A Comprehensive Survey of AI Agents in Healthcare

Gelei Xu[1], Xueyang Li[1], Yixiong Chen[2], Yuying Duan[1], Shuqing Wu[1], Alexander Yu[1], Ching-Hao Chiu[1], Juntong Ni[3], Ningzhi Tang[1], Toby Jia-Jun Li[1], Alan Yuille[2], Wei Jin[3], and Yiyu Shi[1]

[1]University of Notre Dame
[2]Johns Hopkins University
[3]Emory University

November 14, 2025

Abstract

The rapid advancement of large language models has accelerated the development of AI agents. This shift has important implications for healthcare, a domain characterized by intensive knowledge use and complex clinical decisions. At the same time, the high-stakes and safety-critical nature of medical practice poses unique challenges, making general-purpose agentic frameworks often inadequate. The autonomy that defines these agents introduces additional concerns regarding trust, reliability, and alignment with clinical constraints. With research in this area growing exponentially over the past two years, this survey synthesizes more than 200 recent studies and presents a structured map of healthcare agents. It begins with a quantitative landscape analysis of growth trends, technical focus areas, and emerging application domains. A lifecycle taxonomy is then presented, covering the *Perception* of clinical modalities and continuing through core *Agentic Capabilities and Architectures*. This is followed by an outline of the *Application Ecosystem* organized by major stakeholders, and a review of prevailing *Evaluation* frameworks. The survey concludes with current limitations and future research opportunities. A curated list of all related papers is available at: https://github.com/AgenticHealthAI/Awesome-AI-Agents-for-Healthcare/.

A Comprehensive Survey of AI Agents in Healthcare

Gelei Xu[*†], Xueyang Li[*†], Yixiong Chen[*‡], Yuying Duan[*†], Shuqing Wu[*†], Alexander Yu[*†], Ching-Hao Chiu[*†], Juntong Ni[*§], Ningzhi Tang[†], Toby Jia-Jun Li[†], Alan Yuille[‡], Wei Jin[§], Yiyu Shi[†]
[†]{gxu4, xli34, yduan2, swu7, hyu4, cchui3, ntang, toby.j.li, yshi4}@nd.edu
[‡]{ychen646, ayuille1}@jh.edu, [§]{juntong.ni, wei.jin}@emory.edu
[†]University of Notre Dame, Notre Dame, IN, USA
[‡]Johns Hopkins University, Baltimore, MD, USA
[§]Emory University, Atlanta, GA, USA

Abstract

The rapid advancement of large language models has accelerated the development of AI agents. This shift has important implications for healthcare, a domain characterized by intensive knowledge use and complex clinical decisions. At the same time, the high-stakes and safety-critical nature of medical practice poses unique challenges, making general-purpose agentic frameworks often inadequate. The autonomy that defines these agents introduces additional concerns regarding trust, reliability, and alignment with clinical constraints. With research in this area growing exponentially over the past two years, this survey synthesizes more than 200 recent studies and presents a structured map of healthcare agents. It begins with a quantitative landscape analysis of growth trends, technical focus areas, and emerging application domains. A lifecycle taxonomy is then presented, covering the *Perception* of clinical modalities and continuing through core *Agentic Capabilities and Architectures*. This is followed by an outline of the *Application Ecosystem* organized by major stakeholders, and a review of prevailing *Evaluation* frameworks. The survey concludes with current limitations and future research opportunities. A curated list of all related papers is available at https://github.com/AgenticHealthAI/Awesome-AI-Agents-for-Healthcare/.

1 Introduction

The emergence of large language models (LLMs) has advanced AI systems toward the capabilities of artificial general intelligence [53]. A key development in this progression is the rise of AI agents [149, 179], which extends predictive modeling with additional capabilities such as autonomy, planning, memory, and tool use. Instead of generating isolated outputs, agents can pursue goals, decompose them into multi-step plans, learn from prior interactions, and employ external tools to expand their capabilities. This transformation enables AI to operate adaptively in complex and dynamic environments, rather than remaining confined to predefined tasks.

Healthcare presents a strong need for AI agents. It is a knowledge-intensive domain in which clinicians face information overload from heterogeneous sources such as electronic health records (EHRs), imaging, genomics, and medical literature [45, 87, 156]. Agents can integrate fragmented inputs into longitudinal patient profiles through tools and memory. Many clinical tasks also require long-horizon, adaptive decision-making. For instance, in cancer management, therapy cycles involve planning and simulating treatment strategies, with similar demands in chronic and acute care where decisions must adjust to changing patient states [138]. Fragmented workflows and rising expectations for personalized communication further underscore the role of agents as workflow coordinators [48]. By managing analysis, documentation, and communication while maintaining continuity through memory and persona conditioning, agentic systems can act as integrated partners in care delivery.

Despite these prospects, the adoption of AI agents in clinical practice remains more limited than in many other domains. Healthcare is a highly stakes and tightly regulated environment where errors directly affect patient safety [52, 120]. In this context, the defining characteristic of agents, autonomy, introduces new challenges for safety and legal accountability. Clinical systems that access sensitive data such as EHRs, imaging, and patient dialogues must comply with strict regulatory requirements (e.g., HIPAA, GDPR) [50, 175, 197]. When such multimodal data are combined, clinically useful insights must be generated without compromising privacy. Decisions must also reflect the ethical obligation to avoid harm, not just optimize model performance. Moreover, treatment systems should allow clinicians and patients to scrutinize outcomes, establish trust, and retain final authority [21, 112]. These requirements show that healthcare agents cannot simply replicate general-purpose designs [145]. The central challenge is embedding safety, trust, and accountability under uncertainty while still supporting task effectiveness, which reshapes the design space of AI agents in healthcare.

Research on healthcare agents has expanded rapidly, reflected in the substantial increase in publications. In 2025, the number of studies is already more than 130% higher than in 2024. This growth coincides with a broader technological transition driven by advances in reliable tool-use language models [150], multimodal foundation models [104, 172], and interoperable clinical APIs [47]. Together, these developments mark an inflection point where agents can interact with clinical data to coordinate tasks and operate within controlled simulators—capabilities that were previously impractical. Despite this progress, systematic analyses that connect these technical advances to the specific characteristics and constraints of healthcare remain limited. Existing surveys [182] provide useful overviews of architectures and applications, but they cover a narrow set of studies and rarely examine how modality characteristics and technical developments intersect to shape domain-specific challenges for agentic systems. This survey provides the first comprehensive synthesis of AI agents in healthcare, analysing over 200 recent studies and integrating perspectives across perception, agent capabilities, and application domains to offer a unified view of how healthcare agents are designed and deployed.

[*]The authors contributed equally to this research.

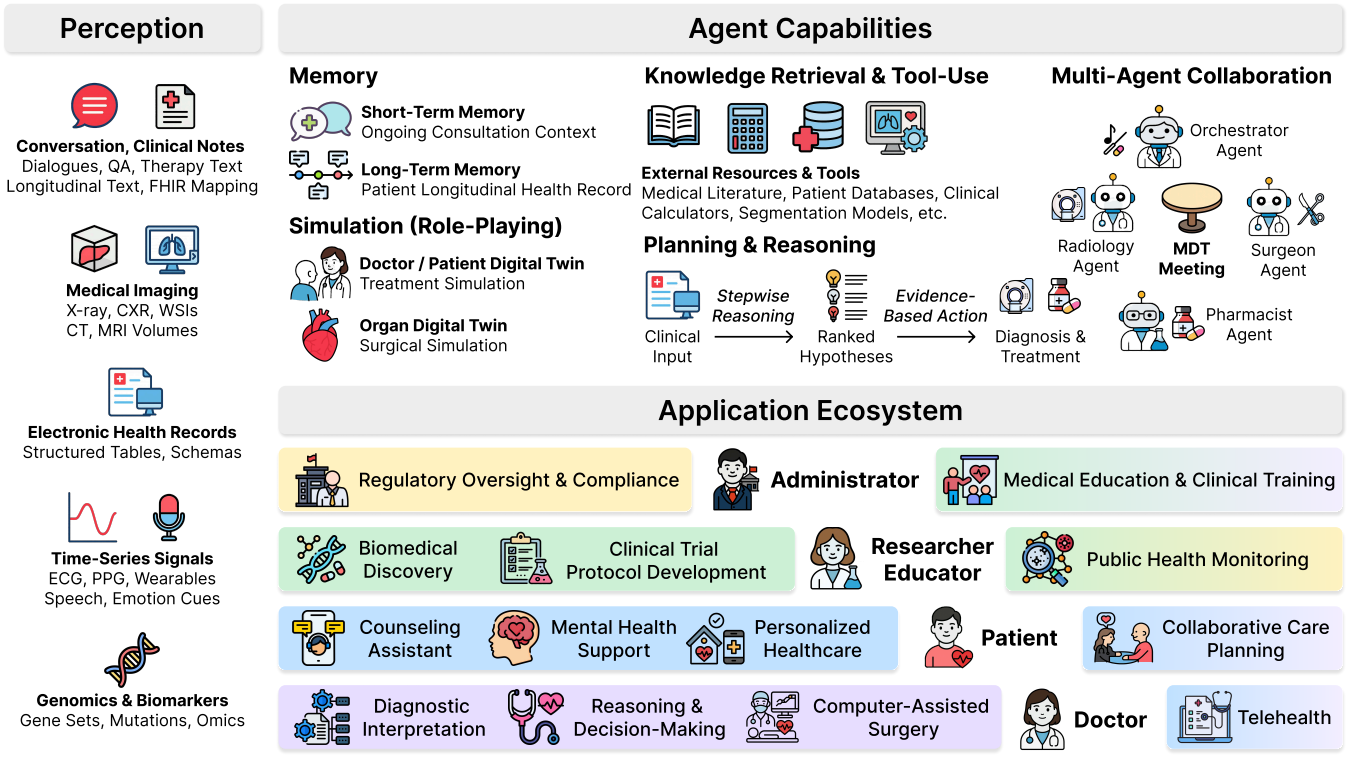


Figure 1: A conceptual framework for AI agents in healthcare. The framework illustrates the flow from data Perception and core Agent Capabilities to a hierarchical Healthcare Agent Application Ecosystem, organized around four central stakeholders (doctor, patient, researcher/educator, and administrator), each linked to both role-specific and collaborative applications.

The remainder of this paper is organized as follows. Section 2 presents a macro-level analysis of current trends in healthcare agent development, and Section 3 outlines the survey methodology. Section 4 discusses perception and environment. Section 5 examines the core agent architectures and technical capabilities. Section 6 reviews applications across the healthcare ecosystem organized by key stakeholders. Section 7 focuses on the evaluation of current agentic systems. Section 8 discusses open challenges and future directions, and Section 9 concludes.

In summary, this survey structures recent work on AI agents in healthcare into a comprehensive framework. For computer scientists, it outlines how safety-critical constraints in medicine reshape core questions of autonomy and evaluation, and proposes guidance for future research. For clinicians, it maps emerging applications to clarify current capabilities and limitations, supporting more informed participation in adoption and oversight. By connecting these perspectives, the survey provides a shared foundation for advancing innovation while maintaining accountability.

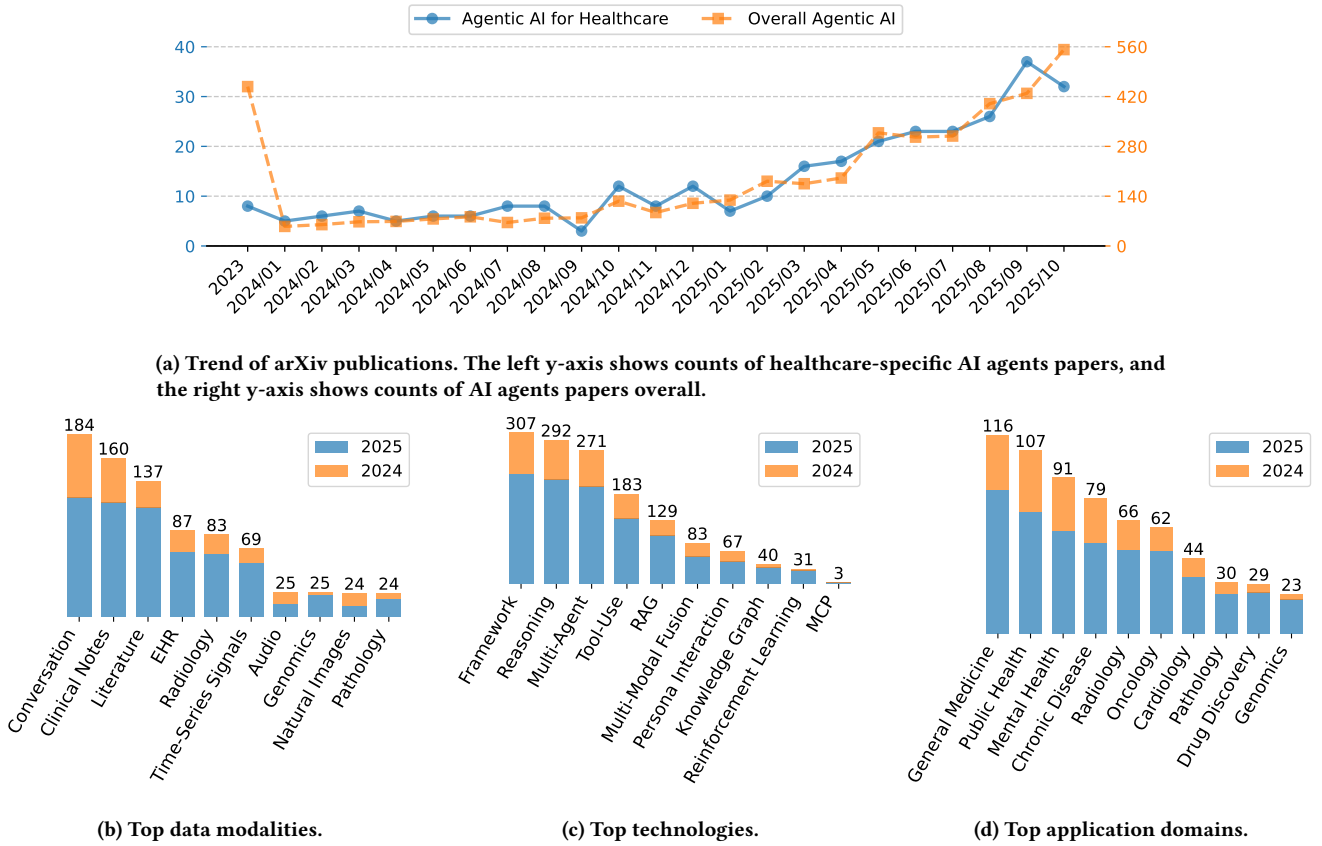
Connections to Existing Surveys. Several recent surveys analyze aspects of AI agents in healthcare. Some review small sets of systems and focus on profiles, planning methods, or ethical safeguards [174, 182]. Others target specific domains such as radiology or outline implementation roadmaps that emphasize privacy and interoperability [12, 131]. Broader surveys of language models and autonomous agents discuss architectures, datasets, and evaluation practices across fields [179, 229], with limited coverage of safety-critical and regulatory factors unique to healthcare. Building on these contributions, our survey synthesizes more than 200 recent studies and introduces a unified taxonomy connecting perception modalities, agent capabilities, and stakeholder applications, alongside evaluation and governance. This domain-oriented synthesis complements prior reviews by grounding technical developments within clinical practice and regulatory requirements.

2 A Landscape of AI Agents in Healthcare

Definition of General Agents. A general AI agent is a computational system that perceives its environment, maintains internal state, plans, and takes actions to achieve tasks [205]. Contemporary definitions highlight three elements: a reasoning module, interfaces for perception, and action mechanisms for invoking tools or APIs. Agents often incorporate memory for contextual continuity, planning modules for multi-step goals, and tool-use capabilities for interacting with external systems [114]. These components enable adaptation to changing conditions and coordinated task execution.

2.1 A Conceptual Framework for AI Agents in Healthcare

To systematically map the complex ecosystem of AI agents in healthcare, we introduce a comprehensive conceptual framework, illustrated in Figure 1. This framework provides a holistic view, detailing



(a) Trend of arXiv publications. The left y-axis shows counts of healthcare-specific AI agents papers, and the right y-axis shows counts of AI agents papers overall.

(b) Top data modalities.

(c) Top technologies.

(d) Top application domains.

Figure 2: Quantitative analysis of the research landscape for healthcare agents, based on a survey of recent literature. (a) Trend of publications, (b) frequency of top data modalities, (c) key technologies, and (d) application domains.

the pipeline from initial data perception and foundational agent capabilities to a hierarchical application ecosystem.

The framework is built upon three core pillars:

- **Perception**: At the entry point of the framework, agents perceive the clinical environment through diverse data modalities. As shown on the left of Figure 1, these inputs range from structured sources like EHR to unstructured data such as text and medical images, as well as high-dimensional genomic data and time-series signals. This multi-modal sensory capability is fundamental to an agent's understanding and will be explored in detail in Section 4.
- **Agent Blocks**: At the top right of the framework lie the foundational technical agent capabilities that enable agentic behavior. These "Agent Blocks" represent the core competencies discussed in section 5. They include 1) **Memory**, both short-term and long-term, for maintaining context and continuity; 2) **Tool-use**, which allows agents to access external knowledge sources and computational utilities; 3) **Simulation**, which helps align agent behavior with clinical norms and user expectations; 4) **Planning & Reasoning** for goal decomposition and action selection; and 5) **Multi-Agent** systems for collaborative problem-solving.
- **Application Ecosystem**: The ecosystem organizes applications by primary stakeholders on the left and cross-stakeholder collaborations on the right. **Doctor**-oriented tasks such as diagnostic interpretation, decision support, and computer-assisted surgery form the base layer. **Patient**-facing applications, including counseling assistants, mental health support, and personalized healthcare, build on this foundation. The next tier comprises **Researcher** and **Educator** activities such as biomedical discovery, clinical trial protocol development, and medical training, followed by **Administrator**-level functions involving regulatory oversight and compliance. The right side illustrates **Multi-Stakeholder Interactions** in which agents coordinate processes such as telehealth consultations, collaborative care planning, and research data access that requires consent. This structure reflects both stakeholder-specific needs and the coordinated workflows present in healthcare practice.

2.2 Quantitative Analysis of the Research Landscape

We conducted a quantitative analysis of recent academic literature, and the key findings are summarized in Figure 2. This analysis

offers a data-driven view of the field's growth trajectory, technical focus areas, and application trends. The results in Figure 2a are obtained from the arXiv trend analysis. The findings in Figures 2b, 2c, and 2d are based on labels assigned during our initial paper screening, with methodological details provided in Section 3.

The research interest in this domain is experiencing exponential growth, as shown in Figure 2a. The broader field of AI agents began rising before 2024, and healthcare-focused work is now rapidly following the same trajectory. This sharp increase reflects growing community interest and underscores the timeliness of this survey.

A breakdown of the surveyed papers reveals distinct patterns of established focus areas and emerging frontiers:

- **Top Data Modalities** (Figure 2b): Textual data, including *Conversation* and *Clinical Notes*, remain the most frequently utilized modalities, reflecting their centrality in clinical practice. EHR and medical images have also been prevalent since 2024. However, the stacked chart reveals that modalities such as ***Time-Series*** and ***Genomics*** exhibit a high proportion of publications from 2025. This indicates their status as rapidly emerging areas of interest, likely driven by the increased availability of wearable sensor data and advances in precision medicine.
- **Top Technologies** (Figure 2c): The technological focus is heavily concentrated on three topics: 1) developing high-level *Frameworks*, 2) enhancing agent *Reasoning*, and 3) designing *Multi-Agent* collaboration paradigms. While these established areas continue to grow, the stacked data reveals a significant recent uptick in research on ***Knowledge Graph*** integration and ***Reinforcement Learning (RL)***. This indicates a growing trend towards more sophisticated methods for agent reasoning in structured knowledge and enabling reward-driven decision-making.
- **Top Application Domains** (Figure 2d): AI agents continues to be widely applied in broad domains like *General Medicine*, *Public Health*, and *Mental Health*, highlighting the significant potential of agentic systems to address large-scale and long-term health management challenges. Critically, the year-over-year breakdown highlights ***Drug Discovery*** and ***Genomics*** as particularly new frontiers. The majority of agent-based research in these domains has appeared in 2025, pointing to a growing application of AI agents in foundational biomedical science where complex hypothesis generation and data interpretation are paramount.

