The Sequential Task Chain paradigm enhances task specialization and accuracy but risks error propagation across steps. The Collaborative Experts paradigm leverages diverse domain knowledge and reduces biases but introduces coordination complexity and higher computational costs. The Iterative Evolution paradigm enables continuous learning and adaptation but requires robust feedback mechanisms to prevent error reinforcement.

Task Decomposition Clinical planning often follows a structured decomposition from high-level objectives to specific actions. A *Single-Agent* model handles tasks autonomously, while a *Sequential Task Chain* approach structures planning into distinct steps, such as data ingestion, hypothesis generation, treatment planning, and risk assessment. Each step interacts with specialized medical tools, ensuring task separation and facilitating precise error correction (Liu et al., 2024a).

Multi-Agent Collaboration Across Departments For complex cases requiring interdisciplinary expertise, a *Collaborative Experts* model assigns specialized agents to clinical areas such as radiology, pathology, and laboratory analysis. These agents

communicate using standardized protocols to aggregate findings and refine diagnoses. This reduces diagnostic uncertainty by integrating insights from multiple specialties (Tang et al., 2024).

Adaptive Planning Architecture A dynamic *Sequential Task Chain* or *Collaborative Experts* framework adjusts decision-making based on real-time data and task complexity. For example, MDAgents framework (Kim et al., 2024a) employs LLMs with predefined medical roles, which function autonomously or in coordination. Planning layer continuously updates clinical strategies, prioritizes urgent cases, and refines past decisions based on new evidence. Federated learning mechanisms further enhance adaptability by integrating diverse clinical experiences (Dutta and Hsiao, 2024).

Iterative Self-Evolution Beyond static workflows, *Iterative Evolution* frameworks enable continuous improvement. These systems maintain an experience base of past cases, refining decision-making over time. Self-improvement mechanisms allow agents to autonomously incorporate new medical data and learn from previous outcomes, progressively enhancing accuracy and reliability (Li et al., 2024b; Du et al., 2024).

3.3 Medical Reasoning

The Medical Reasoning module enhances diagnostic accuracy and transparency by structuring logical inference processes and integrating real-time feedback.

Multi-Step Diagnostic Reasoning Complex cases are analyzed through sequential inference, where Chain-of-Thought methods (Wei et al., 2023) generate step-by-step reasoning, and Tree-of-Thought approaches (Yao et al., 2023a) explore multiple hypotheses in parallel, discarding less probable options. This structured approach improves diagnostic precision (Dutta and Hsiao, 2024).

Reflective Decision-Making To handle clinical uncertainty, the system iteratively refines conclusions by incorporating real-time feedback and expert input. Inspired by the ReAct framework, it alternates between reasoning and action, identifying inconsistencies and improving decision robustness (Yao et al., 2023b; Yue et al., 2024).

Collaborative Group Reasoning A multi-agent reasoning framework assigns specialized agents—such as primary care providers and specialists—to perform independent analyses. Their conclusions are aggregated through consensus mechanisms, mitigating biases and enhancing reliability (Zuo et al., 2025).

Memory-Enhanced Reasoning Integrating long-term memory modules enables agents to accumulate medical knowledge and past clinical experiences, refining decision-making over time. This persistent memory allows the system to adapt to new medical insights, improve reasoning capabilities, and maintain continuity in patient care (Li et al., 2024b). Additionally, experience-based learning mechanisms enable LLM-based agents to update their diagnostic strategies dynamically, leading to more context-aware and personalized medical insights (Jiang et al., 2024).

3.4 External Capacity Enhancement

The external capacity enhancement augments the agent's capabilities by integrating it with real-world clinical data sources and specialized tools.

Perception subsystem processes diverse clinical inputs, including structured electronic health records (EHRs) to access patient histories and clinical parameters. Advanced Optical Character Recognition (OCR) techniques convert scanned documents into text, while models like CLIP analyze medical images, facilitating comprehensive multimodal understanding.

Knowledge Integration connects the agent with external sources such as medical knowledge graphs, drug interaction databases, and clinical guideline repositories. This connection helps the agent verify its inferences with trusted sources, thereby increasing both its accuracy and reliability (Li et al., 2024a; Huang et al., 2024a).

Action layer allows agents to perform clinical tasks by using specialized tools such as medical calculators, electronic health record interfaces, and image analysis software. The system calls additional functionalities when processing complex data, ensuring that its outputs are complete and take the context into account (Shi et al., 2024; Zhu et al., 2024).

4 Application Scenarios

LLM-based agents are applied in various areas of medicine. This section outlines the main application scenarios and provides a summary in Table 1.

4.1 Clinical Decision Support and Diagnosis

In the area of Clinical Decision Support and Diagnosis, multi-agent frameworks based on LLMs improve clinical decision-making by addressing limitations of standalone LLMs. Systems in this

Table 1: Summary of application of LLM-Agents in the medical field.

