A Survey on LLM-based Multi-Agent AI Hospital

Preprint · March 2025

DOI: 10.31219/osf.io/bv5sg_v1

2 authors, including:

Zonghai Yao
University of Massachusetts Amherst
51 PUBLICATIONS **228** CITATIONS

SEE PROFILE

A Survey on LLM-based Multi-Agent AI Hospital

Zonghai Yao
University of Massachusetts, Amherst
zonghaiyao@umass.edu

Hong Yu
University of Massachusetts, Lowell
University of Massachusetts, Amherst
UMass Chan Medical School
Hong_Yu@uml.edu

Abstract

AI hospitals use agent-driven multi-agent systems based on large language models (LLMs) to automate and optimize medical workflows, enabling intelligent agents to understand, reason, and assist in complex medical tasks. Although AI-driven healthcare applications are developing rapidly, research remains fragmented across various scenarios. This survey carefully analyzes 72 studies on LLM-based medical agents published between 2023 and 2025, provides a comprehensive review of AI hospitals, and systematically introduces a structured taxonomy to categorize its core components and applications. Additionally, we explored the key challenges associated with the core components of AI hospitals, including agent roles, interaction patterns, reasoning mechanisms, memory management, and tool integration. Finally, we explore how to further develop the AI hospital into a more meaningful research and practice platform that supports medical simulation, complex problem-solving, evaluation, and synthetic data generation. This, in turn, will accelerate progress in clinical reasoning, decision support, and AI-driven medical innovation, highlighting the crucial role of AI hospitals as the foundational framework for AI-powered healthcare ecosystems. By providing a structured perspective, this survey bridges AI and healthcare research, providing a roadmap for strengthening interdisciplinary collaboration and the practical applicability of AI hospitals.

1 Introduction

With the rapid development of artificial intelligence (AI), large language models (LLMs) have demonstrated impressive agent capabilities, enabling them to operate autonomously in various fields [147]. By 2025, AI agents are expected to significantly improve productivity in industries such as finance [75, 86, 99], education [137, 154], and especially healthcare [125, 39, 74]. Researchers

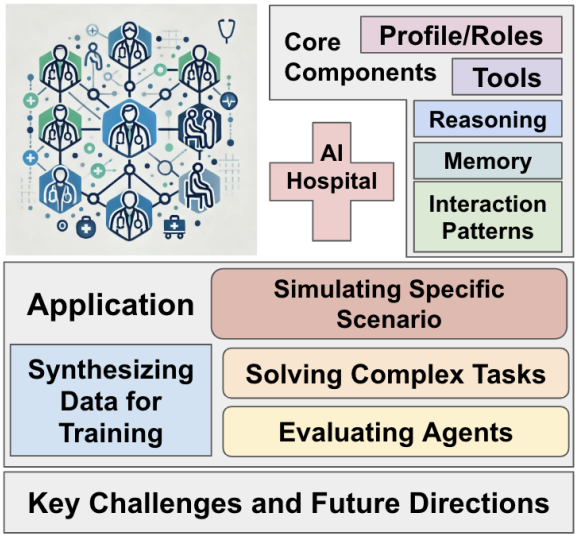

Figure 1: Overview of the LLM-based multi-agent AI hospital system. Figure 2 & 3 present the taxonomy of core components and various applications separately. Table 2 outlines the key challenges and future directions.

are actively exploring how AI agents can collaborate to complete complex tasks, improve decision-making processes, automate evaluations, and simplify workflows [43, 30, 26, 79, 52, 20].

In healthcare, AI agents play a key role in clinical reasoning, expert assessment, and data generation [73, 138, 31, 159, 83, 171]. The concept of AI hospitals (agent hospitals or virtual hospitals) has become a paradigm where multi-agent systems simulate real-world hospital and laboratory workflows. These environments enable systematic observation of LLM behaviors, rigorous models/agents reliability and efficiency assessment, and controlled simulations for generating high-quality training data.

Despite the growing research interest in AI-driven hospitals, existing studies remain fragmented and lack a unified framework that systematically connects different AI hospital simulations across real-world scenarios. While many works explore specific multi-agent applications, they often do so in isolation, without providing a compre-

hensive perspective on how these systems interact and contribute to a cohesive AI hospital ecosystem (see Table 13 for a detailed list of 72 relevant studies). To the best of our knowledge, no existing multi-agent healthcare survey [39, 73, 74, 138] has systematically examined these works through the lens of AI hospitals, analyzing the essential components, key applications, existing limitations, and future directions necessary for their advancement.

This survey systematically organizes and categorizes AI hospital research in the following three key areas, establishing a structured framework to link isolated studies and provide a holistic perspective. By providing a clear analytical framework for AI researchers and medical professionals, it enhances interdisciplinary understanding and promotes deeper integration of AI with healthcare in both research and practical applications.