In summary, our quantitative analysis paints a clear picture of a dynamic research field. While established areas continue to mature, there is clear momentum towards newer modalities like genomics, more advanced technologies like reinforcement learning, and high-impact applications in drug discovery and precision medicine.

3 Survey Methodology

We conducted a systematic literature review to address the research questions on AI agents in healthcare. Following established methodologies [124], the objective was to identify and synthesize relevant studies and provide a comprehensive overview of existing work. We queried five major academic databases, i.e., Google Scholar, PubMed, ACM Digital Library, IEEE Xplore, and ACL Anthology, which together span both artificial intelligence and healthcare research. To capture recent developments in this rapidly evolving area, we collected peer-reviewed publications as well as preprints from sources such as arXiv. The search strategy combined four categories of keywords: agent-related terms (e.g., agent, multi-agent system, virtual assistant, agentic), model-related terms (e.g., large language model, LLM, foundation model), healthcare-related terms (e.g., medical, clinical, health, patient, diagnosis), and modality-related terms (e.g., computed tomography, MRI, ultrasound, pathology).

After the initial collection, papers were screened to retain those most relevant to the review. Papers were included if they (1) addressed a healthcare-related task and (2) employed an agent or proposed a framework with agentic capabilities. The search yielded 1,106 papers. After deduplication and relevance filtering, 682 papers remained. These were assigned multi-label annotations covering technology type, medical domain, task, clinical stage, data modality, evaluation paradigm, datasets, and metrics to support subsequent selection and organization. The final set of included studies comprised 223 publications. Additional references on general-purpose AI agents and foundational methods were cited as needed to contextualize the discussion.

4 Perception and Environment

The effectiveness of a healthcare agent depends on the data it perceives. In clinical contexts, the modality and type of input data (e.g., natural language, EHR, images, signals, genomics, and audio) determine both the agent's functional role and its action space. This section reviews how different modalities are processed by LLMs and how they define distinct action environments. It concludes with a discussion of multi-modal systems that integrate heterogeneous inputs and coordinate reasoning strategies that individual modalities cannot support alone. We summarize the perception modalities, agent roles, and representative systems in Table 1.

4.1 Single Modality

4.1.1 Natural Language Conversation. Natural language, especially in conversational form, is a central modality for healthcare agents. For LLM-based systems, dialogue is not a fixed dataset of facts but an evolving sequence of interactions, including question answering and therapeutic exchanges. The agent's perception focuses on tracking conversational context, inferring user intent across dialogue turns, and capturing subtle semantic variations during clinical discussions. This process forms a collaborative setting where agents actively engage in information exchange. Within such settings, agents can take on various roles. They may serve as clinical investigators, conducting multi-turn dialogues to identify symptoms and refine diagnostic hypotheses, as in DoctorAgent-RL [40]. They may act as virtual clinicians in simulated multidisciplinary teams, where multiple agents deliberate and reach consensus, as demonstrated by MedAgents [171]. In mental health applications, systems such as CACTUS [83] function as counselors, using structured conversation to guide therapeutic interventions.

Open Challenges. A major challenge lies in the mismatch between static training datasets and the variability of real clinical interactions. Agents trained on fixed dialogues often perform well

Modality	Input Data	Perception Methods	Representative Roles	Representative Systems
Natural Language Conversation	Dialogues, QA, Therapy Text	Text Understanding	Clinician	DoctorAgent-RL [40] MedAgents [171] CACTUS [83]
Electronic Health Records	Structured EHR Tables, Schemas	Data Analysis → Text Understanding	Data Analyst	EHRFlow [190] TrustEHRAgent [160]
Clinical Notes	Longitudinal Text, FHIR Mapping	Temporal Extraction → Text Understanding	Diagnostic Assistant	Infherno [42] CARE-AD [93]
2D Medical Imaging	CXR, WSIs	Image Encoding → Token Patching	Radiologist Pathologist	MedRAX [34] PathChat+ [18] CPathAgent [167]
3D Medical Imaging	CT, MRI Volumes	Volume Encoding → Token Patching	Radiologist	CT-Agent [113] AgentMRI [147] MESHAgents [219]
Time-Series Signals	ECG, PPG, Wearables	Statistical Analysis → Text Understanding	Computational Analyst	PHA [58] LLM-PPG [38]
Genomics & Biomarkers	Gene Sets, Mutations, Omics	Knowledge Graph → Statistical Analysis → Text Understanding	Bioinformatics Analyst	HEAL-KGGen [240] GeneAgent [184] AI-HOPE [202]
Audio	Speech, Emotion Cues	Speech-to-Text → Text Understanding, Waveform Analysis	Empathetic Listener	ANNA [130] Survey Agent [74]
Multi-Modality	Text, Image, Signal, Omics	Combination of Above	Coordinator Tool Orchestrator	MAM [233] MedAgent-Pro [185] MMedAgent [87]

Table 1: Overview of perception modalities, agent roles, and representative systems for healthcare agents.

in imitation tasks but lack adaptive reasoning for unexpected patient behaviors [40]. Another challenge is maintaining long-term dialogue coherence and generating clinically meaningful follow-up questions, which require domain-specific reasoning capabilities beyond general conversation.

4.1.2 Electronic Health Records (EHRs). The EHR modality places the agent in a data-centric query environment. The task shifts from general questions such as "What are the symptoms of diabetes?" to specific instructions like "Identify all patients over 65 with a diabetes diagnosis and an HbA1c level above 8% in the past year." The agent's perception depends on its understanding of database schemas and its ability to translate natural language instructions into executable code, such as Python or SQL. Frameworks including EHRFlow [190] and TrustEHRAgent [160] demonstrate this capability by converting user queries into structured code for database interaction. In this setting, the agent functions as a data analyst. EHRFlow uses multiple agents to decompose complex physician requests into smaller analytical tasks, while TrustEHRAgent acts as a verifier that estimates confidence in its outputs and abstains from uncertain predictions to maintain clinical safety.

Open Challenges. Major challenges include ensuring reliability and avoiding code generation errors that can lead to clinically significant outcomes, as discussed in TrustEHRAgent [160]. Bridging the gap between physicians' high-level analytical intent and executable code remains an open problem, which EHRFlow [190] seeks to address. Privacy concerns further require that all computations be executed locally without exposing sensitive patient data.

4.1.3 Clinical Notes. Clinical notes are textual records that capture longitudinal patient narratives. They shift the agent's focus from real-time interaction to temporal extraction and synthesis. The agent must reconstruct a coherent clinical timeline from multiple time-stamped entries written by different clinicians over many years, identifying key events and assembling them into structured sequences. For example:

[Time: -10y, Symptom: "forgetfulness", Source: PCP] → [Time: -5y, Symptom: "getting lost", Source: Neurology, Reporter: wife] → [Time: -2y, Symptom: "apathy", Source: Psychiatry]

This temporal reasoning defines the agent's role as a diagnostic assistant integrating information across specialties. Infherno [42] transforms free-text notes into standardized FHIR resources, requiring both information extraction and validation against external terminologies such as SNOMED CT. CARE-AD [93] employs multiple specialized agents to analyze longitudinal clinical notes, each identifying domain-specific indicators of chronic conditions such as Alzheimer's disease to produce an aggregated risk assessment.

Open Challenges. The main challenge lies in longitudinal reasoning, as early disease indicators are often sparse and distributed across years of documentation, as noted by CARE-AD [93]. Ensuring adherence to complex output schemas such as FHIR also presents difficulties [42]. Additionally, clinical notes contain implicit knowledge and shorthand expressions, requiring contextual understanding for accurate interpretation.

4.1.4 2D Medical Imaging. For healthcare agents, 2D medical imaging introduces a complex perceptual modality. Visual understanding is enabled by a vision encoder that converts pixel data into numerical embeddings, or tokens, which are processed jointly with text by the LLM. The nature of perception and downstream reasoning varies substantially across imaging domains. In chest X-rays (CXR), the agent perceives a single global view and often acts as a radiology assistant, as in MedRAX [34]. The model follows a Reason–Act

process: it first plans diagnostic steps and then invokes specialized tools such as classifiers, grounding models, or segmenters to identify and localize abnormalities. In pathology, whole-slide images (WSIs) pose a different challenge due to their gigapixel scale. Perception becomes an active exploration process that mimics a pathologist's workflow. Systems such as PathChat+ [18] and CPathAgent [167] employ a hierarchical structure where a supervisor agent inspects a low-resolution overview, generates hypotheses, and directs explorer agents to zoom into selected regions for detailed examination. The LLM coordinates these agents, integrating observations across magnifications until a diagnostic conclusion is reached.

Open Challenges. For CXR agents, major issues involve fusing heterogeneous tool outputs and resolving conflicts among them. For pathology agents, efficient navigation of gigapixel slides remains difficult, requiring intelligent search strategies to identify diagnostic regions without exhaustive scanning. Multi-scale reasoning, linking cellular features at high magnification with global tissue structures, remains an open research frontier.

4.1.5 3D Medical Imaging. Three-dimensional imaging modalities such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) extend the agent's perceptual space from single images to volumetric structures. The agent may process these data either as unified 3D volumes or as sequential 2D slices, depending on the vision encoder and token compression strategy. In this environment, the agent functions as an anatomy-aware radiologist capable of understanding complex spatial relationships. CT-Agent [113] adopts a modular design where fine-tuned LoRA plugins serve as specialized tools for different anatomical regions (e.g., lung, heart, pleura). A planning module identifies the target anatomy based on the user's query and dynamically invokes the corresponding tool for localized analysis, enabling report generation and region-specific visual question answering. Agents can also operate as quality control or reconstruction specialists. AgentMRI [147] detects motion or noise artifacts in MRI volumes, determines the degradation type, and selects the appropriate correction model to reconstruct clean images. Other systems, such as MESHAgents [219] and CTPA-Agent [228], integrate 3D imaging features with clinical variables for tasks including phenome-wide association and survival prediction.

Open Challenges. Processing large volumes introduces substantial computational costs due to the number of visual tokens. Capturing cross-slice spatial continuity remains difficult, as agents must recognize that findings across adjacent slices correspond to a single lesion rather than independent abnormalities. Anatomical complexity also demands high specialization; general-purpose models often fail to distinguish subtle regional variations. Modular, tool-based designs such as CT-Agent address this limitation but remain computationally intensive. A broader challenge is integrating volumetric interpretations with other clinical data to achieve coherent, clinically meaningful reasoning.

4.1.6 Time-Series Signals. Physiological signals such as ECG, continuous glucose monitoring, and wearable sensor data capture dynamic temporal processes. For a healthcare agent, perceiving these signals involves structured analysis rather than direct ingestion of raw numerical streams. This modality is inherently temporal. For example, when asked "Am I getting more fit recently?", the agent interprets the query and formulates an analytical plan, as in the Personal Health Agent [58]: (1) filter activity data from the past three months, (2) compute average running speed per week, and (3) fit a regression model to detect trends. Translating natural language intent into a formal analysis plan is the core perceptual act.

This environment positions the agent as a bridge between qualitative queries and quantitative computation. In PHA's Data Science Agent [58], the system functions as a computational analyst that executes its own analytical plan through Python code for data transformation and statistical testing. In contrast, the LLM-Powered Agent for PPG [38] acts as a signal-processing orchestrator, diagnosing artifacts in photoplethysmography signals and invoking appropriate denoising or correction pipelines.

Open Challenges. A central challenge is the limited numerical reasoning and signal-processing capability of LLMs. Directly feeding long or continuous waveforms often exceeds token limits and leads to unreliable results [38]. Contextual grounding is another difficulty: interpreting metrics such as heart rate variability requires comparisons with personal history, population baselines, and concurrent physiological data. Finally, tool-based pipelines depend on accurate signal diagnosis. Misidentifying motion artifacts as noise, for example, can lead to invoking the wrong processing tool and invalidate the entire workflow.

4.1.7 Genomics & Biomarkers. Genomic and biomarker data form a symbolic modality requiring knowledge-based reasoning rather than visual or conversational perception. Inputs such as gene sets, mutation lists (e.g., TP53, KRAS), or multi-omics profiles hold limited meaning without contextual information. The agent's key perceptual act is to recognize biological entities and retrieve their functional relationships from external knowledge sources. Frameworks such as HEAL-KGGen [240] and GeneAgent [184] autonomously ground LLM outputs in curated biomedical resources by querying medical knowledge graphs or domain databases.

This environment situates the agent as a computational biologist automating bioinformatics workflows. AI-HOPE [202] converts natural language queries (e.g., "Compare survival outcomes for FOLFOX-treated patients with and without KRAS mutations") into executable code that performs statistical analysis and generates survival plots. HEAL-KGGen [240] extends this idea through a multi-agent architecture where a generalist routes tasks to specialized agents (e.g., genomics, proteomics). GeneAgent [184] further introduces self-verification, proposing biological hypotheses and validating them against public databases to reduce factual errors.

Open Challenges. A major challenge is factual hallucination, where an agent may incorrectly assert gene–disease links. Self-verification mechanisms, as in GeneAgent, directly mitigate this risk. Knowledge integration also remains difficult, requiring consistent fusion of LLM-based reasoning with structured biomedical graphs and multi-omics data. Finally, it is hard to translate a clinician's high-level question into precise analytical workflows. As exemplified in AI-HOPE, it demands accurate mapping between natural language intent and executable bioinformatics pipelines.

4.1.8 Audio. Audio adds temporal and affective information to the agent's inputs. Perception proceeds in two stages. First, a speech-to-text system transcribes spoken language for LLM processing, as in ANNA [130] and automated survey agents [74]. Second, the agent analyzes the waveform to extract paralinguistic features. Prior

work [1] uses speech emotion recognition to infer states such as sadness, anger, or joy from prosody, pitch, and intensity.

This dual pathway supports empathetic and diagnostic interaction. In therapeutic settings, the agent adapts responses to both content and detected affect, for example, shifting to supportive guidance when sadness is inferred.

Open Challenges Transcription accuracy remains a bottleneck, since word error rate directly affects downstream interpretation [74]. Emotion cues are subjective and culturally variable, which limits reliability [1]. Subtle diagnostic markers in speech are easily confounded by noise, fatigue, and recording conditions, demanding robust models for real-world clinical use [130].

4.2 Multi-Modality

Multimodality represents the most advanced stage of healthcare agents, where perception extends beyond a single data source to a unified understanding of the patient. The agent must process and align heterogeneous inputs—images, clinical text, time-series signals, and genomic data—and identify semantic links among them. For instance, it should connect a mutation identified in a genomic report with the corresponding morphological features observed in a pathology slide. Effective reasoning thus depends on constructing a coherent cross-modal representation of the patient's state.

In this setting, the agent functions as a coordinator within a virtual multidisciplinary team. Frameworks such as MAM [233] and MedAgent-Pro [185] decompose diagnostic reasoning into specialized roles. In MAM, a generalist triages cases and routes data to domain-specific agents (e.g., a radiologist for images, a specialist for clinical notes), while a director synthesizes their outputs into a final decision. MMedAgent [87] operates as a tool orchestrator that selects appropriate models based on the modalities involved—such as invoking a segmentation tool for images or a report generator for text. This coordination enables integrated decision-making beyond any single modality. For example, MedAgent-Pro combines qualitative assessments from fundus images with quantitative measures like cup-to-disc ratios to improve diagnostic accuracy.

Open Challenges. Key challenges include evidence fusion and conflict resolution, such as reconciling contradictions between textual and visual findings. MedAgent-Pro addresses this by weighting the reliability of each source before synthesis. Semantic grounding across modalities is another difficulty, requiring alignment between symbolic data such as visual or physiological features. Finally, the scarcity of comprehensive multimodal datasets—where imaging, genomics, and outcomes are consistently linked—remains a major limitation for developing and validating these systems.

5 Agent Architecture and Capabilities

Compared with single-pass LLMs, agentic systems introduce grounding and control loops that support more reliable and auditable behavior. We describe their operation in healthcare by organizing core components in the order they are typically invoked during an agent's workflow. The process begins with **Knowledge Retrieval**, which provides verifiable evidence from medical literature, patient databases, and structured guidelines. We separate retrieval from tool use because retrieval focuses on acquiring symbolic or textual knowledge, whereas tool use executes external computational utilities such as calculators or segmentation tools. **Memory**, encompassing both short-term context and long-term records, supplies persistent information that complements retrieval. It preserves patient-specific details, maintains continuity across interactions, and stabilizes downstream reasoning. These inputs support **Planning and Reasoning**, where the agent decomposes goals into sub-tasks, forms intermediate hypotheses, and generates auditable rationales. **Tool Use** then executes these plans by invoking external resources and recording the provenance of results. **Simulation** provides controlled, role-conditioned settings for development and evaluation, enabling assessment under clinically meaningful scenarios. **Multi-Agent Collaboration** builds on all prior components by assigning specialist roles for coordinated discussion, debate, cross-checking and decision-making. Table 2 summarizes these architectural elements, representative systems, and the technical contributions of each category.

5.1 Knowledge Retrieval

A key advance from single-pass LLMs to agentic systems is the addition of explicit knowledge retrieval. This is especially critical in healthcare, which requires risk-aware and trustworthy behavior. Among grounding approaches, retrieval-augmented generation (RAG) is most common, in which the agent conditions outputs on authoritative evidence. Knowledge graphs (KGs) canonicalize entities and typed relations, support compositional reasoning, and encode safety constraints; they also strengthen RAG through entity normalization and constraint-guided generation. Beyond retrieval, KGs can support post-generation verification, improving risk awareness. In practice, two variants are common: KG-anchored RAG, which links mentions to canonical entities before retrieval; and text- or tool-anchored RAG, which retrieves directly from clinical sources without a KG. We next describe these variants and knowledge-based verification that leverages retrieval.

5.1.1 KG-anchored RAG. KGs strengthen RAG by enabling entity canonicalization, relation-aware retrieval, and constraint-aware generation. Such systems explicitly map user queries and intermediate variables to canonical biomedical entities and traverse the graph to assemble coherent evidence. A typical workflow includes entity linking, subgraph expansion, retrieval of supporting passages, grounded generation, and span-level citations. For example, AMG-RAG [144] and RUGGED [134] automate KG construction and updating, link new findings to evolving concepts, and couple graph traversal with external retrieval, reducing hallucinations and improving diagnostic accuracy over knowledge-deficient baselines. MedGraphRAG [192] further extends GraphRAG [33] with a triple graph over entities, papers, and lexicon definitions and introduces U-Retrieval. ClinicalRAG [108] also integrates KG-anchored RAG with a multi-agent system, retrieving knowledge from heterogeneous medical knowledge. Anchoring RAG to KGs improves disambiguation, compositional reasoning, and safety checks, though KG curation is costly and may lag behind evolving clinical knowledge.

5.1.2 Other RAG. Text- and tool-anchored RAG retrieves directly from authoritative sources and operational tools, avoiding explicit KG construction while retaining evidence-conditioned generation.