Purpose	Functionality		Work	Framework Type	Tool Use
Clinical Decision Support and Diagnosis	Refine Diagnostic Reasoning		(Dutta and Hsiao, 2024)	Adaptive Planning	-
	Reduce Cognitive Bias		(Ke et al., 2024)	Collaborative Experts	-
	Task Coordination		(Wei et al., 2024b)	Sequential Task Chain	-
	Diagnosis Accuracy		(Kim et al., 2024c)	Collaborative Experts	Yes
			(Tang et al., 2024)	Collaborative Experts	-
	Domain Specific Functionalities	Clinical Trial Outcome Prediction	(Yue et al., 2024)	Sequential Task Chain	Yes
		Patient Interaction Safety	(Mukherjee et al., 2024)	Sequential Task Chain	-
		Prescription Validation	(Van et al., 2024)	Sequential Task Chain	Yes
	Diagnosis Capability		(Yan et al., 2024)	Collaborative Experts	-
	Integrated Modelling		(Fan et al., 2024)	–	-
Clinical Data Analysis and Documentation	Mortality Prediction		(Wang et al., 2024b)	Collaborative Experts	Yes
	Hospital Readmission Analysis				
	Clinical Documentation		(Lee et al., 2024)	Single Agent	-
	Patient Friendly Medical Reports		(Sudarshan et al., 2024)	Iterative Evolution	Yes
	Integrated Simulation		(Li et al., 2024b)	Iterative Evolution	-
Medical Training and Simulation	Evaluated Diagnosis and Treatment Performance		(Yan et al., 2024)	Collaborative Experts	-
	Integrated Simulation		(Fan et al., 2024)	–	-
			(Li et al., 2024b)	Iterative Evolution	-
	Medical Training	Training Environment	(Wei et al., 2024a)	Collaborative Experts	Yes
			(Wu et al., 2024)	Collaborative Experts	-
		Scenario Simulation	(Yu et al., 2024)	Collaborative Experts	Yes
Healthcare Service Optimization	Automation of Non-diagnostic Tasks		(Mukherjee et al., 2024)	Sequential Task Chain	Yes
			(Laymouna et al., 2024)	–	-
	Automation of Diagnostic Tasks		(Chadebecq et al., 2023)	Iterative Evolution	-

area assign specialized roles to agents for intent recognition, diagnostic reasoning, and treatment planning so that healthcare delivery can be both personalized and sensitive to the context. For example, the framework proposed by Dutta and Hsiao (Dutta and Hsiao, 2024) simulates interactions between doctors and patients to refine diagnostic reasoning and has shown better performance on datasets such as MedQA. The system developed by Ke et al. (Ke et al., 2024) reduces cognitive biases in diagnosis by using agents that provide expert opinions and critical evaluations. Other systems, such as *MedAide* (Wei et al., 2024b), coordinate agents across stages including pre-diagnosis, diagnosis, medication, and post-diagnosis, while frameworks such as MDagents (Kim et al., 2024c) and EHRagent (Tang et al., 2024) improve diagnostic accuracy through structured discussions and shared reasoning. Domain-specific applications also show promise. For instance, the work by Yue et al. (Yue et al., 2024) uses multi-agent collaboration to predict clinical trial outcomes by integrating large-scale domain knowledge. The Polaris framework (Mukherjee et al., 2024) combines general communication agents with task-specific agents to ensure safe patient interactions, and the system known as *Rx Strategist* (Van et al., 2024) uses knowledge graphs and multi-stage reasoning to check prescriptions for correct indications, dosages, and drug interactions.

4.2 Clinical Data Analytics and Documentation

In Clinical Data Analytics and Documentation, LLM-based agents show strong performance in processing both structured and unstructured data by using advanced architectures and retrieval-augmented generation techniques. LLM-based agents advance medical data synthesis by generating synthetic datasets that replicate clinical data, balancing data availability and privacy needs through integration with medical ontologies and knowledge graphs (Kumichev et al., 2024; Tang et al., 2023; Li et al., 2023b).

The system *ColaCare* proposed by Wang et al. (Wang et al., 2024b) integrates different agents to perform tasks such as mortality prediction and analysis of hospital readmission, demonstrating improved performance on the MIMIC-III and MIMIC-IV datasets. Beyond MIMIC datasets, other pivotal data sources include eICU, i2b2 corpora, MedNLI, and PMC-derived corpora, which provide critical foundations for developing LLM-based agents in medicine (Pollard et al., 2018; Romanov and Shiv-

ade, 2018).