- **Core Components**: Analyzing the fundamental elements of AI hospitals, including agent roles, interaction patterns, tool integration, memory management, and reasoning mechanisms.
- **Applications**: Investigating how AI hospitals contribute to simulating specific medical scenarios, solving complex tasks, evaluating agents, and generating synthetic data for training medical AI systems.
- **Key Challenges & Future Directions**: Providing targeted discussions on existing challenges in each core component and application, as well as how to further develop the AI hospital into a more meaningful research and practice platform.

The remainder of this survey is structured as follows: § 2 to § 6 discuss the five core components of AI hospitals, Sections § 7 to § 10 cover four key applications, and Section § 11 explores the challenges and future directions facing each subcomponent and the overall AI hospital paradigm.

2 Core Components: Agent Roles

2.1 Patient-Centered Agents

Patient-centric agents are the core group of agents in AI Hospital. These LLM agents use prompt engineering, retrieval-augmented generation (RAG), and fine-tuning to simulate diverse patient characteristics to enhance realism.

Patient Agent powered by LLMs are designed to simulate patients with different demographic backgrounds, health conditions, and communication abilities. These agents enhance clinical education by providing scalable and cost-effective standardized patient (SP) alternatives, enabling medical trainees to practice patient interactions without logistical constraints [85]. Enhancing the realism of patient agents is a focus emphasized by many works. [11] enhances the realism of patient agents by combining psychological and personality-driven attributes. Some methods use the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) to shape patients' communication style, affecting conversation flow and decision-making. AIPatient [164] adopts knowledge graph and reasoning-driven RAG workflow to improve response robustness, ensure consistency between interactions, and adopt a structured action response framework to mimic real-world patient behavior more effectively. EvoPatient [37] introduces an evolutionary learning mechanism where patient agents iteratively improve their conversational capabilities, thereby improving the consistency between generated responses and standardized medical protocols. NoteChat [135] improves the authenticity of patient agent responses by adding a Polish module to ensure consistency in language style and communication methods. In addition, recent methods integrate RAG and supervised fine-tuning techniques to overcome the weakness of traditional prompting-based methods. For example, patient agents have been improved on large-scale LLMs (such as Qwen2.5-72B-Instruct) through LoRA-based fine-tuning, showing significant improvements in hallucination rate (HR) and response relevance (IRR) [92]. Another method, CureFun, enhances the role-playing ability of LLMs by combining Chain of Thoughts (CoT) reasoning and RAG, ensuring the stability of simulated patient interactions while mitigating hallucinations [85] [1].

Psychological Patient Agent (PPA) simulates mental health conditions for AI-driven treatment training. Unlike general patient agents, PPAs must replicate mood changes, cognitive distortions, and treatment resistance. We observed that PPAs appeared significantly more frequently than other unique patients, which should be because compared with the higher requirements for multimodal capabilities of agents in other scenarios, psychological treatment can be completed more in the form of dialogue, which is more suitable for LLMs with-

[1] For further details, see Appendix Table 3.