Architecture	Capability	Representative Systems	Key Contribution
Knowledge Retrieval	Grounding agent decisions in curated evidence	AMG-RAG [144] Path-RAG [121] AI-VaxGuide [215]	Automates integration of existing knowledge into generation to reduce hallucinations and improve diagnostic accuracy over LLM-only baselines.
Memory	Maintaining longitudinal context	AMC [105] REMI [140] EHRAgent [156]	Maintains explicit memory of patient-history or experiment-history to improve decision-making.
Planning and Reasoning	Decomposing tasks, reasoning decisions	STELLA [72] CT-Agents [113] MeNTi [238]	Implements explicit task-decomposition and structured reasoning, integrating tool use and intermediate feedback to iteratively refine plans and conclusions.
Tool Use	Extending the agent's capabilities by integrating and selecting external software tools	GeneAgent [184] DrugAgent [106] AgentMD [71]	Incorporates and dynamically selects domain-specific tools into the agent architecture.
Simulation	Creating and maintaining a controlled clinical persona for the agent	Talk2Care [206] BehaviorSFT [77] Medco [188]	Develops and evaluates methods to encode and enforce agents' role identity.
Multi-Agent Systems	Enabling collaboration between multiple agents	MAC [23] MAM [233] PathChat+ [18]	Routes tasks to domain-specialized agents while using coordination algorithms to formalize team behaviors.

Table 2: Overview of agent architectures, their primary capabilities, representative systems, and key contributions.

Common patterns include query rewriting, dense or sparse retrieval, evidence-conditioned generation, citation of supporting passages, and selective tool invocation. Representative systems include AI-VaxGuide [215], which organizes vaccination protocols into an interactive knowledge base and delivers context-sensitive guidance via agentic RAG; LLM-based CDSS [127] offering medication-safety decision support that applies RAG across multiple specialties to improve prescribing workflows; pharmacovigilance pipelines such as MALADE [30] that orchestrate agents over literature, labels, and regulatory interfaces; and CLADD [82], a collaborative agent framework that retrieves from biomedical knowledge bases to contextualize molecules without domain-specific fine-tuning. End-to-end agentic training further optimizes retrieval and reasoning policies, as demonstrated by Deep-DxSearch [226], which learns a unified retrieval-generation policy over a large medical corpus. Although these methods lack graph-level constraints for multi-hop consistency and the ability of cross-source integration, such systems' strengths include flexibility and rapid adaptation to evolving sources.

Knowledge-based verification Beyond using knowledge retrieval in generation, several systems verify and revise outputs by leveraging retrieved text or accumulated experience. MDTeamGPT [20] stores validated answers and error reflections in vector memories that are retrieved to steer later rounds and are filtered by an LLM-based reviewer. Similarly, ReflecTool [100] builds a long-term memory of successful tool-use trajectories and applies a verifier to improve action choices using prior tool-wise experience. These knowledge-based verifiers improve safety and traceability of systems, though they do not provide formal graph- or rule-level consistency checks.

5.2 Memory

Memory enables agents to maintain consistency and adaptation across extended clinical interactions. Clinical scenarios often unfold over multiple encounters, making it essential for agents to preserve continuity and personalization. Memory modules not only preserve past dialogues but also help coordinate domain knowledge, retrieve external resources, and adjust communication styles. Recent research explores memory from several perspectives, most commonly divided into episodic memory, semantic memory, and long-term or demonstration memory.

5.2.1 Episodic Memory: Capturing Specific Interactions and Context. Episodic memory forms the backbone of long-horizon agents by recording the *what, when, and where* of specific interactions. Researchers argue that for agents to maintain continuity in clinical dialogues, they require an explicit episodic memory system that goes beyond the limitations of standard context windows [66, 137]. In practice, implementing this memory enhances the quality of agent-patient interactions. For instance, health chatbots [73] equipped with long-term memory have been shown to foster a greater sense of intimacy and increase users' willingness to share health information. Similarly, in diagnostic simulations for mental health, AMC [105] samples critical information from dialogue history and medical records, and has achieved substantially higher accuracy. While these applications highlight the benefits for personalization and clinical performance, they also underscore the associated challenges of managing memory scope and ensuring patient privacy.

5.2.2 Semantic and Knowledge-Centric Memory. Semantic memory focuses on organizing general facts, concepts, and relations, often leveraging structured formats like knowledge graphs to make agent reasoning more interpretable. By building a causal knowledge graph of a patient's life events, an agent can generate transparent and personalized medical advice that is grounded in clear reasoning pathways [140]. A powerful emerging trend is the integration of semantic and episodic memory. These hybrid frameworks combine structured, long-term knowledge with specific user interaction histories, allowing the agent to dynamically adapt its behavior to align with user preferences and context [220]. This synergy enables agents to be both knowledgeable and personally attuned.

5.2.3 Long-Term and Demonstration Memory. Another approach involves using long-term memory to store a library of successful past experiences or demonstrations that guide future decisions. In the context of reasoning over electronic health records, an agent can improve its planning by dynamically retrieving relevant past cases from memory to inform its current task [156]. This principle also applies to complex medical tool use, where an agent can learn from a repository of successful problem-solving trajectories to select the right tool and correct its actions [100]. At a larger scale, this concept can be architected into a central memory hub that serves multiple specialized agents, integrating accumulated experience with up-to-date medical knowledge to tackle complex challenges like rare disease diagnosis [224]. By learning from what has worked before, these agents can more effectively navigate complex scenarios.

5.3 Planning and Reasoning

In clinical settings, planning denotes the decomposition of tasks such as diagnosis, therapy selection, and follow-up into actionable steps. Reasoning denotes the analysis of evidence, the application of domain knowledge, and the formation of testable hypotheses. Challenges include high stakes and uncertainty, incomplete or noisy inputs, evolving patient states and literature, and requirements for transparent justification. Recent systems address these challenges by embedding explicit reasoning and planning within single- and multi-agent frameworks [75].

5.3.1 Planning and Task Decomposition Frameworks. Effective agents factor long-term clinical objectives into tractable sub-tasks and reusable procedures, performing **goal decomposition**. For instance, STELLA [72] treats planning as a library of evolving templates, selecting and refining patterns and tools for tasks such as biomedical QA or experimental design. Simulated-patient systems (AIPatient [211]) couple multi-agent roles with retrieval and a knowledge graph to coordinate summarization, question answering, and evaluation within a unified plan. Operational planners extend the same principles beyond diagnosis: MedScrubCrew [146] formulates appointment scheduling as constrained matching and integrates a knowledge graph with agent policies to align patient and provider profiles. Modality-specific decompositions follow naturally; CT-Agents [113] partitions volumetric interpretation by anatomy and dispatches specialized reasoning paths with compact token representation. Long-term memory modules further link plans across encounters, encouraging continuity and stable defaults over time [73]. Taken together, these systems frame planning as program-like composition, with templates, constraints, and memory supporting reuse and adaptation.

Action-oriented planning and reasoning. A complementary design interleaves planning with tool execution so that observations revise subsequent steps. For example, MeNTi [238] introduces nested tool calling in which a meta-controller selects calculators, fills slots, and manages units for clinical computations; MMedAgent [87] trains agents to choose among modality-specific tools for tasks spanning diagnosis and report generation. In procedural domains, SurgBox [191] combines retrieval with a surgical copilot and a long–short memory mechanism to balance immediate assistance with persistent knowledge. Graph-aligned approaches follow the same action–reflection loop: KGARevion and KERAP interleave reasoning with calls to knowledge graphs and calculators under a ReAct-style scheme [164, 195]. Domain-focused agents such as GeneAgent [184] act on retrieved gene-set statistics and revise explanations after verification. Across these designs, planning is tightly coupled to execution, with tool feedback and verification signals guiding iterative refinement.

5.3.2 Reasoning: Enhancing Decision Depth and Reliability. **Sequential reasoning** models emulate a clinician's step-by-step thought process. Chain-of-Diagnosis (CoD) [19] converts diagnostic inference into a stepwise program and produces a calibrated disease distribution. Role-structured multi-agent dialogues [75] assign complementary responsibilities to reduce bias and expose assumptions. CareCall [73] further utilizes long-term memory modules to support continuity across encounters, though such a design raises privacy and governance considerations.

Exploratory and structured reasoning. More complex medical problems require exploring multiple reasoning paths. Graph-based approaches address this by expanding hypotheses and consolidating evidence across branches. Tree-of-Reasoning [135] expands and evaluates differential diagnoses with cross-branch validation. For biomedical question-answering, ESCARGOT [115] organizes thoughts as a dynamic graph aligned with a knowledge graph to reduce unsupported claims and improve transparency. For radiology, CT-Agent [113] decomposes volumetric analysis into organ-specific sub-tasks and uses a planner to dispatch reasoning paths with exemplar retrieval. Furthermore, STELLA [72] augments structured reasoning with a self-evolving template library and a dynamic tool ocean. Graph-anchored retrieval frameworks such as AMG-RAG and Medical Graph RAG integrate reasoning with knowledge-graph traversal for evidence alignment [144, 192].

Reflective and self-correcting reasoning. Safety-critical healthcare applications require agents that can critique and revise their outputs. One line of work verifies claims against trusted resources: GeneAgent [184] queries domain databases to validate gene-function outputs, and KGARevion [164] extracts latent triplets and checks them against a knowledge graph. A second line integrates verification into the reasoning process via coordinated roles: KERAP [195] aligns linkage, retrieval, and prediction agents on a shared graph to produce calibrated outputs. A third line emphasizes continual improvement: RAG-KG-IL [212] combines retrieval with incremental updates to the knowledge base to reduce hallucination over time. Complementarily, evaluation-driven reflection closes the loop: GEMA-Score [222] parses generated radiology reports and provides explanatory feedback, while AutoCT [102] applies search-based refinement to iteratively improve feature sets and decisions. Together, these approaches formalize self-critique and revision as explicit steps, improving reliability and traceability.

5.4 Tool Use

In clinical settings, text-only responses are insufficient; agents must invoke external tools to execute multi-step tasks. Tool use transforms plans into verifiable procedures and yields three practical benefits: improved safety and reliability through grounded computations with full traceability [7, 31, 215]; stronger explainability and accountability via auditable call logs, provenance tags, and intermediate results [49]; and higher productivity by automating

routine steps such as note pre-population, templated synthesis, guideline-concordant ordering, and scheduling [7, 32, 92, 97, 215].

5.4.1 Types of Tool Use in Healthcare. **Retrieval and knowledge tools** sit upstream and ground reasoning in external evidence. Classical RAG systems can query medical literature (PubMed, clinical guidelines), patient EHRs, and specialized databases to retrieve related evidence [31, 215]. STRID [31] further introduces an advanced implementations orchestrate multiple retrieval strategies, combining dense retrieval for semantic similarity and sparse retrieval for keyword matching. Moreover, domain-specific retrieval extends beyond literature: genomic agents query variant databases (ClinVar, gnomAD) [139, 184], drug discovery agents access compound libraries [106, 159], and diagnostic agents retrieve similar cases from institutional repositories [49]. **Computational tools** operate midstream by transforming inputs into quantitative estimates that can drive decisions. Risk stratification agents employ validated calculators (Framingham, ASCVD) [7], radiotherapy agents interface with dose optimization algorithms [181], and genomic agents invoke sequence aligners and variant effect predictors [139, 184]. These tools provide numerical outputs with established clinical interpretation thresholds, reducing recommendation ambiguity. **Workflow integration tools** enact decisions downstream and connect agents to care delivery. FHIR APIs [32, 42] allow agents to read structured clinical data and write orders, care plans, and documentation. Systems like IMAS [44] orchestrate triage workflows, while conversational agents [97] pre-populate intake forms. Privacy-preserving architectures implement role-based access control, ensuring minimal necessary privileges [32]. Orthogonally, **simulation and verification tools** ensure reliability before deployment. Platforms like AgentClinic [151] provide simulated patient encounters, while verification tools check internal consistency [184] and cross-validate against clinical guidelines [215].

5.4.2 Agentic Progression. Agent sophistication is measured by the ability to *select, sequence, and adapt* tool use as contexts evolve, marking a shift from scripted automation to genuine autonomy. At the simplest end, **Single-step invocation** is a deterministic tool use in which an input is mapped to a single tool; systems for structured clinical assessments [130] and radiology report generation [49] often trigger terminology services in this way, which is effective for narrowly defined tasks yet limited when the first call leaves uncertainty. Moving beyond this, **Sequential workflows** chain multiple tools in predefined sequences: genomic interpretation agents [139, 184] proceed from variant calling, through frequency filtering and functional prediction, to clinical association search and report generation, while radiotherapy planning [181] coordinates dose calculation, complication modeling, and quality metric evaluation in iterative cycles; progression is output-dependent and continues or terminates once sufficient certainty is achieved. Further, **Dynamic selection with reasoning** chooses which tools to invoke based on the evolving clinical context; multi-agent diagnostic systems [224, 231] switch among genetic databases, metabolic analyzers, and case repositories as hypotheses shift, and AgentMD [71] learns tool-selection policies from EHR data, identifying calculators and tests that historically yield the greatest diagnostic value for a given presentation, in line with pretest probability clinical reasoning. At a higher level, **Self-verification and adaptive orchestration** implements closed-loop control: GeneAgent [184] automatically invokes verification tools, detects inconsistencies, retrieves additional evidence, and revises conclusions, reporting a 40% reduction in hallucinations; rare-disease agents [224] maintain competing hypotheses with parallel tool invocations and adjudicate among them, and error-recovery mechanisms detect tool failures and retry with alternatives. Finally, **Meta-learning and discovery** captures systems that learn which tools work best in particular scenarios and synthesize new combinations: BioScientistAgent [217] uses reinforcement learning to discover optimal query sequences for drug repurposing, outperforming hand-designed workflows; looking ahead, agents may propose new clinical tools, suggest novel computational methods, and generate synthetic training data, thereby improving the toolkit rather than merely using it.

5.5 Simulation

Persona design is a core technical component for ensuring that AI agents in healthcare are controllable, trustworthy, and effective. A well-designed persona aligns an agent's behavior with clinical norms and user expectations, a critical requirement in high-empathy domains. As LLM agents expand in healthcare, persona-driven conditioning is essential for maintaining system alignment and safety [142]. This section dissects this technical lifecycle, from instantiation and regulation to validation and frontier challenges.

5.5.1 Persona Instantiation and Regulation. Persona instantiation translates an abstract clinical role into a computational representation, often drawing from established practices like Cognitive Behavioral Theory [83]. These personas, such as patients or providers, are already used to improve clinical communication in applications like Talk2Care [206]. Technically, this is achieved through explicit and parametric encoding. Explicit encoding via prompting is the most flexible method, ranging from simple instructional prompts defining a character's traits [199] to more sophisticated techniques like Behavioral Tokenization, which uses special tokens for fine-grained behavioral control as introduced in BehaviorSFT [77]. Other methods, such as LAPI [96], use objective-constrained prompting to align responses with a professional identity.

However, a well-defined persona is only effective if regulated during dynamic interactions. To maintain consistency and combat "persona drift", a risk highlighted by memory failures in commercial chatbots [111], systems require robust memory architectures, as discussed in AnnaAgent [180]. Beyond consistency, effective regulation requires dynamic adaptation, typically achieved by modeling the user's state to inform the agent's strategy. This allows agents to selectively apply clinical techniques like Motivational Interviewing based on user progress [204] or calibrate their feedback style to a clinician's expertise level [15].

5.5.2 Persona Validation and Frontier Challenges. Validating a persona requires a multidimensional protocol spanning technical fidelity, user perception, and task performance. On the technical axis, benchmarks quantify behavioral fidelity against predefined strategies [77]. On the human axis, granting users control over persona configuration improves perceived trust and engagement [227]. These considerations should connect to task outcomes: for example,

Medco [188] employs role-specific agentic copilots to deliver training simulations for medical students, while VChatter [218] adopts therapeutic roles to support exposure-based interventions.

Despite this progress, several frontier challenges remain. First, a persona generalization gap persists: personas validated in one cultural context may fail in another, as generative agent societies diverge from real-world public health attitudes [62]. This gap elevates risks of bias, manipulation, and harm, motivating stronger evaluation frameworks and risk assessments [111, 162, 194]. Second, system-level complexity complicates multi-agent deployments: coordinating heterogeneous personas for composite reasoning remains open, including in role-playing expert systems [171]. Third, embodied interfaces introduce consistency requirements between verbal and non-verbal behavior [117]. Across these settings, practical deployments continue to benefit from human oversight to enforce persona boundaries and safety constraints [86].

5.6 Multi-Agent Systems

LLM-based multi-agent systems (MAS) provide an orchestration layer over the primitives outlined above. By assigning specialist roles and regulating communication, MAS supports collaborative decomposition, cross-checking, and consensus formation. Coordination typically uses role specifications, shared state or memory, and judge or mediator agents that reconcile divergent hypotheses while preserving provenance. The paradigm's utility is demonstrated across diverse domains, from multi-modal cardiology diagnostics [234] and debate-based mental health counseling [84] to synthetic data generation for medical dialogues [5]. Comparative studies indicate advantages over single-agent baselines on complex tasks that benefit from complementary expertise and structured disagreement resolution [10]. In practice, MAS integrates knowledge grounding, planning, and tool use within each role and exchanges intermediate results through constrained protocols, improving robustness and transparency in workflows that mirror team-based clinical practice.

5.6.1 Architectural Patterns & Organizational Structures. Multi-agent architectures in healthcare mainly differ by their organizational structure and distribution of decision authority, falling into hierarchical/centralised, flat/decentralised, and hybrid patterns.

Hierarchical & Centralised Architectures In hierarchical MAS, a central coordinator decomposes tasks and synthesises outcomes from subordinate agents. This structure naturally mirrors clinical team dynamics, where a "manager" or "director" agent assigns specialised diagnostic tasks to workers in fields like cardiology, forensic pathology, and general medicine, before aggregating their analyses into a final report [154, 233, 234]. This pattern is also widely adopted to ensure safety and guideline adherence. For instance, systems often implement tiered oversight where junior agents propose actions, mid-tier agents check for harm, and a senior agent, or a human doctor, provides final approval [20, 27, 78]. Beyond clinical reasoning, centralised architectures orchestrate complex AI pipelines, with a controller coordinating agents dedicated to data handling and model training [39], or a meta-agent synthesising information retrieved by multiple agents from EHRs [186]. These designs ease coordination and align with institutional oversight. Risks include single-point failure, reduced diversity of perspectives, and sensitivity to errors at the coordinator.

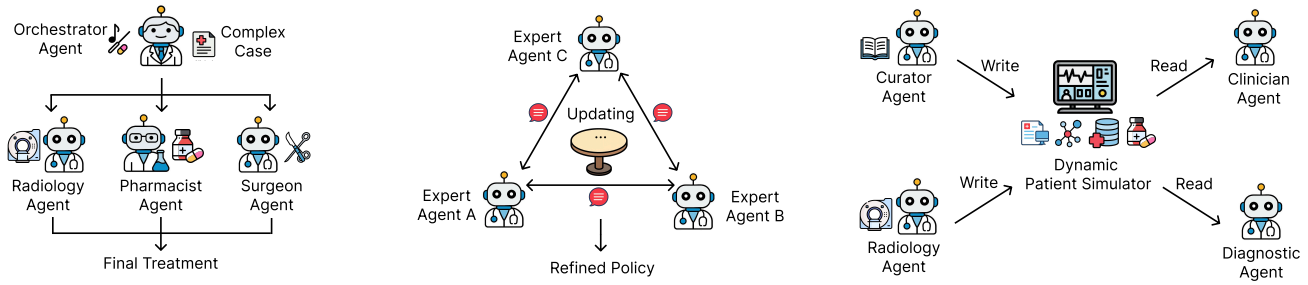
Flat & Decentralised Architectures Flat architectures empower agents with equal status to engage in peer-to-peer debate or voting. This approach is frequently used to tackle complex clinical reasoning, where multiple specialist agents debate evidence to identify conditions, disambiguate similar diseases, or provide empathetic counselling [84, 85, 225, 231]. The same peer-review structure is effective in multimodal reasoning, such as radiology VQA, where context, reasoning, and verification agents collaborate to finalise an answer [207]. Beyond diagnostics, decentralised debates can enhance research creativity through continuous knowledge exchange [213] or improve robustness by intentionally injecting dissent to overcome silent agreement bias in group discussions [183]. Although these frameworks promote viewpoint diversity and resilience, they require sophisticated consensus mechanisms.