The work by Lee et al. (Lee et al., 2024) introduces Sporo AI Scribe to address challenges related to the variability and complexity of clinical documentation. Research by Sudarshan (Sudarshan et al., 2024) shows that technical medical reports can be converted into patient-friendly formats by using iterative self-reflection and retrieval-augmented generation. In addition, *Agent Hospital* (Li et al., 2024b) contributes to simulation systems by generating complete interactions that improve diagnostic and treatment capabilities.

4.3 Medical Training and Simulation

In Medical Training and Simulation, simulation environments are used to test and refine LLM-based agents before their use in clinical practice. LLM-based agents serve as effective tools for training medical personnel by providing interactive case simulations and real-time adaptive feedback, functioning as personalized tutors for clinicians (Wei et al., 2024a; Yu et al., 2024).

Systems such as *ClinicalLab* (Yan et al., 2024) and *AI Hospital* (Fan et al., 2024) evaluate diagnostic and treatment performance by simulating interactions across many specialties and complex healthcare scenarios. The system *Agent Hospital* (Li et al., 2024b) further improves this process by allowing repeated training through large-scale simulations. In the field of medical education, systems such as *MEDCO* (Wei et al., 2024a) support training in diagnostic reasoning and collaborative problem solving, while *AIPatient* (Yu et al., 2024) integrates electronic health records with knowledge graphs to simulate realistic clinical scenarios. The system *SurgBox* (Wu et al., 2024) provides a training environment for surgical procedures with real-time decision support that has been validated against actual surgical records.

4.4 Healthcare Service Optimization

In Healthcare Service Optimization, LLM-based agents improve the delivery of healthcare by automating tasks such as patient education, data collection, and support services. Research shows that automating these non-diagnostic tasks reduces the workload of healthcare professionals while maintaining service quality (Swarms, 2025; Mukherjee et al., 2024; Laymouna et al., 2024). There is also potential for the future automation of certain diagnostic tasks, including endoscopies and surgeries (Chadebecq et al., 2023). These implementations have resulted in measurable improvements in operational efficiency and patient satisfaction.

4.5 Drug Discovery and Molecular Modeling Applications

Beyond clinical contexts, LLM-based agents advance pharmaceutical research through automated drug discovery and molecular modeling. These applications demonstrate enhanced efficiency in complex chemical and biological tasks traditionally requiring extensive manual effort.

LLM-based agents accelerate drug discovery through automated molecular property prediction and synthesis planning. DrugAgent employs multi-agent frameworks for drug-target interaction prediction, achieving 4.92% improvement in ROC-AUC (**?**), demonstrating enhanced efficiency in automating literature analysis and synthesis planning. In molecular modeling, ChemCrow integrates 18 expert-designed tools for autonomous molecular synthesis and property prediction (**?**), while MDCrow utilizes 40 tools for molecular dynamics simulations with chain-of-thought reasoning (**?**). These approaches provide scalable insights for computational chemistry, enabling researchers to explore molecular interactions and optimize drug candidates more efficiently than traditional computational methods.

5 Evaluation and Benchmarking

Evaluating LLM-based medical agents is essential for confirming their reliability, safety, and clinical effectiveness. A comprehensive evaluation framework is required to measure performance across different medical tasks, identify limitations, and guide improvements for clinical applications. A summary of the evaluation metrics and benchmark categories is provided in Table 2 in the appendix.

5.1 Benchmark Categories

Benchmarks for LLM-based medical agents can be divided into three categories.

Static Question-Answering benchmarks evaluate medical knowledge through tasks that have predetermined answers. For example, MedQA (Jin et al., 2020) simulates USMLE-style questions, MedMCQA (Pal et al., 2022) includes 194,000 questions covering 2,400 topics across 21 subjects, PubMedQA (Jin et al., 2019) assesses the understanding of biomedical research, and MMLU (Hendrycks et al., 2021b,a) offers cross-domain

single-choice questions. JAMA Clinical Challenge, which comprises real-world diagnostic cases, provides a critical complement to static QA benchmarks by aligning more closely with clinical practice. Although these datasets are useful for testing factual knowledge, they do not capture the interactive and sequential decision-making seen in clinical practice.

Workflow-based Simulation benchmarks mimic clinical decision-making through multiple stages. For instance, MedChain (Liu et al., 2024a) contains 12,163 cases from 19 specialties and uses 7,338 medical images, AI Hospital (Fan et al., 2024) evaluates interactions between healthcare providers and patients using the MVME dataset, AgentClinic (Schmidgall et al., 2024b) offers versions for both multimodal analysis and dialogue-based scenarios, and ClinicalLab (Yan et al., 2024) tests diagnostic performance across 24 departments and 150 diseases. Workflow-based benchmarks such as MedChain (Liu et al., 2024a) and AgentClinic (Schmidgall et al., 2024b) assess dialogue-based clinical diagnosis, while the MVME benchmark (Fan et al., 2024) evaluates patient-provider interactions, thereby enhancing the clinical relevance of LLM-based agents in dynamic care settings. These benchmarks reflect the dynamics of clinical reasoning and the adaptation required when patient information changes, although their complexity makes standardization challenging.