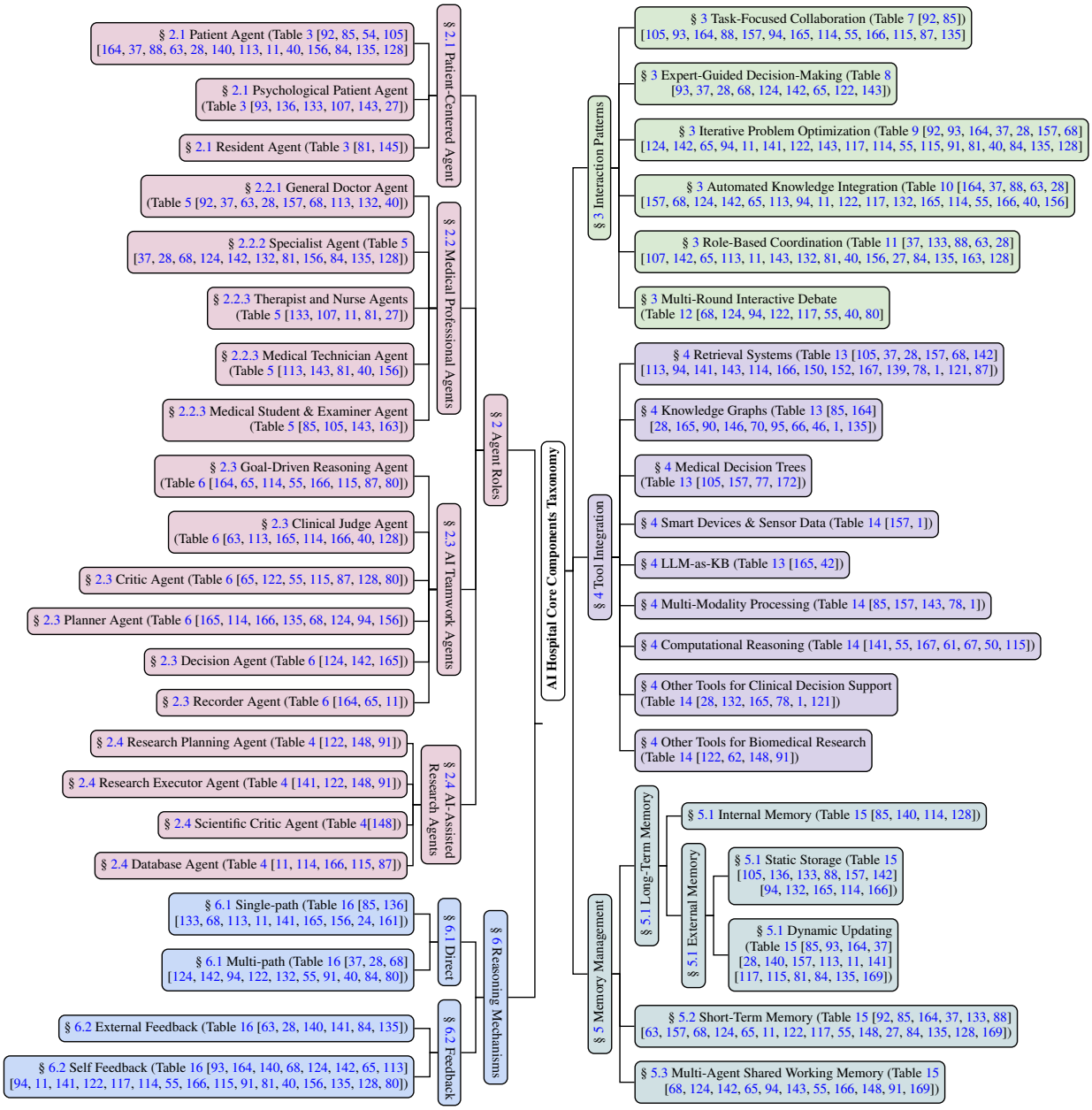

Figure 2: Taxonomy of AI hospital core components. Agent Roles and Reasoning Mechanisms components are on the left, while Interaction Patterns, Tool Integration, and Memory Management are on the right.

out multimodal capabilities in the early stage. As with general patient agents, authenticity is the focus of many studies. For example, [93] used expert-guided prompt engineering to improve compliance with psychological norms and enhance the authenticity of training. PATIENT-Ψ [136] embeds cognitive behavioral therapy (CBT) principles into LLM to achieve structured cognitive modeling. PPA also enhances psychiatry training by simulating honesty, emotional instability, and resistance to treatment, fostering adaptive communication strategies. Patient chatbots complement psychiatric simulations [27]. ClientCAST improves scalability and ethical evaluation by having LLM-generated patients evaluate the treatment process to evaluate AI therapists [133], while MEDCO enables structured diagnostic assessments in a virtual environment [143].

Resident Agents (RAs) model general populations, transitioning into patient agents when they develop illnesses. Unlike static patient agents, RAs simulate real-world disease progression and healthcare-seeking behavior [81]. These agents autonomously undergo triage, consultation, diagnosis, and treatment, tracking recovery and re-engaging with the healthcare system as needed. RAs also contribute to public health simulations, modeling disease spread and policy interventions. RAs in [145] can make health decisions based on symptoms and environmental factors, facilitating epidemiological modeling and intervention strategy evaluation.

2.2 Medical Professional Agents

Medical Professional Agents play a crucial role in AI hospital simulations, including General Doctor

Agents, Specialist Agents, Therapist Agents, Nurse Agents, Medical Technician Agents, Medical Student & Examiner Agents. These agents, empowered by LLMs, can perform tasks such as patient consultation, medical history collection, clinical reasoning, diagnostic decision-making, emotional support, care coordination, and auxiliary examinations, enhancing the automation and intelligence of medical workflows [2].

2.2.1 General Doctor Agent

General practitioner agents, often called primary care physicians (PCPs) or attending physician agents, play a crucial role in AI-driven healthcare simulations by performing initial patient assessments and overseeing the diagnostic process.