Hybrid Architectures. Hybrid designs blend hierarchical coordination with peer-to-peer collaboration to balance efficiency and diversity. Some systems create specialised peer groups for tasks like data extraction and captioning, which are then overseen by a higher-level filtering or review agent [168, 216]. Others combine different functional agents, such as those for retrieval, knowledge graph integration, and scoring, to produce more reliable and explainable outputs [212, 222]. A third pattern uses hybrid structures to simulate cognitive modularity, where distinct agents handle analysis, synthesis, and validation in a coordinated workflow, or engage in an "inner dialogue" to guide users. These approaches have shown significant accuracy improvements in complex domains like neurological problem-solving [24, 135, 161]. Hybrids thus capture both the coordination advantages of hierarchies and the creative diversity of decentralised debates.

5.6.2 Coordination & Communication Mechanisms. Coordination is how multi-agent systems decompose complex tasks, interact, and converge on solutions. As shown in Figure 3, three dominant mechanisms appear: (a) task decomposition and allocation, (b) debate-based consensus, and (c) environment-driven coordination.

Task Decomposition & Allocation. Hierarchical MAS naturally excel at breaking down complex workflows into manageable subtasks. This is often seen in diagnostic pipelines, where a primary agent decomposes a case into sub-problems for specialist agents (e.g., GP, radiologist) and integrates their findings, as introduced by MAC [23], MAP [27], MAM [233], and FEAT [154]. The approach also extends to automating complex technical and analytical workflows, such as configuring machine learning pipelines[39], generating descriptions for pathology images[80], or performing thematic analysis of interview transcripts [198]. More advanced systems introduce dynamic team formation, where specialist agents can join or leave the collaboration based on evolving information needs, enabling a more adaptive diagnostic process [153].

Debate-Based Consensus. Debate mechanisms leverage the argumentative and reasoning capabilities of LLMs to refine solutions. In clinical settings, structured debates help disambiguate similar diseases, formulate empathetic counselling responses, or synthesize evidence from data-driven and knowledge-driven perspectives [84, 225, 231]. A key benefit of this adversarial process is the mitigation of biases and hallucinations. By encouraging agents to challenge each other, systems can avoid premature convergence and surface conflicting evidence [85]. Some frameworks explicitly



(a) Task Decomposition and Allocation **(b) Debate-Based Consensus** **(c) Environment-Driven Coordination**

Figure 3: Three coordination mechanisms for healthcare multi-agent system: (a) Task decomposition and allocation via an orchestrator assigning sub-tasks. (b) Debate-based consensus among peer agents. (c) Environment-driven coordination through a shared, persistent environment (e.g., a patient simulator).

introduce a "dissenter" agent to disrupt groupthink [183], while others use dedicated validation agents to cross-examine diagnoses, both leading to more robust and accurate outcomes [161, 207].

Environment-Driven Coordination. Many MAS coordinate asynchronously by reading and writing to a shared environment that persists state and mirrors clinical workflows. One line of work employs dynamic patient simulators: agents conduct note-grounded dialogues with synthetic patients [5] and query MIMIC-based simulators for vitals and labs, updating a shared case representation over time [153]. A second line uses shared knowledge structures, such as PathChat+ [18], so digital pathology agents log interim report findings to refine diagnoses collaboratively, while curator agents like RAG-KG-IL [212] incrementally add literature-derived facts to a central KG that others use as a stable context to curb hallucinations. Finally, existing structured records function as the environment: agents extract features from radiology reports [222], retrieve EHR histories to condition recommendations [186], and iterate over imaging repositories to surface confounders for association studies [219]. Collectively, these designs replace direct messaging with a persistent, queryable state that supports reproducible multi-agent clinical workflows.

5.6.3 Key Technical Challenges & Future Directions. Despite impressive progress, LLM-based MAS face several open challenges. **Shared Context & State Consistency** remains central: systems that externalize state with structured reasoning trees or validation agents mitigate contradictions relative to ad-hoc histories [85, 135, 161], and integrating persistent knowledge graphs provides a common factual substrate that reduces hallucinations [212]. **Communication Overhead & Protocol Design** calls for constrained interfaces—well-defined APIs, concise list-based messages, and tiered pipelines—to control cost and error propagation [39, 78, 80]; in parallel, dynamic team formation improves efficiency [153]. **System-Level Evaluation & Alignment** requires end-to-end metrics because component-level gains can degrade overall performance [10]; useful signals include expert-correlated scoring and guideline concordance across datasets [27, 222, 233], with layered oversight and deliberation to detect harms and mitigate bias [78]. **Agent Frameworks & Scalability** benefit from modular designs and benchmarked team composition [39, 80, 153], yet long-horizon coordination and history management introduce overhead in clinician-in-the-loop settings [20, 69, 198]; advancing scheduling, memory, and resource allocation remains a key direction.

Future work must focus on developing persistent memory for consistent patient context, designing efficient communication protocols, and devising comprehensive, system-level evaluation metrics that account for safety and ethics. Emerging directions like dynamic team formation [153], deeper integration of knowledge graphs [212], and fairness-aware deliberation [69, 78] are promising. By addressing these challenges, LLM-based MAS are poised to become integral partners in the future of healthcare delivery.

6 Applications Across the Healthcare Ecosystem

This section reviews practical applications of AI agents across the healthcare ecosystem. Unlike the preceding technical taxonomy, it is organized around the primary stakeholders these systems serve. This role-based view helps clinicians, patients, researchers, educators, and administrators identify relevant developments and understand how agent roles, autonomy, and human oversight differ across contexts. Table 3 summarizes the main application areas of healthcare agents, describing their typical roles, core capabilities, autonomy levels, and representative systems across stakeholders.

6.1 Supporting Doctors: Enhancing Clinical Workflow

Doctors are key decision-makers in complex clinical environments. For them, AI agents evolve from passive tools to active collaborators. These systems assist with adaptive decision-making by integrating distributed data and coordinating tasks. Their goal is to enhance clinical judgment while preserving human oversight. This section follows a typical workflow: diagnosis and decision support, documentation, and workflow automation.

6.1.1 Diagnosis & Decision Support. Clinical diagnosis relies on integrating patient data with medical knowledge. Recent agentic systems move beyond information retrieval and image analysis, treating each module as a callable tool to enable higher-level planning and reasoning within clinical workflows. This section outlines

Application Domain (Stakeholder)	Agent Role	Key Agentic Capability	Autonomy Level	Representative Systems
Diagnosis & Decision Support (Doctors)	Diagnostic Assistant MDT Simulator	Planning & Reasoning Tool-Use Multi-Agent Systems	Human-in-the-Loop (Mandatory Verification)	AgentMD [71] TxAgent [46] CT-Agent [113] ColaCare [186]
Clinical Documentation (Doctors)	Clinical Scribe Reporter	Knowledge Retrieval (RAG) Tool-Use (FHIR/API)	Human-on-the-Loop (Final Review & Sign-Off)	Infherno [42] PathChat+ [18] FRAME [209]
Workflow Automation (Doctors)	Digital Assistant Executor	Planning (NL-to-Code) Tool-Use	Procedural Autonomy (Rule-based Execution)	EHRFlow [190] CDR-Agent [193] SurgBox [191]
Mental Health & Counseling (Patients)	Therapeutic Counselor	Simulation (Persona) Memory (Long-Term)	Autonomous (with Crisis Escalation Protocol)	AutoCBT [196] CAMI [204] MAGI [81] VChatter [218]
Patient Education (Patients)	Health coach Educator	Knowledge Retrieval (RAG) Memory Simulation (Persona)	Autonomous (Informational / Coaching)	AI-VaxGuide [215] CareCall [73]
Biomedical Research (Researchers)	Research Assistant	Planning Tool-Use Knowledge Retrieval	Autonomous (Research Sandbox / Hypothesis Gen)	Stella [72] BioScientistAgent [217] DrugAgent [106]
Medical Education (Educators)	Virtual Patient Colleague	Simulation (Persona) Multi-Agent Systems	Autonomous (Simulation Environment)	AgentClinic [151] MEDCO [188]
Hospital Automation (Administrators)	Administrative Coordinator	Multi-Agent Systems Planning	Highly Autonomous (Back-Office Automation)	MedScrubCrew [146] ORDIRS-Agent [155]
Institutional Governance (Administrators)	Compliance Logistics Analyst	Knowledge Retrieval (RAG) Planning	Human-in-the-Loop (Expert Assistive)	TrialGenie [90]

Table 3: Overview of application domains, agent roles, autonomy levels, and representative systems for healthcare agents.

three directions: (1) constructing reliable informational foundations for diagnosis, (2) supporting diagnostic reasoning, and (3) using simulation to model patients or clinicians in diagnostic tasks.

Building the Informational Foundation. A core challenge in diagnosis is extracting accurate and relevant insights from fragmented clinical data. Agents address this by transforming unstructured inputs into standardized and interpretable representations such as FHIR resources [42]. They can extract key variables like Gleason scores from pathology notes [178] or psychosocial factors from interviews [125], creating structured patient profiles for downstream reasoning.

To make these profiles trustworthy, agents must keep their outputs aligned with the continually evolving body of medical knowledge. For example, AI-VaxGuide [215] integrates retrieval-augmented generation with immunization guidelines, while other systems ground treatment recommendations in standards such as those of the American Diabetes Association [3]. To reduce hallucination, GeneAgent [184] incorporates self-verification loops that cross-check claims against authoritative databases.

Using external tools further enhances agents' capacity to produce evidence-grounded insights, as shown by AgentMD [71], which converts free-text notes into structured risk scores using validated calculators, and TxAgent [46], which orchestrates over 200 resources, from drug databases to genomic repositories, to generate personalized treatment plans. Similarly, oncology-focused agents combine vision models, image segmentation, and guideline retrieval to improve diagnostic accuracy [41].

Together, these systems illustrate how agents convert raw data into structured, evidence-grounded recommendations, establishing a dependable foundation for clinical decision support [34, 37, 87].

Diagnostic Reasoning. Beyond data preparation, agents are increasingly applied to diagnostic reasoning. They adapt analytical strategies across modalities and coordinate information to support accurate and explainable decisions.

In pathology, agents analyze whole-slide images at multiple magnifications and distribute tasks among specialized modules to improve cancer classification and generate visually grounded reports [18, 80, 167]. Large-scale foundation models trained on millions of slides further enable zero-shot classification and biomarker prediction [152]. In radiology, CT-Agent [113] decomposes 3D volumes into anatomical sub-tasks for visual question answering, while AT-CXR [98] introduces adaptive triage that determines when to automate and when to defer to radiologists under uncertainty.

Beyond imaging, agents interpret genomic results into guideline-based recommendations [177, 202, 240] and use longitudinal health records to forecast disease trajectories [49, 93, 109]. Knowledge graphs further enhance reasoning by encoding links among symptoms, diseases, and treatments. They enable multi-step reasoning across biological entities, combine evidence from multiple sources, and integrate diverse data types to achieve diagnostic accuracy comparable to human experts [195, 234, 239].

Clinical Simulation. A key strength of AI agents is their ability to simulate interactions in complex healthcare environments. *Patient-centered simulations* create digital twins from longitudinal

records, enabling systems like EHR2Path [133] to forecast hospital trajectories instead of predicting isolated outcomes. OrganAgents [16] extend this concept by modeling patients' organs as interconnected physiological subsystems, allowing clinicians to test counterfactual scenarios (e.g., delaying treatment) and observe downstream effects. These simulations support reasoning in critical care and other high-risk settings [141].

Doctor-centered simulations replicate collaboration by coordinating virtual specialists who deliberate and converge on treatment strategies, achieving diagnostic accuracy comparable to human teams [161, 225]. Platforms such as ColaCare [186] demonstrate how agents from different specialties can coordinate to reach consensus on treatment decisions, thereby enhancing patient safety. Building on this direction, infrastructures like ClinicalLab [200] establish standardized benchmarks for evaluation, while frameworks such as DynamiCare [153] dynamically recruit virtual specialists to address case-specific needs. Together, these systems demonstrate how simulation integrates reasoning and collaboration into a cost-efficient and safe approach to clinical decision support.

6.1.2 Clinical Documentation & Reporting. Clinical documentation and reporting convert computational outputs into medico-legal records that must remain faithful to source data and verifiable against evidence. Early neural systems for radiology reporting [103] and conversational summaries [122] showed feasibility but lacked transparency and verification. These limitations motivated agentic approaches, which decompose documentation into modular, evidence-based workflows to ensure traceability and auditability. This shift is illustrated in the following two domains.

Structuring Clinical Interactions. Encounter notes are a natural entry point where the main challenge is turning unstructured, multi-party conversations into structured, actionable information. Agents go beyond simple transcription by generating context-aware representations that integrate external knowledge and reasoning [2, 173]. Some systems convert text directly into interoperable formats, such as Infherno [42], which embeds FHIR resources into documentation.

In real-world use, Mo was deployed with over 900 patients and improved clarity and physician satisfaction, with 95% of ratings classified as good/excellent [107]. Similar pipelines are also applied outside of documentation. FRAME [209], for instance, produces scientific manuscripts with quality comparable to human authors.

Grounding Reports in Visual Evidence. Another direction centers on generating formal reports from multimodal data, where each claim must be supported by evidence. Agents replicate the stepwise workflow of human experts, introducing checkpoints and tool integration for verification. In radiology, multi-agent "councils" distribute tasks for retrieval, image analysis, drafting, and review, improving accuracy and reducing hallucinations [208, 216]. Pathology adopts similar strategies at gigapixel scale. PathChat+ [18] uses hierarchical pipelines to navigate whole-slide images at multiple magnifications, keeping diagnostic reports visually grounded.

6.1.3 Workflow Automation. Clinical documentation records both patient states and clinician actions, forming a foundation for workflow automation. Building on this, intelligent agents streamline practice across three levels: data management, clinical knowledge application, and support for human interaction.

Data Access and Organization. At the data level, agents provide natural-language interfaces to complex healthcare databases, simplifying querying and retrieval. EHRFlow [190] breaks clinician queries into sub-tasks, selects tools, and generates privacy-preserving SQL and FHIR queries with iterative debugging. EHRAgent [156] treats multi-table reasoning as a tool-use planning task supported by a code interpreter. ClinicalAgent [214] lets clinicians ask natural-language questions about clinical trials and retrieves evidence from knowledge graphs, safety reports, enrollment data, and predictive models to assess feasibility and safety. Talk2Biomodels (T2B) [187] extends this idea to systems biology, enabling users to query and simulate biological models through natural language. Together, these systems reduce manual effort and error in retrieval, allowing clinicians to access longitudinal data through simple queries.

Application of Clinical Knowledge. At the knowledge level, agents apply established medical rules and protocols, shifting from retrieval to rule-based execution. CDR-Agent [193] demonstrates this by applying validated decision rules in emergency care to reduce unnecessary imaging. In radiotherapy, LLM agents automate treatment planning and achieve better organ protection than manual methods [201]. Similar methods appear in MRI reconstruction, where planning agents translate expert heuristics into workflows that optimize image quality [17]. In public health, the Decision-Language Model (DLM) [11] encodes clinical and equity priorities into reward functions that guide intervention allocation. Across these domains, agents function as dependable executors of evidence-based procedures rather than independent decision-makers.

Support Human Interaction. At the interface level, agents mediate interactions between clinicians, patients, and software systems. Web-based agents with visual perception can navigate graphical EHR platforms to perform data entry and retrieval automatically [165]. Other systems support communication by translating technical radiology reports into patient-friendly language, with over 81% requiring no correction [165]. Higher-level platforms such as MedicalOS [235] translate natural-language instructions into software commands for retrieving records or managing examinations, reducing navigation effort. In surgical settings, systems like SurgBox [191] act as real-time copilots, coordinating information and assisting intraoperative decisions.

6.2 Empowering Patients: Fostering Engagement and Wellness

Patients are increasingly active in managing their own health, creating a need for personalized and continuous support beyond the clinic. For this group, agentic systems provide scalable and on-demand engagement by using memory and persona conditioning to support care delivery. Direct interaction also introduces specific risks, particularly in areas such as mental health, where patient vulnerability and the consequences of inappropriate interventions require strong safeguards. This discussion is organized by intervention type and associated risk, beginning with therapeutic applications and then turning to supportive systems for patient education and self-management.

6.2.1 Mental Health and Counseling. Mental health applications represent one of the most sensitive domains in healthcare. Agentic

systems in this area span a spectrum, from tools that assist clinicians in assessment to therapeutic platforms that engage directly with patients. The latter requires robust safeguards given patient vulnerability, the possibility of crisis situations, and the serious consequences of inappropriate interventions.

Clinician Support for Assessment and Diagnosis. Agentic systems are being developed as auxiliary tools for clinicians, primarily to structure complex assessment protocols and analyze diverse behavioral signals under professional oversight. Multi-agent frameworks can operationalize standardized psychiatric protocols into reliable computational workflows. A representative example is MAGI [81], which converts the MINI interview into a guided process managed by specialized agents that handle navigation, adaptive questioning, and diagnosis generation. Debate-based architectures extend this idea by using agentic argumentation to clarify ambiguous conditions [84]. Beyond structured interviews, systems have been proposed to extract diagnostic signals from varied sources, including patient narratives for identifying cognitive distortions [65], social media activity for detecting disorders [88], and multimodal signals including facial expressions for recognizing emotional states [117]. Other efforts focus on non-verbal inputs. PsyDraw [221], for instance, analyzes children's House-Tree-Person drawings with multi-agent feature extraction and achieves consistency with professional assessments. Additional systems cluster around knowledge-graph reasoning for interpretable differential diagnoses [132], interactive assessment formats [203], and risk-focused tools such as MentalRAG, which monitors patient data to identify suicidal ideation and notify clinicians [158].

Patient-Facing Therapeutic Interventions. When agentic systems shift from assisting clinicians to delivering therapy directly, the safety requirements become more stringent. Empathy and therapeutic alliance are central in this context. CAMI [204] infers patient states and applies motivational interviewing to guide conversations, and AutoCBT [196] delivers structured cognitive behavioral therapy sessions through a multi-agent framework. Other applications target specific needs: VChatter [218] simulates social interactions for exposure therapy, while PDC30 [28] offers psychoeducational support for dementia caregivers. Safety is addressed through continuous monitoring, as in EmoAgent [230], and through escalation protocols embedded in conversational frameworks [2]. A further line of work explores multi-role counseling agents, exemplified by MIND [24], which distributes therapeutic responsibilities across different LLMs to emulate collaborative counseling teams.

Evaluation and Governance. Given the high stakes of patient-facing systems, evaluation and governance are critical. Emerging taxonomies identify risks specific to AI-assisted psychotherapy, including threats to therapeutic alliance and failures in crisis management [162]. Benchmarks such as ESC-Judge [99] provide systematic assessments of empathy and therapeutic quality. Challenges remain, including disparities in model performance across demographic groups, which raise fairness concerns [232]. At the same time, offline-capable models [76] expand access by enabling support in low-connectivity environments. Ensuring that these systems augment rather than replace professional clinical judgment remains central to their responsible use.

6.2.2 Patient Education and Self-Management. AI agents are increasingly deployed as interactive health coaches that explain diseases, treatments, and prevention strategies to improve health literacy. A defining capability of these systems is personalization. They adapt language, tone, and cultural framing to individual patient profiles. Personas that are consistent with clinical norms and can be adjusted by users have been shown to foster trust and engagement [77, 96, 227]. When combined with empathic responses and autonomy-supportive phrasing, these systems help patients make informed decisions and adhere more consistently to care plans.