Automated Evaluation frameworks are developed to reduce reliance on human evaluators. For example, AI-SCE (Mehandru et al., 2024) uses an OSCE-based framework for systematic evaluation, and RJUA-SPs (Liu et al., 2024b) applies automated evaluation methods in urology using standardized patients and retrieval-augmented techniques.

5.2 Metrics for Task-specific Evaluation

Exact Match Metrics are used for tasks with clear correct answers, such as multiple-choice questions. In these tasks, accuracy, precision, and recall are calculated by directly comparing the model outputs with reference answers. Benchmarks such as MedQA (Jin et al., 2020) and MedMCQA (Pal et al., 2022) often use these metrics. While these metrics are effective for assessing factual knowledge, they may not be sufficient for tasks that involve complex reasoning or detailed explanations.

Semantic Similarity Metrics are applied to text generation tasks, such as writing clinical reports or diagnostic summaries. These metrics assess how well the meaning of the generated text matches that of the reference text. Metrics such as BLEU (Papineni et al., 2002), which measures n-gram overlap, ROUGE (Lin, 2004), which evaluates summarization quality, and BertScore (Zhang et al., 2020), which uses contextual embeddings to capture semantic relationships, have been applied in benchmarks such as ClinicalLab and MedChain.

LLM-based Evaluation Metrics use language models themselves to evaluate outputs based on factors such as coherence, relevance, and reasoning quality. For example, ChatCoach (Huang et al., 2024a) uses LLMs to assess the effectiveness of communication and decision making in patient consultations, while the Retrieval-Augmented Evaluation framework (Liu et al., 2024b) used in RJUA-SPs measures the alignment of outputs with standard clinical pathways. This approach provides a scalable and adaptable method for assessing complex, multi-step clinical tasks.

Human evaluation is also crucial for assessing clinical relevance and usability of LLM-based medical agents. Studies have shown that expert evaluation using standardized protocols can uncover subtle nuances and potential risks that automated metrics alone cannot capture (Chiu and Chung, 2024; Mehandru et al., 2024).

6 Discussions

Integrating Large Language Model (LLM)-based agents into medical workflows presents both challenges and opportunities. While previous work has achieved successes, the field remains in its early stages. Several significant challenges persist, and many opportunities require further exploration to fully realize their potential in healthcare applications. The following sections discuss these challenges and opportunities.

6.1 Technical Challenges

6.1.1 Hallucination Management

LLM hallucinations—instances where models generate incorrect or misleading information—pose a significant risk in medical contexts, potentially leading to erroneous diagnoses and treatments (Huang et al., 2024b). Benchmarks such as *MedHallBench* (Zuo and Jiang, 2024) and *HaluEval* (Li et al., 2023a) highlight the need for reliable verification systems and error prevention mechanisms, especially in multi-agent scenarios where mistakes can propagate. Future research should

focus on developing verification systems and dynamic error-correction methods that continuously update models with real-time, validated medical knowledge.

6.1.2 Multimodal and Multilingual Integration

LLM-based agents must process various data types, including clinical texts and medical images, and handle variability in medical terminology across different languages and cultures (Li et al., 2024a; Mehandru et al., 2024). Variations in documentation standards and regional practices add to this complexity. It is crucial to develop models that can reliably operate in both multilingual and multimodal contexts.

6.1.3 Cross-Department Integration

Healthcare environments encompass various departments, such as emergency, outpatient, and long-term care, each with its own workflows and documentation standards (Qiu et al., 2024). Achieving interoperability and accurate data exchange among these settings is challenging. Future work should focus on developing universal standards and adaptive interfaces that harmonize terminology and processes across departments, ensuring effective communication among LLM-based agents.

6.2 Evaluation Challenges

Evaluating LLM-based medical agents is challenging. Traditional static benchmarks, which focus on fixed question-answering tasks, do not capture the dynamic and interactive aspects of clinical workflows, such as sequential decision-making, adaptive reasoning, and effective communication with patients and clinicians (Jin et al., 2020; Schmidgall et al., 2024b). Moreover, many medical applications require integrating heterogeneous data types, including text records, images, and laboratory results, which calls for evaluation frameworks that accurately simulate multimodal interactions (Liu et al., 2024a; Fan et al., 2024). Standard language metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) assess only textual overlap and do not reflect clinical outcomes such as diagnostic accuracy. Additionally, dataset biases—such as the overrepresentation of specific conditions—can limit the generalizability of evaluation results across different healthcare settings (Yan et al., 2024). Future research should develop integrated, multimodal evaluation frameworks that combine quantitative measures with qualitative clinical assessments and establish standardized clinical performance metrics while reducing dataset biases (Mehandru et al., 2024).

6.3 Implementation Barriers

Integrating LLM-based agents into healthcare requires addressing interoperability among heterogeneous EHR systems through standards like HL7 FHIR, and ensuring compliance with regulatory requirements including HIPAA, FDA guidelines for AI/ML-based Software as a Medical Device, and GDPR (HL7 International; U.S. Department of Health & Human Services; U.S. Food and Drug Administration; European Commission, 2016).

6.3.1 System Integration Complexity

Large-scale systems like the Polaris healthcare system (Mukherjee et al., 2024), which involve millions of professionals and established decision-making processes, illustrate the complexity of integration. Although many LLM frameworks prove valuable in specific applications, their broader integration does not always lead to improvements in operational efficiency.

6.3.2 Resource Allocation Dilemma

Developing and maintaining LLM-based agents requires significant computational resources, resulting in high costs for medical institutions. Such investments may produce systems that are not entirely reliable, raising concerns about their cost-effectiveness.

6.4 Ethical and Privacy Concerns

6.4.1 Patient-Centered Design

Medical diagnosis systems powered by LLM agents currently receive limited feedback from patients and caregivers, despite the importance of including their perspectives in decision-making (Kim et al., 2024b). Most existing frameworks focus solely on interactions with physicians. A more responsible approach would integrate patient narratives, physician observations, and caregiver input to support a truly patient-centered process.

6.4.2 Algorithmic Bias

Both general-purpose and medically fine-tuned LLMs can exhibit various biases, including social and cognitive biases. Quantitative studies demonstrate the scope of this challenge: differential privacy approaches can balance data utility and

confidentiality (Dyda et al., 2021), while the BiasMedQA benchmark revealed that some medical LLMs achieve as low as 50% accuracy in handling specific biases (Schmidgall et al., 2024a). Medical agents must be designed to make responsible decisions, and reducing bias is essential for achieving this goal.

6.4.3 Privacy and Security Threats

Sensitive data used for training may be exposed during text generation or extracted through techniques such as inference attacks (Kandpal et al., 2023) or data extraction (Carlini et al., 2021). It is critical to protect sensitive information in accordance with regulations like GDPR (EU) (GDPR, 2016) and HIPAA (USA) (Act, 1996) when deploying medical agents. Data collection for developing LLM agents must prioritize privacy protection. (Dou et al., 2024) suggests using LLMs for autonomous data generation and labeling as a means to protect privacy. In addition, privacy-preserving data processing methods, such as differential privacy, can add controlled noise to data so that individual records do not significantly influence overall results while preserving data utility.

6.5 Future Opportunities and Application

6.5.1 Inspiration of O1 and R1 for Medical Reasoning

The evolution of LLM-based medical agents can benefit from insights drawn from DeepSeek R1 and inference-time scaling strategies. DeepSeek R1 (Guo et al., 2025) has shown that reinforcement learning combined with long-chain reasoning leads to more accurate and context-aware medical decision-making, offering potentials for improving autonomous medical agents (Faray de Paiva et al., 2025). By continuously optimizing AI-generated diagnoses and treatment recommendations through iterative self-evolution, LLM-based agents can better integrate multimodal clinical data, including electronic health records, medical images, and laboratory findings (Xu et al., 2024). Inference-time scaling, allowing LLMs more reasoning time, has been shown to improve performance in complex tasks such as differential diagnosis and treatment planning (Huang et al., 2025), consistent with the hypothetico-deductive method used in clinical reasoning. Future research should explore how LLM-based agents can dynamically adjust inference time based on task complexity while incorporating reinforcement learning-based optimization techniques to enhance adaptability and reliability in clinical settings.

6.5.2 Integration with Physical Systems

Expanding LLM-based agents from virtual applications to integration with physical systems represents a significant step in medical care. While LLMs excel at data analysis and decision support, connecting them with physical systems such as medical robots could enable direct patient care. Such systems might combine language processing with physical inputs to support tasks like surgical assistance and patient monitoring. For example, empowering nursing robots (Zhao et al., 2025) is one potential approach. However, this integration raises challenges regarding safety and real-time performance. Addressing technical limitations, ensuring system reliability, and resolving ethical concerns are necessary for successful integration. Hardware systems must accurately execute LLM outputs because errors could endanger patient safety, and high costs or technical complexity may limit system availability. Future work should focus on improving the integration of LLM-based agents with physical systems and on creating practical implementation frameworks.