Evidence-Grounded and Personalized Education. To provide reliable guidance, patient-facing agents rely on RAG to ground explanations in current knowledge rather than latent model memory. AI-VaxGuide [215] exemplifies this approach by transforming vaccination guidelines into an interactive knowledge base and delivering context-sensitive, multilingual answers. Beyond accuracy, long-term patient engagement requires continuity and adaptation. Memory mechanisms allow agents to recall personal details, prior concerns, and progress, which strengthens trust and encourages disclosure [73]. Planning capabilities enable gradual, adaptive goal setting for chronic condition management, such as incremental lifestyle adjustments or regular mental health check-ins. Together, these strategies illustrate how personalization, evidence-grounding, and longitudinal support converge in patient education.

Connected Self-Management and Care Coordination. Agentic systems increasingly extend their role by linking self-management with clinical workflows and remote monitoring. Wearables and home devices provide real-time inputs, while FHIR-aware multi-agent frameworks [32] allow privacy-preserving integration with EHR systems, enabling agents to log observations and trigger follow-ups with least-privilege access. Patient-reported narratives can also be converted into structured FHIR resources for clinical use [42]. Beyond integration, agents facilitate remote care coordination by automating follow-ups and collecting structured updates that support triage and escalation [97]. These capabilities show how patient education tools evolve into connected platforms that sustain self-management while ensuring timely links to professional care.

6.3 Advancing Medical Science and Education

Researchers and educators form the foundation of the healthcare ecosystem, advancing knowledge and training clinicians. Their work involves challenges such as forming hypotheses from large, heterogeneous biomedical data and building scalable, interactive training environments. AI agents contribute in two ways: as a research accelerator that orchestrates discovery workflows in fields such as genomics and drug development, and as a simulator that models patients or peers for clinical training. This section is organized around these two functions: knowledge creation in biomedical research and knowledge dissemination in medical education.

6.3.1 Biomedical Research and Discovery. Biomedical research depends on systematic hypothesis formation, careful experimental planning, and the integration of diverse data and literature. The scale and complexity of genomic and proteomic data, high-resolution microscopy, and large biomedical text corpora pose specific challenges. Systems based on LLMs are beginning to assist clinicians

and researchers by interpreting these data and linking them to decision-making processes [159].

Hypothesis Generation and Experimental Design. A central role for agents in research is scientific reasoning and hypothesis development. Stella [72] identifies literature gaps and proposes new hypotheses with corresponding experiments. BioScientistAgent [217] applies a similar approach to drug discovery by integrating evidence for drug repurposing and suggesting testable interventions. DrugAgent [106] further extends this paradigm by translating high-level discovery concepts into executable code for in silico experimentation. In wet-lab contexts, CRISPR GPT [139] automates guide RNA design, system selection, and protocol drafting, providing an end-to-end pipeline for gene-editing experiments. Biomni [64] similarly demonstrates autonomous hypothesis generation and experimental design capabilities, interpreting complex multi-modal biomedical data to propose testable biological mechanisms and generate experimentally verifiable protocols.

Knowledge Graphs and Adaptive Graphs. Agents also advance discovery by interpreting genomic and structured data with explicit knowledge grounding. GeneAgent [184] enhances gene-set analysis by cross-checking results with biomedical databases, improving reliability in preclinical scenarios. Knowledge graphs provide a basis for auditable reasoning. Biomni [64] performs KG-aware inference across gene-disease-drug-pathway relations to support transparent research decisions. Related approaches combine retrieval-augmented generation with multi-hop KG reasoning [159] and employ generate-verify-revise loops to improve accuracy in interpreting scientific evidence [164].

Multimodal Modeling and Imaging Foundations. In multimodal research, agents integrate imaging with other biomedical data to uncover novel associations. PRISM2 [152], trained on 700,000 whole-slide images and paired reports, provides a foundation for pathology research by enabling zero-shot classification and biomarker prediction, forming a basis for downstream agent pipelines. Building on such foundations, MESHAgents [219] use multi-agent reasoning over cardiovascular imaging to link image features with risk factors, improving classification and supporting causal analysis and explainability. Biomni [64] interprets complex, multi-modal biomedical datasets and autonomously generates experimentally testable protocols.

6.3.2 Medical Education and Training. AI agents are reshaping medical education from static content delivery toward interactive, practice-based learning. Large language models support personalized curricula and adaptive learning plans [4]. Agentic systems extend these capabilities by simulating interactions from clinical practice. By adopting roles such as virtual patients, expert coaches, or interdisciplinary colleagues, they create dynamic environments for training communication, reasoning, and team coordination.

Clinical Encounters Simulation. A central application is the simulation of clinical encounters for skill development. These range from conversational practice, where one agent acts as a patient and another provides structured feedback, to full diagnostic encounters that require history-taking and reasoning under uncertainty. Representative examples include ChatCoach [63], which refines consultation skills, systems that support training in sensitive tasks such as breaking bad news [189], and adolescent health education delivered through interactive narrative games [163]. Fidelity is often achieved by grounding simulated patients in illness scripts or de-identified EHR data, with feedback aligned to rubrics such as OSCE checklists [60, 211]. Advanced platforms such as AgentClinic [151] extend this approach by allowing learners to interact with multimodal patients, order tests, and make diagnostic decisions, giving instructors a means to identify reasoning gaps.

Collaborative and Procedural Training. Another application focuses on team-based and procedural training. MEDCO [188] creates interdisciplinary environments where agents represent patients, physicians, and radiologists, enabling learners to practice collaboration across roles. SurgBox [191] addresses surgical education by coordinating agents across perioperative phases and acting as a real-time copilot, giving surgeons a controlled environment to rehearse complex procedures.

The growing availability of such systems introduces new challenges for medical educators. The focus is shifting from content delivery to designing high-fidelity simulation scenarios. AI-driven workflows assist in defining objectives, generating patient cases, and creating debriefing plans, which lowers development effort while ensuring alignment with established training standards [9].

6.4 Optimizing Healthcare Administration

Beyond direct clinical care, healthcare depends on an administrative layer that includes hospital managers, regulators, and logistical staff responsible for operational efficiency and regulatory compliance. These stakeholders oversee workflows such as scheduling, billing, and compliance reporting. For this domain, agentic AI, particularly multi-agent systems, acts as an autonomous coordinator and workflow executor, capable of automating end-to-end processes like appointment matching or regulatory adherence. The associated risks are mainly operational and legal rather than clinical. This section examines applications at two levels: internal hospital automation and system-level institutional governance and logistics.

6.4.1 Hospital Operational Automation. Agentic systems are being adopted to improve hospital operations by automating logistical workflows, digital administration, and financial reporting. These applications aim to ease the burden of routine tasks while supporting more effective use of institutional resources.

Hospital Workflow & Operational Efficiency. Agents can also be used to improve internal hospital operations or resource use. ORDiRS-Agent [155] leverages digital twin representations and reasoning segmentation to analyze operating room workflows from video streams. It decomposes high-level queries into subtasks, enabling actionable insights into bottlenecks, staffing usage, or resource occupancy. [136] enables privacy-preserving workflow analysis by converting raw video into de-identified digital twin abstractions before applying event detection pipelines. These methods help hospital management by identifying inefficiencies, optimizing staff allocation, and planning room schedules.

Administrative Task Automation. Some works directly target the automation of backend administrative workflows within healthcare institutions. One framework [48] proposes automating general administrative tasks in healthcare via LLM agents, such as

document generation, scheduling, or internal coordination. MedScrubCrew [146] is a multi-agent framework that automates patient–provider appointment matching and scheduling, optimizing resource usage across provider profiles and patient preferences. These systems lower overhead for hospital staff, reduce delays, and improve service capacity.

Coding & Reporting. A further domain is the use of agents for coding, reporting, and regulatory compliance tasks. "Code Like Humans" [118] is a multi-agent solution for medical coding, facilitating revenue cycle management and billing processes with minimal human intervention. A multi-agent approach for International Classification of Diseases (ICD) coding [94] pushes this further in large-scale, automatic ICD assignment. In oncology settings, [176] demonstrates the feasibility of LLMs for registry submission and reporting tasks under real-world constraints. These agents act as intermediaries between raw clinical data and institutional reporting systems, reducing manual work and error rates.

6.4.2 Institutional Governance and System Logistics. A related direction targets institutional and systemic challenges. Agent-based systems help navigate regulations, monitor compliance, and coordinate inter-institutional logistics. Studies also explore their use in supply chain management, clinical research oversight, and safety standard implementation.

Regulation, Policy, and Safety Oversight. These works target regulators, policy makers, or institutional governance. For example, [43] proposed that the future of LLM-based health apps depends on regulators enforcing safety standards, highlighting the need for governance frameworks to ensure safe deployment in healthcare settings. More broadly, [128] calls for regulatory science innovation to handle the unique challenges posed by generative AI and LLMs in health and medicine, advocating for adaptive policies, regulatory sandboxes, and international harmonization. These works emphasize that agents cannot be purely technical — their deployment must be embedded in legal, ethical, and institutional ecosystems.

Supply Chain, Compliance, and Research Oversight. Some works address larger-scale institutional or cross-institutional administration. A negotiation agent [6] for medical supply chains integrates LLMs and blockchain to coordinate deliveries, maintain resilience, and handle contracts in uncertain settings. [55] builds an agentic system for assessing medical device compliance across different legal jurisdictions, useful for manufacturers interacting with varied regulatory bodies. [101] proposes an automated protocol adherence system, useful in coordinating research compliance and internal governance. TrialGenie [90] empowers automated design of clinical trial protocols using agentic intelligence combined with real-world data, reducing overhead in research administration.

7 Evaluation Framework

Evaluating healthcare agents requires a layered framework that links technical accuracy to clinical impact. A rigorous protocol should assess (1) *task and agentic performance*, such as planning and collaboration within established workflows; (2) *simulation, clinical integration, and governance*, i.e., how agents operate safely in clinical environments and comply with institutional standards; and (3) *LLM-as-a-judge*, scalable assessment of open-ended output, validated through expert-aligned meta-evaluation. Together, these layers connect model performance with real-world clinical effectiveness.

7.1 Task and Agentic Performance Metrics

Evaluating agentic AI in healthcare requires moving from metrics for isolated tasks to a holistic assessment of performance within complex clinical workflows. Traditional *task metrics* remain the foundation of performance evaluation. Metrics such as accuracy/F1 for classification [26], exact match for VQA tasks [25], Dice/IoU for segmentation [67], and (BLEU, ROUGE, BERTScore) for report generation [8] measure whether individual components can reliably solve well-defined problems.

However, recent benchmarks extend this evaluation beyond isolated tasks, reflecting *real-world clinical workflows* that involve multi-step reasoning and decision-making. AgentClinic [151] simulates patient visits where agents manage incomplete information and multilingual cases, reporting metrics such as task completion, dialogue efficiency, and guideline adherence. MedAgentBench [70] provides a virtual EHR environment with hundreds of physician-authored tasks, measuring success across administrative and clinical activities. Other benchmarks, including MedAgentBoard [236], CliBench [110], evaluate multi-step workflows such as EHR automation and diagnosis planning.

Building on these workflow-level evaluations, recent work has introduced *agent-specific metrics* that capture both outcomes and processes. Beyond evaluating "what" output an agent produces (accuracy, safety, efficiency), they also consider "how" the result is obtained (planning, tool use, collaboration). Recent studies consistently adopt the following metrics:

- **Planning Execution Quality** evaluates how consistently an agent executes its plan under constraints such as incomplete information or required tool use, and how well the final outcome aligns with the intended reasoning path [70, 151, 156]. For example, HealthFlow [237] examines self-evolving planning and tracks how success rates improve with step budgets and converge over iterations, providing a framework for assessing planning quality over time.
- **Tool-Use Quality** evaluates an agent's ability to orchestrate external tools, e.g., correct API selection, valid parameters, successful execution [57, 151]. Vision-based systems such as CT-Agent [113] and CPathAgent [167] assess both tool-use success and the grounding of decisions in visual evidence to ensure faithfulness. MedOrch [57] extends this by providing transparent step traces that facilitate auditing its tool usage.
- **Efficiency and Cost** quantify the computational and interaction resources required by an agent. Key indicators include latency, token usage, and the number of interaction turns needed to complete a task. MedAgentsBench [170] standardizes evaluation across performance, cost, and latency for complex clinical questions. Interactive benchmarks such as MEDIQ [95] additionally assess information-seeking efficiency as a subgoal of agent performance.
- **Collaboration** introduces metrics for evaluating cooperation in multi-agent pipelines. MedAgentBoard [236], for example, reports collaboration gain and planning stability when comparing with alternative approaches.

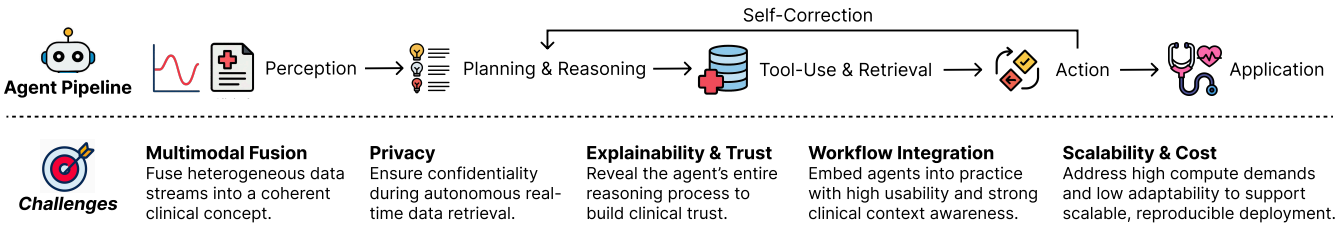


Figure 4: A high-level overview of key challenges facing AI agents in healthcare.

7.2 Simulation, Clinical Integration, and Governance Evaluation

Evaluating AI agents in healthcare goes beyond task accuracy to assess integration into clinical workflows and alignment with institutional rules. This section reviews how evaluation evolves from simulations to clinical studies and governance frameworks.

Simulation-based evaluations test whether agentic systems can operate safely and effectively in multi-actor settings. Agent Hospital [91] simulates a hospital with patient, nurse, and doctor agents, measuring sequential decision success and unsafe behaviors such as constraint violations across full care cycles. The MATEC pilot [29] similarly integrates a team of specialized agents into sepsis workflows and reports clinician ratings on usefulness and accuracy, capturing user feedback at early development stages.

Clinically integrated studies extend evaluation by using medical standards and expert agreement as proxies for decision quality. AMEGA [36] tests whether LLM-generated diagnoses and treatment plans follow established clinical guidelines across specialties. PRISM [54], a clinical trial-matching system, evaluates agents on real-world EHR data, measuring how accurately they identify eligible patients and align with expert judgments.

Safety and governance-oriented evaluations further incorporate reliability, privacy, and accountability into assessment frameworks. A HIPAA-aware agentic framework [123] defines privacy compliance and policy enforcement as measurable dimensions, ensuring governance is represented alongside technical metrics. Recent reviews [22] emphasize the need for integrated evaluation frameworks that connect model-level metrics with human and system outcomes, synthesizing dimensions such as guideline adherence, workflow efficiency, safety, and usability.

7.3 LLM-as-a-Judge

As clinical tasks become more open-ended, evaluating the quality of generated text increasingly relies on LLMs themselves as evaluators. This shift, often called *LLM-as-a-judge*, aims to assess both semantic accuracy and clinical evidence grounding. Early systems used structured, clinically informed metrics such as RadGraph F1 and RadCliQ [68, 210], which align better with radiologist judgments. Recent work introduces LLM-enhanced evaluators that interpret clinical text directly. GREEN [129] uses an LLM to detect and explain clinically significant errors, providing category-level counts that correlate with expert reviews. GEMA-Score [222] combines structured extraction (NER-based F1) with an LLM scoring module for completeness and readability.

Meta-evaluation of LLM judges. The reliability of LLM judges is a growing concern. Surveys [51, 89] identify issues such as prompt sensitivity, bias, and inconsistent scoring, and recommend reporting expert agreement (Spearman or Kendall correlation), prompt robustness, and the use of supporting evidence. To validate, resources such as RadEvalX [14] and ReXrank [13] provide expert annotations and leaderboards to benchmark how well automated metrics, including LLM-based ones, align with human evaluation. Later studies [169] suggest grounding judgments in predefined rubrics, using multiple judges for consensus, and including human audits to reduce inconsistency. In practice, we recommend future papers employing LLM judges should report (1) agreement with clinicians, (2) judging prompts and rubrics, (3) evidence-conditioning protocols, and (4) judging cost relative to human review. These practices improve transparency and reproducibility across agent evaluation.

Ultimately, integrating human-grounded meta-evaluation with automated LLM judging creates a hybrid paradigm that balances scalability and trustworthiness, ensuring that progress in AI agents reflect clinical improvement rather than metric inflation [157]. While promising, large-scale evaluation pipelines still face computational and accessibility constraints. Future work on lighter and open models could make clinical evaluation more practical.

8 Future Challenges and Opportunities

The challenges of agentic AI systems stem from their operation across the entire clinical pipeline. Figure 4 shown a typical healthcare agent system that process multimodal data, performs planning and reasoning, accesses external tools and information, and executes actions within clinical workflows. Each stage introduces specific points of failure, including multimodal fusion, real-time data governance, decision transparency, workflow integration, and scalability. The following sections examine these challenges and identify opportunities for improving the reliability and deployability of AI agents in healthcare.

8.1 Reliable Fusion Across Modalities

Effective multimodal fusion is essential to clinical reasoning and decision-making. In sequential decision systems, fusion errors are not simple misclassifications—they can propagate into a chain of faulty downstream actions, from ordering contraindicated tests to missing critical interventions. The challenge lies not only in integrating heterogeneous data streams but also in bridging the semantic gap between raw signals and clinically meaningful concepts [135, 228]. Moreover, these systems must remain reliable

under incomplete information, which requires more than data imputation [143]. They must reason under uncertainty and proactively plan missing data acquisition, as a clinician would during diagnosis.

8.2 Clinical Workflow Integration

Integrating advanced AI into clinical practice requires moving from decision support to active collaboration. Systems that act and interact must maintain strong contextual awareness to ensure their behavior is safe and relevant, such as recognizing the urgency of an ICU compared with the pace of an outpatient clinic [200]. Achieving this balance requires careful human-AI interaction design rooted in *mixed-initiative interaction* [61]. The system should dynamically negotiate control with users, taking initiative when appropriate and deferring when human judgment is essential [79, 116].

The collaborative potential of agentic systems depends on interoperability. Their ability to perceive context and execute tasks relies on stable communication with diverse hospital IT systems; when integration is fragile, deployment fails. Adoption also depends on usability: clinicians should be able to operate these systems without programming skills or complex setup [56, 59]. Interfaces should support low-friction workflows, role-appropriate abstractions, and safe defaults that reduce cognitive and operational load.

8.3 Architectural Safeguards for Data Privacy

The dynamic and interactive nature of advanced systems introduces new challenges for data privacy. Unlike models trained on static datasets, systems that autonomously query patient data at runtime create a continuous need for governance. The focus thus shifts from securing stored data to regulating real-time data transactions. In addition, data acquisition in healthcare is slow and costly due to consent processes and regulatory approvals, limiting the scale and diversity of training and evaluation datasets [119]. Consequently, privacy-preserving machine learning must be complemented by architectural safeguards that regulate system behavior [166]. Principles such as least-privilege access control [148], immutable audit trails for each data query, and real-time information anonymization are essential to ensure that autonomy does not compromise patient confidentiality [123, 126].