6.5.3 Advancements in Training Simulation

Current medical LLM agents often use simulated hospitals for training, such as the *Agent Hospital* framework, which enables the autonomous evolution of doctor agents through synthetic patient interactions (Li et al., 2024b). Extending these simulations to include educational medical games could improve training data generation and learning experiences, even though challenges in data quality remain. AI-driven patient simulations that provide structured feedback have demonstrated effectiveness in enhancing clinical decision-making (Brügge et al., 2024), but validating data generated by such games remains resource intensive.

7 Conclusion

This survey examines LLM-based agents in medicine, covering their architectures, applications, and challenges. While these agents enhance diagnostics, data analysis, and clinical workflows, issues remain in hallucination management, multimodal integration, and medical reasoning accuracy. Future work should focus on real-time error correction, improved multimodal fusion, and hybrid reasoning to enhance reliability and clinical utility.

Limitations

This survey has several inherent limitations that should be considered. Due to the rapid development of LLM-based medical agents, our review primarily covers works published between 2022 and early 2025, which means future developments may introduce new architectures and approaches not captured in this analysis. Additionally, while we aimed for comprehensive coverage, we focused mainly on English-language publications in major academic databases such as PubMed, ACM Digital Library, arXiv, and Google Scholar. Valuable work published in other languages or regional databases may not be included in our analysis. These limitations reflect the inherent constraints of conducting a survey in a rapidly evolving field rather than shortcomings in the reviewed research itself.

References

Accountability Act. 1996. Health insurance portability and accountability act of 1996. *Public law*, 104:191.

David Bani-Harouni, Nassir Navab, and Matthias Keicher. 2024. Magda: Multi-agent guideline-driven diagnostic assistance. *Preprint*, arXiv:2409.06351.

Emilia Brügge, Sarah Ricchizzi, Malin Arenbeck, Marius Niklas Keller, Lina Schur, Walter Stummer, Markus Holling, Max Hao Lu, and Dogus Darici. 2024. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education*, 24.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

F Chadebecq, L.B. Lovat, and D. Stoyanov. 2023. Artificial intelligence and automation in endoscopy and surgery. *Nature Reviews Gastroenterology & Hepatology*, 20:171–182. Accepted: 03 October 2022, Published: 09 November 2022, Issue Date: March 2023.

Jiajun Chai, Sicheng Li, Yuqian Fu, Dongbin Zhao, and Yuanheng Zhu. 2025. Empowering LLM agents with zero-shot optimal decision-making through q-learning. In *The Thirteenth International Conference on Learning Representations*.

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. 2024. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *Preprint*, arXiv:2401.03428.

Edwin Kwan-Yeung Chiu and Tom Wai-Hin Chung. 2024. Protocol for human evaluation of artificial intelligence chatbots in clinical consultations. *medRxiv*, pages 2024–03.

Chengfeng Dou, Ying Zhang, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2024. Exploring llm-based data annotation strategies for medical dialogue preference alignment. *Preprint*, arXiv:2410.04112.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haohao Ying. 2024. Llms can simulate standardized patients via agent coevolution. *Preprint*, arXiv:2412.11716.

Abhishek Dutta and Yen-Che Hsiao. 2024. Adaptive reasoning and acting in medical language agents. *Preprint*, arXiv:2410.10020.

A Dyda, M Purcell, S Curtis, et al. 2021. Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. *Patterns (N Y)*, 2(12):100366.

European Commission. 2016. Data protection in the eu. Accessed: 2025-04-02.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Preprint*, arXiv:2402.09742.

Lisle Faray de Paiva, Gijs Luijten, Behrus Puladi, and Jan Egger. 2025. How does deepseek-r1 perform on usmle? *medRxiv*, pages 2025–02.

GDPR GDPR. 2016. General data protection regulation. *Regulation (EU)*, 679.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

HL7 International. Fhir overview. `https://www.hl7.org/fhir/overview.html`. Accessed: 2025-04-02.

Hengguan Huang, Songtao Wang, Hongfu Liu, Hao Wang, and Ye Wang. 2024a. Benchmarking large language models on communicative medical coaching: a novel system and dataset. *Preprint*, arXiv:2402.05547.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. O1 replication journey – part 3: Inference-time scaling for medical reasoning. *Preprint*, arXiv:2501.06458.

Xun Jiang, Feng Li, Han Zhao, Jiaying Wang, Jun Shao, Shihao Xu, Shu Zhang, Weiling Chen, Xavier Tang, Yize Chen, Mengyue Wu, Weizhi Ma, Mengdi Wang, and Tianqiao Chen. 2024. Long term memory: The foundation of ai self-evolution. *Preprint*, arXiv:2410.15665.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A Choquette-Choo, and Zheng Xu. 2023. User inference attacks on large language models. *arXiv preprint arXiv:2310.09266*.