8.4 Explainability for Trustworthy AI

To earn clinical trust, explanations must move beyond predictions to reveal the reasoning process behind each recommendation. When systems initiate actions, clinicians need transparency into their goals, alternative strategies, and the rationale for their final decisions [19, 75]. Such transparency is essential for safely delegating clinical tasks. Trust also depends on a system's ability to communicate uncertainty. A reliable collaborator should recognize its knowledge limits and express confidence not only in its findings but also in the expected outcomes of its recommendations. This "explainable uncertainty" helps calibrate clinicians' reliance on the system and supports safe, effective human-AI collaboration [35].

8.5 Adaptability and Reproducibility for Scalability

The substantial computational cost of developing advanced systems poses a scalability challenge closely tied to generalization. High upfront investments are difficult to justify if systems remain brittle and fail to adapt to new clinical environments, a common issue known as domain shift. A promising approach to improving scalability is to move beyond this static design, requiring architectures that are both powerful and adaptable. For example, self-evolving systems can discover new tools, integrate emerging knowledge, and refine their strategies over time [20].

Scalability also depends on reproducibility. When code, models, or datasets are unavailable, proprietary artifacts and opaque configurations prevent replication and hinder cross-site benchmarking. Together, adaptability and reproducibility determine whether these technologies can scale efficiently and equitably across diverse healthcare settings [72, 223].

9 Conclusion

This survey has reviewed the current landscape of AI agents in healthcare through an analysis of more than 200 recent studies. We proposed a taxonomy spanning perception of clinical modalities, agent capabilities and architectures, application domains, and evaluation approaches. The review shows that clinical requirements for safety, reliability, and accountability are shaping agent design in ways that diverge from general-purpose systems. Despite this progress, open challenges remain. Technically, robust multimodal integration, seamless workflow compatibility with clinical IT, and computational scalability are open problems. At the governance level, safeguarding data privacy and achieving process-level explainability are necessary to build trust in clinical settings. Addressing these challenges is essential for reliable deployment of agentic systems in practice. The trajectory of research points toward enhancing human–AI collaboration, with the objective of developing systems that are autonomous, verifiable, and safely integrated into patient care. This survey provides a structured foundation to support further research and guide responsible adoption.

References

[1] Mahyar Abbasian, Iman Azimi, Mohammad Feli, Amir M Rahmani, and Ramesh Jain. 2024. Empathy through multimodality in conversational interfaces. *arXiv preprint arXiv:2405.04777* (2024).

[2] Mahyar Abbasian, Iman Azimi, Amir M Rahmani, and Ramesh Jain. 2025. Conversational health agents: a personalized large language model-powered agent framework. *JAMIA Open* 8, 4 (2025), ooaf067.

[3] Mahyar Abbasian, Zhongqi Yang, Elahe Khatibi, Pengfei Zhang, Nitish Nagesh, Iman Azimi, Ramesh Jain, and Amir M Rahmani. 2024. Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1–4.

[4] Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. 2023. Large language models in medical education: opportunities, challenges, and future directions. *JMIR medical education* 9, 1 (2023), e48291.

[5] Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic arabic medical dialogues using advanced multi-agent llm techniques. In *Proceedings of The Second Arabic Natural Language Processing Conference*. 11–26.

[6] Mariam ALMutairi and Hyungmin Kim. 2025. Resilient Multi-Agent Negotiation for Medical Supply Chains: Integrating LLMs and Blockchain for Transparent Coordination. *arXiv preprint arXiv:2507.17134* (2025).

[7] Umair Ayub, Syed Arsalan Ahmed Naqvi, Salman Ayub Jajja, Muhammad Umar Afzal, Ji-Eun Irene Yum, Kaneez Zahra Rubab Khakwani, Chitta Baral, Sumanta Kumar Pal, Neeraj Agarwal, Abhishek Tripathi, et al. 2025. A large language model (LLM)-based multi-agent framework for risk stratification and treatment recommendations in localized prostate cancer (locPCa).

[8] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid,

Anja Thieme, et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15016–15027.

[9] Francesco L. Barra, Giulia Rodella, Andrea Costa, Andrea Scalogna, Lorenzo Carenzo, Alessandro Monzani, and Francesco Della Corte. 2025. From prompt to platform: an agentic AI workflow for healthcare simulation scenario design. *Advances in Simulation (London)* 10, 1 (2025), 29. doi:10.1186/s41077-025-00357-z

[10] Suhana Bedi, Iddah Mlauzi, Daniel Shin, Sanmi Koyejo, and Nigam H. Shah. 2025. The Optimization Paradox in Clinical AI Multi-Agent Systems. (2025). arXiv:2506.06574 [cs.AI] doi:10.48550/arXiv.2506.06574

[11] Nikhil Behari, Edwin Zhang, Yunfan Zhao, Aparna Taneja, Dheeraj Nagaraj, and Milind Tambe. 2024. A decision-language model (dlm) for dynamic restless multi-armed bandit tasks in public health. *Advances in Neural Information Processing Systems* 37 (2024), 3964–4002.

[12] Christian Bluethgen, Dave Van Veen, Daniel Truhn, Jakob Nikolas Kather, Michael Moor, Malgorzata Polacin, Akshay Chaudhari, Thomas Frauenfelder, Curtis P Langlotz, Michael Krauthammer, et al. 2025. Agentic Systems in Radiology: Design, Applications, Evaluation, and Challenges. *arXiv preprint arXiv:2510.09404* (2025).

[13] A. R. Calamida et al. 2024. ReXrank: A Public Leaderboard for AI-Powered Radiology Report Generation. *arXiv preprint arXiv:2411.15122* (2024). https://arxiv.org/abs/2411.15122

[14] A. R. Calamida and collaborators. 2024. RadEvalX: Radiology Report Generation Models Evaluation Dataset. PhysioNet. https://physionet.org/content/rad-eval-x/

[15] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C Nascimento. 2023. Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–20.

[16] Rihao Chang, He Jiao, Weizhi Nie, Honglin Guo, Keliang Xie, Zhenhua Wu, Lina Zhao, Yunpeng Bai, Yongtao Ma, Lanjun Wang, et al. 2025. Organ-Agents: Virtual Human Physiology Simulator via LLMs. *arXiv preprint arXiv:2508.14357* (2025).

[17] Yuchou Chang, Zhiqiang Li, Huy Anh Pham, and Gulfam Ahmed Saju. 2024. Intelligent Agent Planning for Optimizing Parallel MRI Reconstruction via A Large Language Model. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 1–4. doi:10.1109/EMBC53108.2024.10782629

[18] Chengkuan Chen, Luca L. Weishaupt, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Guillaume Jaume, Ming Y. Lu, and Faisal Mahmood. 2025. Evidence-Based Diagnostic Reasoning with Multi-Agent Copilot for Human Pathology. (2025). arXiv:2506.20964 [cs.CV] doi:10.48550/arXiv.2506.20964

[19] Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301* (2024).

[20] Kai Chen, Xinfeng Li, Tianpei Yang, Hewei Wang, Wei Dong, and Yang Gao. 2025. MDTeamGPT: A Self-Evolving LLM-Based Multi-Agent Framework for Multi-Disciplinary Team Medical Consultation. (2025). arXiv:2503.13856 [cs.AI] doi:10.48550/arXiv.2503.13856

[21] Kai Chen, Taihang Zhen, Hewei Wang, Kailai Liu, Xinfeng Li, Jing Huo, Tianpei Yang, Jinfeng Xu, Wei Dong, and Yang Gao. 2025. MedSentry: Understanding and Mitigating Safety Risks in Medical LLM Multi-Agent Systems. *arXiv preprint arXiv:2505.20824* (2025).

[22] Xiaolan Chen et al. 2025. Evaluating large language models and agents in healthcare. *iMED* (2025). doi:10.1016/j.imed.2025.03.002

[23] Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine* 8, 1 (2025), 159.

[24] Yujia Chen, Changsong Li, Yiming Wang, Tianjie Ju, Qingqing Xiao, Nan Zhang, Zifan Kong, Peng Wang, and Binyu Yan. 2025. MIND: Towards Immersive Psychological Healing with Multi-Agent Inner Dialogue. (2025). arXiv:2502.19860 [cs.CL] doi:10.48550/arXiv.2502.19860

[25] Yixiong Chen, Wenjie Xiao, Pedro RAS Bassi, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. 2025. Are Vision Language Models Ready for Clinical Diagnosis? A 3D Medical Benchmark for Tumor-centric Visual Question Answering. *arXiv preprint arXiv:2505.18915* (2025).

[26] Yixiong Chen, Chunhui Zhang, Li Liu, Cheng Feng, Changfeng Dong, Yongfang Luo, and Xiang Wan. 2021. USCL: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 627–637.

[27] Zhen Chen, Zhihao Peng, Xusheng Liang, Cheng Wang, Peigan Liang, Linsheng Zeng, Minjie Ju, and Yixuan Yuan. 2025. MAP: Evaluation and Multi-Agent Enhancement of Large Language Models for Inpatient Pathways. (2025). arXiv:2503.13205 [cs.AI] doi:10.48550/arXiv.2503.13205

[28] Sheung-Tak Cheng and Peter HF Ng. 2025. The PDC30 Chatbot—Development of a Psychoeducational Resource on Dementia Caregiving Among Family Caregivers: Mixed Methods Acceptability Study. *JMIR aging* 8 (2025), e63715.

[29] Andrew Cho, Jonathan S. H. Woo, et al. 2025. The Application of MATEC (Multi-AI Agent Team Care) Framework in Sepsis Care. *arXiv preprint arXiv:2503.16433* (2025). https://arxiv.org/abs/2503.16433

[30] Jihye Choi, Nils Palumbo, Prasad Chalasani, Matthew M. Engelhard, Somesh Jha, Anivarya Kumar, and David Page. 2024. MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance. arXiv:2408.01869 [cs.CL] https://arxiv.org/abs/2408.01869

[31] Mohita Chowdhury, Yajie Vera He, Jared Joselowitz, Aisling Higham, and Ernest Lim. 2025. ASTRID–An Automated and Scalable TRIaD for the Evaluation of RAG-based Clinical Question Answering Systems. *arXiv preprint arXiv:2501.08208* (2025).

[32] Carmen De Maio, Giuseppe Fenza, Domenico Furno, Teodoro Grauso, and Vincenzo Loia. 2024. A multi-agent architecture for privacy-preserving natural language interaction with fhir-based electronic health records. In *2024 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 1–6.

[33] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130 [cs.CL] https://arxiv.org/abs/2404.16130

[34] Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673* (2025).

[35] Xiuyi Fan. 2025. Position Paper: Integrating Explainability and Uncertainty Estimation in Medical AI. *arXiv preprint arXiv:2509.18132* (2025).

[36] Dennis Fast, Lisa C. Adams, Felix Busch, Conor Fallon, et al. 2024. Autonomous medical evaluation for guideline adherence of large language models. *npj Digital Medicine* (2024). doi:10.1038/s41746-024-01356-6

[37] Nima Fathi, Amar Kumar, and Tal Arbel. 2025. AURA: A Multi-modal Medical Agent for Understanding, Reasoning and Annotation. In *International Workshop on Agentic AI for Medicine*. Springer, 105–114.

[38] Mohammad Feli, Iman Azimi, Pasi Liljeberg, and Amir M Rahmani. 2025. An LLM-Powered Agent for Physiological Data Analysis: A Case Study on PPG-based Heart Rate Estimation. *arXiv preprint arXiv:2502.12836* (2025).

[39] Jinghao Feng, Qiaoyu Zheng, Chaoyi Wu, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. M^3Builder: A Multi-Agent System for Automated Machine Learning in Medical Imaging. (2025). arXiv:2502.20301 [cs.CV] doi:10.48550/arXiv.2502.20301

[40] Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. 2025. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. *arXiv preprint arXiv:2505.19630* (2025).

[41] Dyke Ferber, Omar SM El Nahhas, Georg Wölflein, Isabella C Wiest, Jan Clusmann, Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, et al. 2025. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature cancer* (2025), 1–13.

[42] Johann Frei, Nils Feldhus, Lisa Raithel, Roland Roller, Alexander Meyer, and Frank Kramer. 2025. Infherno: End-to-end Agent-based FHIR Resource Synthesis from Free-form Clinical Notes. *arXiv preprint arXiv:2507.12261* (2025).

[43] Oscar Freyer, Isabella Catharina Wiest, Jakob Nikolas Kather, and Stephen Gilbert. 2024. A future role for health applications of large language models depends on regulators enforcing safety standards. *The Lancet Digital Health* 6, 9 (2024), e662–e672.

[44] Agasthya Gangavarapu and Ananya Gangavarapu. 2024. IMAS: A Comprehensive Agentic Approach to Rural Healthcare Delivery. arXiv:2410.12868 [cs.AI] https://arxiv.org/abs/2410.12868

[45] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with AI agents. *Cell* 187, 22 (2024), 6125–6151.

[46] Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025. TxAgent: An AI agent for therapeutic reasoning across a universe of tools. *arXiv preprint arXiv:2503.10970* (2025).

[47] Roberta Gazzarata, Joao Almeida, Lars Lindsköld, Giorgio Cangioli, Eugenio Gaeta, Giuseppe Fico, and Catherine E Chronaki. 2024. HL7 Fast Healthcare Interoperability Resources (HL7 FHIR) in digital healthcare ecosystems for chronic disease management: Scoping review. *International journal of medical informatics* 189 (2024), 105507.

[48] Senay A Gebreab, Khaled Salah, Raja Jayaraman, Muhammad Habib ur Rehman, and Samer Ellaham. 2024. Llm-based framework for administrative task automation in healthcare. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, 1–7.

[49] Alon Gorenshtein, Moran Sorka, Mohamed Khateb, Dvir Aran, and Shahar Shelly. 2025. Agent-guided AI-powered interpretation and reporting of nerve conduction studies and EMG (INSPIRE). *Clinical Neurophysiology* (2025), 2110792.

[50] Lawrence O Gostin, Laura A Levit, and Sharyl J Nass. 2009. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. (2009).

[51] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594* (2024).

[52] Xiongbin Gui, Hanlin Lv, Xiao Wang, Longting Lv, Yi Xiao, and Lei Wang. 2025. Enhancing hepatopathy clinical trial efficiency: a secure, large language model-powered pre-screening pipeline. *BioData Mining* 18, 1 (2025), 42.

[53] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[54] Shashi Kant Gupta, Aditya Basu, Mauro Nievas, Jerrin Thomas, et al. 2024. PRISM: Patient Records Interpretation for Semantic Clinical Trial Matching using Large Language Models. *npj Digital Medicine* (2024). doi:10.1038/s41746-024-01274-7

[55] Yu Han, Aaron Ceross, and Jeroen HM Bergmann. 2025. Standard Applicability Judgment and Cross-jurisdictional Reasoning: A RAG-based Framework for Medical Device Compliance. *arXiv preprint arXiv:2506.18511* (2025).

[56] Masooma Hassan, Andre Kushniruk, and Elizabeth Borycki. 2024. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. *JMIR Human Factors* 11 (2024), e48633.

[57] Yexiao He, Ang Li, Boyi Liu, Zhewei Yao, and Yuxiong He. 2025. MedOrch: Medical Diagnosis with Tool-Augmented Reasoning Agents for Flexible Extensibility. *arXiv preprint arXiv:2506.00235* (2025). https://arxiv.org/abs/2506.00235

[58] A Ali Heydari, Ken Gu, Vidya Srinivas, Hong Yu, Zhihan Zhang, Yuwei Zhang, Akshay Paruchuri, Qian He, Hamid Palangi, Nova Hammerquist, et al. 2025. The Anatomy of a Personal Health Agent. *arXiv preprint arXiv:2508.20148* (2025).

[59] Vasco Gerardo Hinostroza Fuentes, Hezerul Abdul Karim, Myles Joshua Toledo Tan, and Nouar AlDahoul. 2025. AI with agency: a vision for adaptive, efficient, and ethical healthcare. *Frontiers in Digital Health* 7 (2025), 1600216.

[60] Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, Moritz Mahling, et al. 2024. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Medical Education* 10, 1 (2024), e59213.

[61] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[62] Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. 2025. Can A Society of Generative Agents Simulate Human Behavior and Inform Public Health Policy? A Case Study on Vaccine Hesitancy. *arXiv preprint arXiv:2503.09639* (2025).

[63] Hengguan Huang, Songtao Wang, Hongfu Liu, Hao Wang, and Ye Wang. 2024. Benchmarking large language models on communicative medical coaching: a dataset and a novel system. In *Findings of the Association for Computational Linguistics ACL 2024*. 1624–1637.

[64] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv* (2025).

[65] Shuya Huang, Peng Xu, Richard Fok, Sumbul Ghosh, and Muhao Chen. 2024. ERD: A Framework for Improving LLM Reasoning for Cognitive Distortion Classification. *arXiv preprint arXiv:2403.14255* (2024). http://arxiv.org/abs/2403.14255v1

[66] Alexis Huet, Zied Ben Houidi, and Dario Rossi. 2025. Episodic memories generation and evaluation benchmark for large language models. *arXiv preprint arXiv:2501.13121* (2025).

[67] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* 18, 2 (2021), 203–211.

[68] Shubham Jain et al. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *arXiv preprint arXiv:2106.14463* (2021). https://arxiv.org/abs/2106.14463

[69] Cong Jiang and Xiaolei Yang. 2024. Agents on the Bench: Large Language Model Based Multi-Agent Framework for Trustworthy Digital Justice. (2024). arXiv:2412.18697 [cs.AI] doi:10.48550/arXiv.2412.18697

[70] Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. MedAgentBench: A Realistic Virtual EHR Environment to Benchmark Medical LLM Agents. *NEJM AI* (2025). https://ai.nejm.org/doi/full/10.1056/AIdbp2500144

[71] Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, et al. 2024. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *arXiv preprint arXiv:2402.13225* (2024).

[72] Ruofan Jin, Zaixi Zhang, Mengdi Wang, and Le Cong. 2025. STELLA: Self-Evolving LLM Agent for Biomedical Research. *arXiv preprint arXiv:2507.02004* (2025).

[73] Eunkyung Jo, Yuin Jeong, SoHyun Park, Daniel A Epstein, and Young-Ho Kim. 2024. Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[74] Kurmanbek Kaiyrbekov, Nicholas J Dobbins, and Sean D Mooney. 2025. Automated Survey Collection with LLM-based Conversational Agents. *arXiv preprint arXiv:2504.02891* (2025).

[75] 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. *Journal of Medical Internet Research* 26 (2024), e59439.

[76] Sungjun Kim, Hyungjoo Lee, Jungsun Park, and Edward Choi. 2025. An Offline Mobile Conversational Agent for Mental Health Support: Learning from Emotional Dialogues and Psychological Texts with Student-Centered Evaluation. *arXiv preprint arXiv:2507.10580* (2025). http://arxiv.org/abs/2507.10580v1

[77] Yubin Kim, Zhiyuan Hu, Hyewon Jeong, Eugene Park, Shuyue Stella Li, Chanwoo Park, Shiyun Xiong, MingYu Lu, Hyeonhoon Lee, Xin Liu, et al. 2025. BehaviorSFT: Behavioral Token Conditioning for Clinical Agents Across the Proactivity Spectrum. *arXiv preprint arXiv:2505.21757* (2025).

[78] Yubin Kim, Hyewon Jeong, Chanwoo Park, Eugene Park, Haipeng Zhang, Xin Liu, Hyeonhoon Lee, Daniel McDuff, Marzyeh Ghassemi, Cynthia Breazeal, Samir Tulebaev, and Hae Won Park. 2025. Tiered Agentic Oversight: A Hierarchical Multi-Agent System for Healthcare Safety. (2025). arXiv:2506.12482 [cs.AI] doi:10.48550/arXiv.2506.12482

[79] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems* 37 (2024), 79410–79452.