Yuhe Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Yilin Ning, Irene Li, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: Simulation study. *J Med Internet Res*, 26:e59439.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024a. Mdagents: An adaptive collaboration of llms for medical decision-making. *Preprint*, arXiv:2404.15155.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024b. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Cristina Grau-Vilchez, et al. 2024c. A demonstration of adaptive collaboration of large language models for medical decision-making. *Preprint*, arXiv:2411.00248.

Anastasiya Kiseleva, Dimitris Kotzinos, and Paul De Hert. 2022. Transparency of ai in healthcare as a multilayered system of accountabilities: Between legal requirements and technical limitations. *Frontiers in Artificial Intelligence*, 5.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. 2024. *MedSyn: LLM-Based Synthetic Medical Text Generation Framework*, page 215–230. Springer Nature Switzerland.

M Laymouna, Y Ma, D Lessard, T Schuster, K Engler, and B Lebouché. 2024. Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *Journal of Medical Internet Research*, 26:e56930.

Chanseo Lee, Sonu Kumar, Kimon A. Vogt, and Sam Meraj. 2024. Improving clinical documentation with ai: A comparative study of sporo ai scribe and gpt-4o mini. *Preprint*, arXiv:2410.15528.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. Mmedagent: Learning to use medical tools with multi-modal agent. *Preprint*, arXiv:2407.02483.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Preprint*, arXiv:2405.02957.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Rumeng Li, Xun Wang, and Hong Yu. 2023b. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7129–7143, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. 2024a. Medchain: Bridging the gap between llm agents and clinical practice through interactive sequential benchmarking. *Preprint*, arXiv:2412.01605.

Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu, Ming Zhang, Xiaowei Ma, Xiangguo Lv, Liya Ma, Zhiqiang Zhang, Wei Xue, Yiran Huang, and Jinjie Gu. 2024b. Towards automatic evaluation for llms' clinical capabilities: Metric, data, and algorithm. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pages 5466–5475, New York, NY, USA. Association for Computing Machinery.

Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J Butte, and Ahmed Alaa. 2024. Evaluating large language models as agents in the clinic. *NPJ digital medicine*, 7(1):84.

Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, et al. 2024. Polaris: A safety-focused llm constellation architecture for healthcare. *Preprint*, arXiv:2403.13313.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Y. Park et al. 2024. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making*, 24(1):72.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13.

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. 2024. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420.

Sandeep Reddy. 2024. Generative ai in healthcare: an implementation science informed translational path on application, integration and governance. *Implementation Science*, 19(1):27.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach*. Pearson.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024a. Addressing cognitive bias in medical language models. *arXiv preprint arXiv:2402.08113*.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024b. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D. Wang. 2024. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. *Preprint*, arXiv:2401.07128.

Joshua Strong, Qianhui Men, and Alison Noble. 2024. Towards human-ai collaboration in healthcare: Guided deferral systems with large language models. *Preprint*, arXiv:2406.07212.

Malavikha Sudarshan, Sophie Shih, Estella Yee, Alina Yang, John Zou, Cathy Chen, Quan Zhou, Leon Chen, Chinmay Singhal, and George Shih. 2024. Agentic llm workflows for generating patient-friendly medical reports. *Preprint*, arXiv:2408.01112.

Lei SUN, An'an WANG, Yimin SONG, Jing DONG, Xiaoli LIU, Hong LIANG, Lixuan LI, Xinyu SONG, Yong FAN, Zhilong JIA, et al. 2024. Applications, challenges, and prospects of large language models in the field of clinical medicine.

Swarms. 2025. Unlocking efficiency and cost savings in healthcare: How swarms of llm agents can revolutionize medical operations and save millions. Accessed: 2025-02-04.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *Preprint*, arXiv:2303.04360.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, et al. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. *Preprint*, arXiv:2311.10537.

U.S. Department of Health & Human Services. Hipaa privacy rule. `https://www.hhs.gov/hipaa/for-professionals/privacy/index.html`. Accessed: 2025-04-02.

U.S. Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device. `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device`. Accessed: 2025-04-02.

Phuc Phan Van, Dat Nguyen Minh, An Dinh Ngoc, and Huy Phan Thanh. 2024. Rx strategist: Prescription verification using llm agents system. *Preprint*, arXiv:2409.03440.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Zixiang Wang, Yinghao Zhu, Huiya Zhao, et al. 2024b. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. *Preprint*, arXiv:2410.02551.

Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. 2024a. Medco: Medical education copilots based on a multi-agent framework. *Preprint*, arXiv:2408.12496.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jinjie Wei, Dingkang Yang, Yanshu Li, Qingyao Xu, Zhaoyu Chen, Mingcheng Li, Yue Jiang, Xiaolu Hou, and Lihua Zhang. 2024b. Medaide: Towards an omni medical aide via specialized llm-based multi-agent collaboration. *Preprint*, arXiv:2410.12532.

Jinlin Wu, Xusheng Liang, Xuexue Bai, and Zhen Chen. 2024. Surgbox: Agent-driven operating room sandbox with surgery copilot. *Preprint*, arXiv:2412.05187.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey. *Preprint*, arXiv:2309.07864.

Shaochen Xu, Yifan Zhou, Zhengliang Liu, Zihao Wu, Tianyang Zhong, Huaqin Zhao, Yiwei Li, Hanqi Jiang, Yi Pan, Junhao Chen, Jin Lu, Wei Zhang, Tuo Zhang, Lu Zhang, Dajiang Zhu, Xiang Li, Wei Liu, Quanzheng Li, Andrea Sikora, Xiaoming Zhai, Zhen Xiang, and Tianming Liu. 2024. Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios. *Preprint*, arXiv:2411.14461.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, Li Zhu, and Xuandong Zhao. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *Preprint*, arXiv:2406.13890.

Wenxin YAN, Jian HU, Huatang ZENG, Min LIU, and Wannian LIANG. 2025. The application of large language models in primary healthcare services and the challenges. *Chinese General Practice*, 28(01):1.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Huizi Yu, Jiayan Zhou, Lingyao Li, et al. 2024. Aipatient: Simulating patients with ehrs and llm powered agentic workflow. *Preprint*, arXiv:2409.18924.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, IUI '22, page 841–852, New York, NY, USA. Association for Computing Machinery.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. 2024. R-judge: Benchmarking safety risk awareness for LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1467–1490, Miami, Florida, USA. Association for Computational Linguistics.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. *Preprint*, arXiv:2404.14777.

Luyao Zhang, Jianhua Shu, Jili Hu, Fangfang Li, Junjun He, Peng Wang, and Yiqing Shen. 2024. Exploring the potential of large language models in radiological imaging systems: Improving user interface design and functional capabilities. *Electronics*, 13(11).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Zhendong Zhao, Xiaotian Yue, Jiexin Xie, Chuanhong Fang, Zhenzhou Shao, and Shijie Guo. 2025. A dual-agent collaboration framework based on llms for nursing robots to perform bimanual coordination tasks. *IEEE Robotics and Automation Letters*, pages 1–8.

Yanxin Zheng, Wensheng Gan, Zefeng Chen, et al. 2024. Large language models for medicine: A survey. *Preprint*, arXiv:2403.10100.

Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2024. Menti: Bridging medical calculator and llm agent with nested tool calling. *Preprint*, arXiv:2410.13610.

Kaiwen Zuo and Yirui Jiang. 2024. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *arXiv preprint arXiv:2412.18947*.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2025. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis. *Preprint*, arXiv:2412.16833.

A Appendix

A.1 Paper Collection

All papers included in this review were identified through a systematic search of major academic databases such as PubMed, ACM Digital Library, arXiv, and Google Scholar. Keywords such as "large language model", "medical agent", "clinical decision support", and "healthcare AI" were used to select relevant studies published between 2022 and early 2025. This process initially identified about 300 articles, from which 80 were shortlisted based on title and abstract screening for relevance and quality. After a full-text review, approximately 60 studies specifically addressing LLM-based agents in medical contexts were selected for this survey.

A.2 Evaluation and Benchmarking

Table 2 demonstrates common evaluation benchmarks and metrics for the llm-based agents in medicine.

Table 2: Common Evaluation Benchmarks and Metrics

Evaluation Attribute	Genre	Specific Names	Related Work
Benchmarks	Static Q&A Benchmarks	*MedQA*	(Jin et al., 2020)
		MedMCQA	(Pal et al., 2022)
		Pub-MedQA	(Jin et al., 2019)
		MMLU	(Hendrycks et al., 2021b) (Hendrycks et al., 2021a)
	Workflow-Based Simulation Benchmarks	*MedChain*	(Liu et al., 2024a)
		AI Hospital	(Fan et al., 2024)
		AgentClinic	(Schmidgall et al., 2024b)
		ClinicalLab	(Yan et al., 2024)
	Automated Evaluation Frameworks	*AI-SCE*	(**?**)
		RJUA-SPs	(Liu et al., 2024b)
Metrics for Task-Specific Evaluation	Exact Match Metrics	Accuracy	–
		Precision	–
		Recall	–
	Semantic Similarity Metrics	**BLEU**	(Papineni et al., 2002)
		ROUGE	(Lin, 2004)
		BertScore	(Zhang et al., 2020)
	LLM-Based Evaluation Metrics	*ChatCoach*	(Huang et al., 2024a)
		Retrival-Augmented Evaluation Framework	(Liu et al., 2024b)