[80] Ngoc Bui Lam Quang, Nam Le Nguyen Binh, Thanh-Huy Nguyen, Le Thien Phuc Nguyen, Quan Nguyen, and Ulas Bagci. 2025. GMAT: Grounded Multi-Agent Clinical Description Generation for Text Encoder in Vision-Language MIL for Whole Slide Image Classification. (2025). arXiv:2508.01293 [cs.CV] doi:10.48550/arXiv.2508.01293

[81] Jiaming Lee, Yunfei Chen, Kwang Lee, et al. 2025. MAGI: Multi-Agent Guided Interview for Psychiatric Assessment. *Findings of the Association for Computational Linguistics: ACL* (2025). https://aclanthology.org/2025.findings-acl.1278/

[82] Namkyeong Lee, Edward De Brouwer, Ehsan Hajiramezanali, Tommaso Biancalani, Chanyoung Park, and Gabriele Scalia. 2025. RAG-Enhanced Collaborative LLM Agents for Drug Discovery. arXiv:2502.17506 [cs.LG] https://arxiv.org/abs/2502.17506

[83] Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, et al. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. *arXiv preprint arXiv:2407.03103* (2024).

[84] Yeonji Lee, Sangjun Park, Kyunghyun Cho, and JinYeong Bak. 2024. MentalAgora: A Gateway to Advanced Personalized Care in Mental Health through Multi-Agent Debating and Attribute Control. *arXiv preprint arXiv:2407.02736* (2024).

[85] Yeawon Lee, Xiaoyang Wang, and Christopher C Yang. 2025. Automated Clinical Problem Detection from SOAP Notes using a Collaborative Multi-Agent LLM Architecture. *arXiv preprint arXiv:2508.21803* (2025).

[86] Brenna Li, Amy Wang, Patricia Strachan, Julie Anne Séguin, Sami Lachgar, Karyn C Schroeder, Mathias S Fleck, Renee Wong, Alan Karthikesalingam, Vivek Natarajan, et al. 2024. Conversational AI in health: Design considerations from a Wizard-of-Oz dermatology case study with users, clinicians and a medical LLM. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–10.

[87] Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. 2024. Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483* (2024).

[88] Guodong Li, Zixuan Meng, Fulei Yuan, Xiaobo Peng, Angyan Li, Yang Xiang, and Qi Zheng. 2024. Detecting mental disorder on social media: a ChatGPT-augmented explainable approach. *arXiv preprint arXiv:2401.17477* (2024). http://arxiv.org/abs/2401.17477v2

[89] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *arXiv preprint arXiv:2412.05579* (2024).

[90] Haoyang Li, Weishen Pan, Suraj Rajendran, Chengxi Zang, and Fei Wang. 2025. TrialGenie: Empowering Clinical Trial Design with Agentic Intelligence and Real World Data. *medRxiv* (2025), 2025–04.

[91] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, et al. 2024. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint arXiv:2405.02957* (2024). https://arxiv.org/abs/2405.02957

[92] Ming Li and Keliang Zhang. 2025. A multi-agent system based on HNC for domain-specific machine translation. *Scientific Reports* 15, 1 (2025), 20820.

[93] Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. 2025. CARE-AD: a multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine* 8, 1 (2025), 541.

[94] Rumeng Li, Xun Wang, and Hong Yu. 2024. Exploring llm multi-agents for icd coding. *arXiv preprint arXiv:2406.15363* (2024).

[95] Shiyang Li et al. 2024. MEDIQ: Question-Asking LLMs and a Benchmark for Medical Information-Seeking. In *NeurIPS 2024.* https://proceedings.neurips.cc/paper_files/paper/2024/file/32b80425554e081204e5988ab1c97e9a-Paper-Conference.pdf

[96] Wenwen Li, Kangwei Shi, and Yidong Chai. 2025. AI chatbots as professional service agents: developing a professional identity. *arXiv preprint arXiv:2501.14179* (2025).

[97] Xueshen Li, Xinlong Hou, Nirupama Ravi, Ziyi Huang, and Yu Gan. 2025. A two-stage proactive dialogue generator for efficient clinical information collection using large language model. *Expert Systems with Applications* 287 (2025), 127833.

[98] Xueyang Li, Mingze Jiang, Gelei Xu, Jun Xia, Mengzhao Jia, Danny Chen, and Yiyu Shi. 2025. AT-CXR: Uncertainty-Aware Agentic Triage for Chest X-rays. arXiv:2508.19322 [eess.IV] https://arxiv.org/abs/2508.19322

[99] Zhichao Li, Yue Zhou, Jiahao Liu, Shuai Wang, and Chong Chen. 2025. ESC-Judge: A Framework for Comparing Emotional Support Conversational Agents. *arXiv preprint arXiv:2505.12531* (2025). http://arxiv.org/abs/2505.12531v1

[100] Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. 2025. ReflecTool: Towards Reflection-Aware Tool-Augmented Clinical Agents. arXiv:2410.17657 [cs.CL] https://arxiv.org/abs/2410.17657

[101] Gregor Lichtner, Claudia Spies, Carlo Jurth, Thomas Bienert, Anika Mueller, Oliver Kumpf, Vanessa Piechotta, Nicole Skoetz, Monika Nothacker, Martin Boeker, et al. 2023. Automated monitoring of adherence to evidenced-based clinical guideline recommendations: design and implementation study. *Journal of Medical Internet Research* 25 (2023), e41177.

[102] Fengze Liu, Haoyu Wang, Joonhyuk Cho, Dan Roth, and Andrew W Lo. 2025. AUTOCT: Automating Interpretable Clinical Trial Prediction with LLM Agents. *arXiv preprint arXiv:2506.04293* (2025).

[103] Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. 2023. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine* 6, 1 (2023), 226.

[104] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[105] Kunyao Liu, Yiming Yang, Binbin Liu, Pengfei Li, and Bing Liu. 2024. Depression Diagnosis Dialogue Simulation: Self-improving Psychiatrist with Tertiary Memory. *arXiv preprint arXiv:2409.15084* (2024). http://arxiv.org/abs/2409.15084v2

[106] Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. 2024. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *arXiv preprint arXiv:2411.15692* (2024).

[107] Antoine Lizée, Pierre-Auguste Beaucoté, James Whitbeck, Marion Doumeingts, Anaël Beaugnon, and Isabelle Feldhaus. 2025. Conversational Medical AI: Ready for Practice. arXiv:2411.12808 [cs.AI] https://arxiv.org/abs/2411.12808

[108] Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. 2024. ClinicalRAG: enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024).* 64–68.

[109] Zhiyao Luo and Tingting Zhu. 2025. Are Large Language Models Dynamic Treatment Planners? An In Silico Study from a Prior Knowledge Injection Angle. *arXiv preprint arXiv:2508.04755* (2025).

[110] Mingde Ma et al. 2024. CliBench: A Multifaceted and Multigranular Evaluation of Clinical Diagnosis with LLMs. *arXiv preprint arXiv:2406.09923* (2024). https://arxiv.org/abs/2406.09923

[111] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2024. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, Vol. 2023. 1105.

[112] Neil Mallinar, A Ali Heydari, Xin Liu, Anthony Z Faranesh, Brent Winslow, Nova Hammerquist, Benjamin Graef, Cathy Speed, Mark Malhotra, Shwetak Patel, et al. 2025. A Scalable Framework for Evaluating Health Language Models. *arXiv preprint arXiv:2503.23339* (2025).

[113] Yuren Mao, Wenyi Xu, Yuyang Qin, and Yunjun Gao. 2025. CT-Agent: A Multimodal-LLM Agent for 3D CT Radiology Question Answering. *arXiv preprint arXiv:2505.16229* (2025). https://arxiv.org/abs/2505.16229

[114] Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584* (2024).

[115] Nicholas Matsumoto, Hyunjun Choi, Jay Moran, Miguel E Hernandez, Mythreye Venkatesan, Xi Li, Jui-Hsuan Chang, Paul Wang, and Jason H Moore. 2025. ESCARGOT: an AI agent leveraging large language models, dynamic graph of thoughts, and biomedical knowledge graphs for enhanced reasoning. *Bioinformatics* 41, 2 (2025), btaf031.

[116] Bertalan Mesko. 2023. The ChatGPT (generative artificial intelligence) revolution has made artificial intelligence approachable for medical professionals. *Journal of medical Internet research* 25 (2023), e48392.

[117] Joaquin Molto, Jonathan Fields, Ubbo Visser, and Christine Lisetti. 2024. An LLM-powered Socially Interactive Agent with Adaptive Facial Expressions for Conversing about Health. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*. 75–77.

[118] Andreas Motzfeldt, Joakim Edin, Casper L Christensen, Christian Hardmeier, Lars Maaløe, and Anna Rogers. 2025. Code Like Humans: A Multi-Agent Solution for Medical Coding. *arXiv preprint arXiv:2509.05378* (2025).

[119] Lama Moukheiber, Mira Moukheiber, Dana Moukheiiber, Jae-Woo Ju, and Hyung-Chul Lee. 2025. EchoQA: A Large Collection of Instruction Tuning Data for Echocardiogram Reports. arXiv:2503.02365 [cs.AI] https://arxiv.org/abs/2503.02365

[120] Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, et al. 2024. Polaris: A safety-focused llm constellation architecture for healthcare. *arXiv preprint arXiv:2403.13313* (2024).

[121] Awais Naeem, Tianhao Li, Huang-Ru Liao, Jiawei Xu, Aby M Mathew, Zehao Zhu, Zhen Tan, Ajay Kumar Jaiswal, Raffi A Salibian, Ziniu Hu, et al. 2024. Path-RAG: Knowledge-Guided Key Region Retrieval for Open-ended Pathology Visual Question Answering. *arXiv preprint arXiv:2411.17073* (2024).

[122] Varun Nair, Elliot Schumacher, Geoffrey Tso, and Anitha Kannan. 2023. DERA: enhancing large language model completions with dialog-enabled resolving agents. *arXiv preprint arXiv:2303.17071* (2023).

[123] Subash Neupane et al. 2025. Towards a HIPAA Compliant Agentic AI System in Healthcare. *arXiv preprint arXiv:2504.17669* (2025). https://arxiv.org/abs/2504.17669

[124] Alison Nightingale. 2009. A guide to systematic literature reviews. *Surgery (Oxford)* 27, 9 (2009), 381–384.

[125] Diogo AP Nunes, Dan Furrer, Sara Berger, Guillermo Cecchi, Joana Ferreira-Gomes, Fani Neto, David Martins de Matos, A Vania Apkarian, and Paulo Branco. 2025. Advancing the prediction and understanding of placebo responses in chronic back pain using large language models. *medRxiv* (2025).

[126] Humza Nusrat, Bing Luo, Ryan Hall, Joshua Kim, Hassan Bagher-Ebadian, Anthony Doemer, Benjamin Movsas, and Kundan Thind. 2025. Autonomous Radiotherapy Treatment Planning Using DOLA: A Privacy-Preserving, LLM-Based Optimization Agent. *arXiv preprint arXiv:2503.17553* (2025).

[127] Jasmine Chiat Ling Ong, Liyuan Jin, Kabilan Elangovan, Gilbert Yong San Lim, Daniel Yan Zheng Lim, Gerald Gui Ren Sng, Yuhe Ke, Joshua Yi Min Tung, Ryan Jian Zhong, Christopher Ming Yao Koh, Keane Zhi Hao Lee, Xiang Chen, Jack Kian Chng, Aung Than, Ken Junyang Goh, and Daniel Shu Wei Ting. 2024. Development and Testing of a Novel Large Language Model-Based Clinical Decision Support Systems for Medication Safety in 12 Clinical Specialties. arXiv:2402.01741 [cs.CL] https://arxiv.org/abs/2402.01741

[128] Jasmine Chiat Ling Ong, Yilin Ning, Mingxuan Liu, Yian Ma, Zhao Liang, Kuldev Singh, Robert T Chang, Silke Vogel, John CW Lim, Iris Siu Kwan Tan, et al. 2025. Regulatory science innovation for generative AI and large language models in health and medicine: a global call for action. *arXiv preprint arXiv:2502.07794* (2025).

[129] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, et al. 2024. GREEN: Generative Radiology Report Evaluation and Error Notation. *arXiv preprint arXiv:2405.03595* (2024). https://arxiv.org/abs/2405.03595

[130] Serguei Pakhomov, Jacob Solinsky, Martin Michalowski, and Veronika Bachanova. 2023. A Conversational Agent for Early Detection of Neurotoxic Effects of Medications through Automated Intensive Observation. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024.* World Scientific, 24–38.

[131] Dhavalkumar Patel, Ganesh Raut, Satya Narayan Cheetirala, Benjamin Glicksberg, Matthew A Levin, Girish Nadkarni, Robert Freeman, Eyal Klang, and Prem Timsina. 2025. AI Agents in Modern Healthcare: From Foundation to Pioneer–A Comprehensive Review and Implementation Roadmap for Impact and Integration in Clinical Settings. (2025).

[132] M Pavithra and A Indhuja. 2025. Synergistic joint model of knowledge graph and llm for enhancing xai-based clinical decision support systems. *Mathematics* 13, 6 (2025), 949. https://www.mdpi.com/2227-7390/13/6/949

[133] Chantal Pellegrini, Ege Özsoy, David Bani-Harouni, Matthias Keicher, and Nassir Navab. 2025. From EHRs to Patient Pathways: Scalable Modeling of Longitudinal Health Trajectories with LLMs. *arXiv preprint arXiv:2506.04831* (2025).

[134] Alexander R Pelletier, Joseph Ramirez, Baradwaj Simha Sankar, Irsyad Adam, Yu Yan, Dylan Steinecke, Wei Wang, Karol E Watson, and Peipei Ping. 2025. Evidence-based knowledge synthesis and hypothesis validation: Navigating biomedical knowledge bases via explainable ai and agentic systems. *Journal of Visualized Experiments (JoVE)* 220 (2025), e67525.

[135] Qi Peng, Jialin Cui, Jiayuan Xie, Yi Cai, and Qing Li. 2025. Tree-of-Reasoning: Towards Complex Medical Diagnosis via Multi-Agent Reasoning with Evidence Tree. *arXiv preprint arXiv:2508.03038* (2025).

[136] Alejandra Perez, Han Zhang, Yu-Chun Ku, Lalithkumar Seenivasan, Roger Soberanis, Jose L Porras, Richard Day, Jeff Jopling, Peter Najjar, and Mathias Unberath. 2025. Privacy-Preserving Operating Room Workflow Analysis using Digital Twins. *arXiv preprint arXiv:2504.12552* (2025).

[137] Mathis Pink, Qinyuan Wu, Vy Ai Vo, Javier Turek, Jianing Mu, Alexander Huth, and Mariya Toneva. 2025. Position: Episodic Memory is the Missing Piece for Long-Term LLM Agents. *arXiv preprint arXiv:2502.06975* (2025).

[138] Chongyu Qu, Allen J Luna, Thomas Z Li, Junchao Zhu, Junlin Guo, Juming Xiong, Kim L Sandler, Bennett A Landman, and Yuankai Huo. 2025. Cohort-Aware Agents for Individualized Lung Cancer Risk Prediction Using a Retrieval-Augmented Model Selection Framework. *arXiv preprint arXiv:2508.14940* (2025).

[139] Yuanhao Qu, Kaixuan Huang, Ming Yin, Kanghong Zhan, Dyllan Liu, Di Yin, Henry C Cousins, William A Johnson, Xiaotong Wang, Mihir Shah, et al. 2025. CRISPR-GPT for agentic automation of gene-editing experiments. *Nature Biomedical Engineering* (2025), 1–14.

[140] Vishal Raman, Abhijith Ragav, et al. 2025. REMI: A Novel Causal Schema Memory Architecture for Personalized Lifestyle Recommendation Agents. *arXiv preprint arXiv:2509.06269* (2025).

[141] Sina Rashidian, Nan Li, Jonathan Amar, Jong Ha Lee, Sam Pugh, Eric Yang, Geoff Masterson, Myoung Cha, Yugang Jia, and Akhil Vaid. 2025. AI Agents for Conversational Patient Triage: Preliminary Simulation-Based Evaluation with Real-World EHR Data. *arXiv preprint arXiv:2506.04032* (2025).

[142] Weijieying Ren, Jingxi Zhu, Zehao Liu, Tianxiang Zhao, and Vasant Honavar. 2025. A Comprehensive Survey of Electronic Health Record Modeling: From Deep Learning Approaches to Large Language Models. *arXiv preprint arXiv:2507.12774* (2025).

[143] Zhiwei Ren, Junbo Li, Minjia Zhang, Di Wang, Xiaoran Fan, and Longfei Shangguan. 2025. *Toward Sensor-In-the-Loop LLM Agent: Benchmarks and Implications*. Association for Computing Machinery, New York, NY, USA, 254–267. https://doi.org/10.1145/3715014.3722082

[144] Mohammad Reza Rezaei, Reza Saadati Fard, Jayson L Parker, Rahul G Krishnan, and Milad Lankarany. 2025. Agentic Medical Knowledge Graphs Enhance Medical Question Answering: Bridging the Gap Between LLMs and Evolving Medical Knowledge. *arXiv preprint arXiv:2502.13010* (2025).

[145] Lars Riedemann, Maxime Labonne, and Stephen Gilbert. 2024. The path forward for large language models in medicine is open. *npj Digital Medicine* 7, 1 (2024), 339.

[146] Jose M Ruiz Mejia and Danda B Rawat. 2025. MedScrubCrew: A Medical Multi-Agent Framework for Automating Appointment Scheduling Based on Patient-Provider Profile Resource Matching. In *Healthcare*, Vol. 13. MDPI, 1649.

[147] Gulfam Ahmed Sajua, Marjan Akhib, and Yuchou Chang. 2025. AgentMRI: A Vison Language Model-Powered AI System for Self-regulating MRI Reconstruction with Multiple Degradations. *Journal of imaging informatics in medicine* (2025), 1–19.

[148] Jerome H Saltzer and Michael D Schroeder. 1975. The protection of information in computer systems. *Proc. IEEE* 63, 9 (1975), 1278–1308.

[149] Ranjan Sapkota, Konstantinos I Roumeliotis, and Manoj Karkee. 2025. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468* (2025).

[150] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36 (2023), 68539–68551.

[151] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960* (2024).

[152] George Shaikovski, Eugene Vorontsov, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H Bernhard, Ran A Godrich, Juan A Retamero, et al. 2025. PRISM2: Unlocking Multi-Modal General Pathology AI with Clinical Dialogue. *arXiv preprint arXiv:2506.13063* (2025).

[153] Tianqi Shang, Weiqing He, Charles Zheng, Lingyao Li, Li Shen, and Bingxin Zhao. 2025. DynamiCare: A Dynamic Multi-Agent Framework for Interactive and Open-Ended Medical Decision-Making. *arXiv preprint arXiv:2507.02616* (2025).

[154] Chen Shen, Wanqing Zhang, Kehan Li, Erwen Huang, Yingbin Wang, Jiyong Sun, Jiayi Yu, Yangyang Zhang, Zixin Li, Yuxuan Chen, Yi Zhu, Wenze Ding, Shouhong Wang, Hong Li, Shaowen Yao, Yilan Ruan, and Guanhua Du. 2025. FEAT: A Multi-Agent Forensic AI System with Domain-Adapted Large Language Model for Automated Cause-of-Death Analysis. (2025). arXiv:2508.07950 [cs.CV] doi:10.48550/arXiv.2508.07950

[155] Yiqing Shen, Chenjia Li, Bohan Liu, Cheng-Yi Li, Tito Porras, and Mathias Unberath. 2025. Operating room workflow analysis via reasoning segmentation over digital twins. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 415–424.

[156] 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. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, Vol. 2024. 22315.

[157] Soorya Ram Shimgekar, Shayan Vassef, Abhay Goyal, Navin Kumar, and Koustuv Saha. 2025. Agentic ai framework for end-to-end medical data inference. *arXiv preprint arXiv:2507.18115* (2025).

[158] Francisco RE Silva, Pedro A Santos, and João Dias. 2025. MentalRAG: Developing an Agentic Framework for Therapeutic Support Systems. In *11th International Conference on Information and Communication Technologies for Ageing Well and e-Health, ICT4AWE 2025*. Science and Technology Publications, Lda, 46–57.

[159] Kevin Song, Andrew Trotter, and Jake Y Chen. 2025. Llm agent swarm for hypothesis-driven drug discovery. *arXiv preprint arXiv:2504.17967* (2025).

[160] Yongwoo Song, Minbyul Jeong, and Mujeen Sung. 2025. Trustworthy Agents for Electronic Health Records through Confidence Estimation. *arXiv preprint arXiv:2508.19096* (2025). https://arxiv.org/abs/2508.19096

[161] Moran Sorka, Alon Gorenshtein, Dvir Aran, and Shahar Shelly. 2025. A Multi-Agent Approach to Neurological Clinical Reasoning. *arXiv preprint arXiv:2508.14063* (2025).

[162] Ian Steenstra and Timothy W Bickmore. 2025. A Risk Taxonomy for Evaluating AI-Powered Psychotherapy Agents. *arXiv preprint arXiv:2505.15108* (2025).

[163] Ian Steenstra, Prasanth Murali, Rebecca B Perkins, Natalie Joseph, Michael K Paasche-Orlow, and Timothy Bickmore. 2024. Engaging and entertaining adolescents in health education using llm-generated fantasy narrative games and virtual agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[164] Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2024. KGARevion: an AI agent for knowledge-intensive biomedical QA. *arXiv preprint arXiv:2410.04660* (2024).

[165] 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. *arXiv preprint arXiv:2408.01112* (2024).

[166] Thanathip Suenghataiphorn, Pojsakorn Danpanichkul, Narisara Tribuddharat, and Narathorn Kulthamrongsri. 2025. Toward Real-Time Detection of Drug-Induced Liver Injury Using Large Language Models: A Feasibility Study from Clinical Notes. *Journal of Clinical and Experimental Hepatology* (2025), 102627.

[167] Yuxuan Sun, Yixuan Si, Chenglu Zhu, Kai Zhang, et al. 2025. CPathAgent: An Agent-based Foundation Model for Interpretable High-Resolution Pathology Image Analysis Mimicking Pathologists' Diagnostic Logic. *arXiv preprint arXiv:2505.20510* (2025). https://arxiv.org/abs/2505.20510

[168] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. 2024. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. *arXiv preprint arXiv:2407.00203* (2024).

[169] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 952–966.

[170] Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, et al. 2025. MedAgentsBench: Benchmarking Thinking Models and Agent Frameworks for Complex Medical Reasoning. *arXiv preprint arXiv:2503.07459* (2025). https://arxiv.org/abs/2503.07459

[171] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537* (2023).

[172] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical AI. *Nejm Ai* 1, 3 (2024), AIoa2300138.

[173] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature* (2025), 1–9.

[174] Shubham Vatsal, Harsh Dubey, and Aditi Singh. 2025. AGENTIC AI IN HEALTHCARE & MEDICINE: A SEVEN-DIMENSIONAL TAXONOMY FOR EMPIRICAL EVALUATION OF LLM-BASED AGENTS. *Authorea Preprints* (2025).

[175] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.

[176] Chen-Kai Wang, Cheng-Rong Ke, Ming-Siang Huang, Inn-Wen Chong, Yi-Hsin Yang, Vincent S Tseng, and Hong-Jie Dai. 2024. Using large language models for efficient cancer registry coding in the real hospital setting: A feasibility study. In *Biocomputing 2025: Proceedings of the Pacific Symposium*. World Scientific, 121–137.

[177] Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196* (2025).

[178] Jiayi Wang, Jacqueline Jil Vallon, Neil Panjwani, Xi Ling, Sushmita Vij, Sandy Srinivas, John Leppert, Mark K Buyyounouski, and Mohsen Bayati. 2025. Agent-Based Feature Generation from Clinical Notes for Outcome Prediction. *arXiv preprint arXiv:2508.01956* (2025).

[179] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.

[180] Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. 2025. AnnaAgent: Dynamic Evolution Agent System with Multi-Session Memory for Realistic Seeker Simulation. *arXiv preprint arXiv:2506.00551* (2025).

[181] Qingxin Wang, Zhongqiu Wang, Minghua Li, Xinye Ni, Rong Tan, Wenwen Zhang, Maitudi Wubulaishan, Wei Wang, Zhiyong Yuan, Zhen Zhang, et al. 2025. A feasibility study of automating radiotherapy planning with large language model agents. *Physics in Medicine & Biology* 70, 7 (2025), 075007.

[182] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211* (2025).

[183] Yihan Wang, Qiao Yan, Zhenghao Xing, Lihao Liu, Junjun He, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. 2025. Silence is Not Consensus: Disrupting Agreement Bias in Multi-Agent LLMs via Catfish Agent for Clinical Decision Making. (2025). arXiv:2505.21503 [cs.AI] doi:10.48550/arXiv.2505.21503

[184] Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. 2025. GeneAgent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods* (2025), 1–9.

[185] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025. MedAgent-Pro: Towards Evidence-Based Multi-Modal Medical Diagnosis via Reasoning Agentic Workflow. *arXiv preprint arXiv:2503.18968* (2025).

[186] Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, et al. 2025. 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.

[187] Lilija Wehling, Gurdeep Singh, Ahmad Wisnu Mulyadi, Rakesh Hadne Sreenath, Henning Hermjakob, Tung Nguyen, Thomas Rückle, Mohammed H. Mosa, Henrik Cordes, Tommaso Andreani, Thomas Klabunde, Rahuman S. Malik-Sheriff, and Douglas McCloskey. 2025. Talk2Biomodels: AI agent-based open-source LLM initiative for kinetic biological models. *bioRxiv* (2025). arXiv:https://www.biorxiv.org/content/early/2025/03/12/2025.03.11.642548.full.pdf doi:10.1101/2025.03.11.642548

[188] Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. 2024. Medco: Medical education copilots based on a multi-agent framework. In *European Conference on Computer Vision*. Springer, 119–135.

[189] Dan Weisman, Alanna Sugarman, Yue Ming Huang, Lillian Gelberg, Patricia A Ganz, and Warren Scott Comulada. 2025. Development of a GPT-4–Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study. *JMIR Form Res* 9 (17 Apr 2025), e65670. doi:10.2196/65670

[190] Hao Wu, Yinghao Zhu, Zixiang Wang, Xiaochen Zheng, Ling Wang, Wen Tang, Yasha Wang, Chengwei Pan, Ewen M Harrison, Junyi Gao, et al. 2024. EHRFlow: A Large Language Model-Driven Iterative Multi-Agent Electronic Health Record Data Analysis Workflow. In *KDD'24 Workshop: Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

[191] Jinlin Wu, Xusheng Liang, Xuexue Bai, and Zhen Chen. 2024. Surgbox: Agent-driven operating room sandbox with surgery copilot. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2041–2048.

[192] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical Graph RAG: Evidence-based Medical Large Language Model via Graph Retrieval-Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 28443–28467.

[193] Zhen Xiang, Aliyah R Hsu, Austin V Zane, Aaron E Kornblith, Margaret J Lin-Martore, Jasmanpreet C Kaur, Vasuda M Dokiparthi, Bo Li, and Bin Yu. 2025. CDR-Agent: Intelligent Selection and Execution of Clinical Decision Rules Using Large Language Model Agents. *arXiv preprint arXiv:2505.23055* (2025).

[194] Ying Xiao, Jie Huang, Ruijuan He, Jing Xiao, Mohammad Reza Mousavi, Yepang Liu, Kezhi Li, Zhenpeng Chen, and Jie M Zhang. 2025. AMQA: An Adversarial Dataset for Benchmarking Bias of LLMs in Medicine and Healthcare. *arXiv preprint arXiv:2505.19562* (2025).

[195] Yuzhang Xie, Hejie Cui, Ziyang Zhang, Jiaying Lu, Kai Shu, Fadi Nahab, Xiao Hu, and Carl Yang. 2025. KERAP: A Knowledge-Enhanced Reasoning Approach for Accurate Zero-shot Diagnosis Prediction Using Multi-agent LLMs. *arXiv preprint arXiv:2507.02773* (2025).

[196] Ancheng Xu, Di Yang, Renhao Li, Jingwei Zhu, Minghuan Tan, Min Yang, Wanxin Qiu, Mingchen Ma, Haihong Wu, Bingyu Li, et al. 2025. Autocbt: An autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling. *arXiv preprint arXiv:2501.09426* (2025).

[197] Gelei Xu, Yawen Wu, Zhenge Jia, Jingtong Hu, and Yiyu Shi. 2025. Fair Dermatological Disease Diagnosis Through Auto-weighted Federated Learning and Performance-Aware Personalization. In *MICCAI Workshop on Fairness of AI in Medical Imaging*. Springer, 167–176.

[198] Huimin Xu, Kun Li, Zijian Zhang, Ying Wang, Xinhang Zhang, Yuntian Gong, Jiahua Zhang, Jiahong Li, Yi Jiang, and Min Yang. 2025. TAMA: A Human–AI Collaborative Thematic Analysis Framework Using Multi-Agent LLMs for Clinical Interviews. (2025). arXiv:2503.20666 [cs.AI] doi:10.48550/arXiv.2503.20666

[199] Nicholas Yan and Gil Alterovitz. 2024. A general-purpose AI avatar in healthcare. *arXiv preprint arXiv:2401.12981* (2024).

[200] Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, et al. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *arXiv preprint arXiv:2406.13890* (2024).

[201] Dongrong Yang, Xin Wu, Yibo Xie, Xinyi Li, Qiuwen Wu, Jackie Wu, and Yang Sheng. 2025. Zero-Shot Large Language Model Agents for Fully Automated Radiotherapy Treatment Planning. *arXiv preprint arXiv:2510.11754* (2025).

[202] Ei-Wen Yang and Enrique Velazquez-Villarreal. 2025. AI-HOPE: An AI-Driven conversational agent for enhanced clinical and genomic data integration in precision medicine research. *Bioinformatics* 41, 7 (2025), btaf359.

[203] Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. Psychogat: A novel psychological measurement paradigm through interactive fiction games with llm agents. *arXiv preprint arXiv:2402.12326* (2024).

[204] Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Kit Phey Leng, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-peng Lim. 2025. Cami: A counselor agent supporting motivational interviewing through state inference and topic exploration. *arXiv preprint arXiv:2502.02807* (2025).

[205] Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, et al. 2025. A survey of ai agent protocols. *arXiv preprint arXiv:2504.16736* (2025).

[206] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Shao Zhang, Ethan Rogers, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2023. Talk2Care: Facilitating asynchronous patient-provider communication with large-language-model. *arXiv preprint arXiv:2309.09357* (2023).

[207] Ziruo Yi, Jinyu Liu, Ting Xiao, and Mark V. Albert. 2025. A Multi-Agent System for Complex Reasoning in Radiology Visual Question Answering. (2025). arXiv:2508.02841 [cs.CV] doi:10.48550/arXiv.2508.02841

[208] Ziruo Yi, Ting Xiao, and Mark V Albert. 2025. A Multimodal Multi-Agent Framework for Radiology Report Generation. *arXiv preprint arXiv:2505.09787* (2025).

[209] Chengzhang Yu, Yiming Zhang, Zhixin Liu, Zenghui Ding, Yining Sun, and Zhanpeng Jin. 2025. FRAME: Feedback-Refined Agent Methodology for Enhancing Medical Research Insights. *arXiv preprint arXiv:2505.04649* (2025).

[210] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, et al. 2023. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns* 4, 9 (2023). doi:10.1016/j.patter.2023.100813

[211] Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Jingxian He, Wenyue Hua, Mingyu Jin, et al. 2024. Simulated patient systems are intelligent when powered by large language model-based AI agents. *arXiv preprint arXiv:2409.18924* (2024).

[212] Hong Qing Yu and Frank McQuade. 2025. Rag-kg-il: A multi-agent hybrid framework for reducing hallucinations and enhancing llm reasoning through rag and incremental knowledge graph learning integration. *arXiv preprint arXiv:2503.13514* (2025).

[213] Weilun Yu, Shixiang Tang, Yonggui Huang, Nanqing Dong, Li Fan, Honggang Qi, Wei Liu, Xiaoli Diao, Xi Chen, and Wanli Ouyang. 2025. Dynamic Knowledge Exchange and Dual-Diversity Review: Concisely Unleashing the Potential of a Multi-Agent Research Team. (2025). arXiv:2506.18348 [cs.AI] doi:10.48550/arXiv.2506.18348

[214] Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024. 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.

[215] Abdellah Zeggai, Ilyes Traikia, Abdelhak Lakehal, and Abdennour Boulesnane. 2025. AI-VaxGuide: An Agentic RAG-Based LLM for Vaccination Decisions. *arXiv preprint arXiv:2507.03493* (2025).

[216] Fang Zeng, Zhiliang Lyu, Quanzheng Li, and Xiang Li. 2024. Enhancing LLMs for Impression Generation in Radiology Reports through a Multi-Agent System. (2024). arXiv:2412.06828 [cs.CV] doi:10.48550/arXiv.2412.06828

[217] Fan Zhang, Yalong Zhao, Weihan Zhang, and Lipeng Lai. 2025. BioScientist Agent: Designing LLM-Biomedical Agents with KG-Augmented RL Reasoning Modules for Drug Repurposing and Mechanistic of Action Elucidation. *bioRxiv* (2025), 2025–08.

[218] Han Zhang, KaWing Tsang, and Zhenhui Peng. 2025. VChatter: Exploring Generative Conversational Agents for Simulating Exposure Therapy to Reduce Social Anxiety. *arXiv preprint arXiv:2506.03520* (2025).

[219] Weitong Zhang, Yifan Li, Ruina Li, Cristiane Balsanelli, Syed Raza, Jeffrey De Fauw, Xiaoyi Ying, Nima Kazem, Diana Dey, Lisa Peng, Valentina Carapella, Mingxing Xie, Ting Zhang, Emily Mostofsky, Omar Sorour, Douglas Lee, Suma Sarma, Shing Lee, Wenjia Bai, M. Jorge Cardoso, and Mehul Desai. 2025. Multi-Agent Reasoning for Cardiovascular Imaging Phenotype Analysis. (2025). arXiv:2507.03460 [cs.CV] doi:10.48550/arXiv.2507.03460

[220] Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, et al. 2025. Personaagent: When large language model agents meet personalization at test time. *arXiv preprint arXiv:2506.06254* (2025).

[221] Yiqun Zhang, Xiaocui Yang, Xiaobai Li, Siyuan Yu, Yi Luan, Shi Feng, Daling Wang, and Yifei Zhang. 2024. PsyDraw: A Multi-Agent Multimodal System for Mental Health Screening in Left-Behind Children. arXiv:2412.14769 [cs.CL] https://arxiv.org/abs/2412.14769

[222] Zhenxuan Zhang, Kinhei Lee, Weihang Deng, Huichi Zhou, et al. 2025. GEMA-Score: Granular Explainable Multi-Agent Score for Radiology Report Evaluation. *arXiv preprint arXiv:2503.05347* (2025). https://arxiv.org/abs/2503.05347

[223] Huiya Zhao, Yinghao Zhu, Zixiang Wang, Yasha Wang, Junyi Gao, and Liantao Ma. 2025. ConfAgents: A Conformal-Guided Multi-Agent Framework for Cost-Efficient Medical Diagnosis. *arXiv preprint arXiv:2508.04915* (2025).

[224] Weike Zhao, Chaoyi Wu, Yanjie Fan, Xiaoman Zhang, Pengcheng Qiu, Yuze Sun, Xiao Zhou, Yanfeng Wang, Ya Zhang, Yongguo Yu, et al. 2025. An Agentic System for Rare Disease Diagnosis with Traceable Reasoning. *arXiv preprint arXiv:2506.20430* (2025).

[225] Yutian Zhao, Huimin Wang, Yefeng Zheng, and Xian Wu. 2025. A Layered Debating Multi-Agent System for Similar Disease Diagnosis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. 539–549.

[226] Qiaoyu Zheng, Yuze Sun, Chaoyi Wu, Weike Zhao, Pengcheng Qiu, Yongguo Yu, Kun Sun, Yanfeng Wang, Ya Zhang, and Weidi Xie. 2025. End-to-End Agentic RAG System Training for Traceable Diagnostic Reasoning. arXiv:2508.15746 [cs.CL] https://arxiv.org/abs/2508.15746

[227] Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. 2025. Customizing emotional support: How do individuals construct and interact with LLM-powered chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[228] Zhusi Zhong, Yuli Wang, Jing Wu, Wen-Chi Hsu, Vin Somasundaram, Lulu Bi, Shreyas Kulkarni, Zhuoqi Ma, Scott Collins, Grayson Baird, et al. 2025. Vision-language model for report generation and outcome prediction in CT pulmonary angiogram. *NPJ Digital Medicine* 8, 1 (2025), 432.

[229] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112* (2023).

[230] Kailai Zhou, Chen Qian, Qianxi Liu, Wei Zhang, and Zhou Zhao. 2025. EmoAgent: Assessing and Safeguarding Human-AI Interaction for Mental Health Safety. *arXiv preprint arXiv:2504.09689* (2025). http://arxiv.org/abs/2504.09689v3

[231] Xinyang Zhou, Yongyong Ren, Qianqian Zhao, Daoyi Huang, Xinbo Wang, Tingting Zhao, Zhixing Zhu, Wenyuan He, Shuyuan Li, Yan Xu, et al. 2025. An LLM-Driven Multi-Agent Debate System for Mendelian Diseases. *arXiv preprint arXiv:2504.07881* (2025).

[232] Yue Zhou, Barbara Di Eugenio, and Lu Cheng. 2024. Unveiling performance challenges of large language models in low-resource healthcare: A demographic fairness perspective. *arXiv preprint arXiv:2412.00554* (2024).

[233] Yucheng Zhou, Lingran Song, and Jianbing Shen. 2025. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. *arXiv preprint arXiv:2506.19835* (2025).

[234] Yuan Zhou, Peng Zhang, Mengya Song, Alice Zheng, Yiwen Lu, Zhiheng Liu, Yong Chen, and Zhaohan Xi. 2024. Zodiac: A cardiologist-level llm framework for multi-agent diagnostics. *arXiv preprint arXiv:2410.02026* (2024).

[235] Jared Zhu and Junde Wu. 2025. MedicalOS: An LLM Agent based Operating System for Digital Healthcare. *arXiv preprint arXiv:2509.11507* (2025).

[236] Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2025. MedAgentBoard: Benchmarking Multi-Agent Collaboration with Conventional Methods for Diverse Medical Tasks. *arXiv preprint arXiv:2505.12371* (2025). https://arxiv.org/abs/2505.12371

[237] Yinghao Zhu, Yifan Qi, Zixiang Wang, Lei Gu, Dehao Sui, Haoran Hu, Xichen Zhang, Ziyi He, Liantao Ma, and Lequan Yu. 2025. HealthFlow: A Self-Evolving AI Agent with Meta Planning for Autonomous Healthcare Research. *arXiv preprint arXiv:2508.02621* (2025). https://arxiv.org/abs/2508.02621

[238] 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. *arXiv preprint arXiv:2410.13610* (2024).

[239] Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2025. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis. In *AAAI Bridge Program on AI for Medicine and Healthcare*. PMLR, 195–204.

[240] Kaiwen Zuo, Zixuan Zhong, Peizhou Huang, Shiyan Tang, Yuyan Chen, and Yirui Jiang. 2025. HEAL-KGGen: A Hierarchical Multi-Agent LLM Framework with Knowledge Graph Enhancement for Genetic Biomarker-Based Medical Diagnosis. *bioRxiv* (2025), 2025–06.