Several studies have explored different aspects of these agents, from their questioning and information-gathering strategies to their diagnostic and documentation capabilities. One study [92] systematically analyzed the questioning strategies of different LLM physician models in patient assessments, comparing GPT-4o, GPT-4o-mini, and Claude-3.5-sonnet in five rounds of consultation. The study identified four main types of questioning—chief complaint inquiry, symptom elaboration, related symptom exploration, and family or medical history collection—showing that GPT-4o exhibited the most balanced approach, resulting in improved diagnostic accuracy, while other models tended to be overly detailed or introduce symptoms too early. Another study [37] introduced EvoPatient, a framework that enables PCP agents to autonomously learn query patterns and optimize diagnostic strategies through multiple rounds of interaction. This study highlights the potential of PCP agents to improve initial screening, multidisciplinary consultations, and patient education, demonstrating their superiority over traditional baseline models in history collection and symptom analysis. The CRAFT-MD framework [63] evaluates the reasoning ability of LLMs in simulated clinical conversations, showing that while models such as GPT-4 perform well in static case-based evaluations, they have significantly lower diagnostic accuracy in interactive multi-turn conversations.

Other studies have explored adaptive multi-agent collaboration, such as MDAgents [68], where PCP agents handle less complex cases while more specialized agents take over complex diagnoses. Similarly, DrHouse [157] uses an LLM-driven virtual doctor system to iteratively refine diagnoses over multiple rounds of consultations by combining symptom inquiry with real-time sensor data. RareAgents [28] proposes a novel approach to rare disease diagnosis, where primary care physician agents dynamically form multidisciplinary teams (MDTs) and integrate knowledge from domain-specific experts. In this setting, general practitioner agents play a key role in conducting initial patient assessments and providing critical input for further expert-driven evaluations. AMSC [132] further emphasizes the role of general practitioner agents in diagnosis, modeling an open-source LLM as a primary care physician responsible for initial symptom analysis and multiple-choice question answering (MCQA) tasks, highlighting the differences between general practitioner agents and domain-specific expert models. In addition, [40] points out that integrating general practitioner agents into an AI hospital environment requires strong order-following capabilities, as these agents must autonomously extract symptoms, recommend diagnostic tests, and synthesize findings to identify diseases accurately.

2.2.2 Specialist Agent

Specialist agents, representing domain-specific medical experts such as cardiologists, radiologists, and hematologists, play a vital role in handling complex cases and contributing expert knowledge to diagnostic and treatment decision-making. Unlike PCP, specialist agents require high-precision reasoning, deep medical expertise, and the ability to collaborate effectively in MDTs. For example, the RareAgents [28] demonstrated the benefits of predefined specialist pools, where expert-defined roles significantly enhanced diagnostic accuracy and treatment reliability compared to LLMs autonomously selecting specialists. This highlights the importance of explicitly structured expertise in handling complex medical tasks. Similarly, MDAgents [68] introduced a recruitment mechanism where specialists are selected based on case complexity, operating independently or as part of a coordinated MDT. Within the MEDAGENTS framework [124], specialist agents contributed to a multi-stage reasoning process, where individual experts provided in-depth analysis in their respective domains before an MDT agent consolidated their findings into a comprehensive medical report. The MDT agent was pivotal in resolving disagreements among specialists, ensuring robust and

[2] For further details, see Appendix Table 5.

well-balanced medical decisions through iterative discussions and consensus-building mechanisms. Similar approaches have been explored in AI Hospital environments [156], where multiple specialist agents powered by different LLMs engage in real-time consultations to collaboratively determine diagnoses and treatment plans. Interestingly, [132] found that introducing specialist agents trained in NIH-curated disease knowledge demonstrated significantly improved diagnostic recall compared to general doctor agents. This highlights their effectiveness in leveraging domain-specific knowledge for enhanced diagnostic precision.

2.2.3 Other Medical Professional Agent

Therapist Agent provides emotional support, psychological intervention and psychotherapy [133, 107, 27]. **Nurse Agent** facilitates triage, basic care and patient coordination. The three-step interaction model (action selection, feedback, response generation) in [11] simplifies outpatient reception and ensures effective patient guidance and administrative support. Nurse agent in [81] assists in triage and routine clinical intervention to improve workflow efficiency. **Medical Technician Agents** aid diagnostic procedures, ensuring accurate test results. Examiner Agents in [113] return predefined laboratory and imaging findings, maintaining realism in simulated diagnoses. LLM-based radiology simulations interpret X-rays, CT, MRI, and ultrasound, providing structured diagnostic reports [143]. ClinicalAgent [156] integrates diagnostic agents within MDT frameworks, employing biochemists for laboratory tests and radiologists for imaging diagnostics, refining medical workflows. Examiner Agents in [40] followed structured workflows to validate physician requests, ensuring reliable test outcomes. **Medical students & Examiner agent** Simulate clinical training to improve medical history collection and diagnostic skills. CureFun combines structured case graphs, RAGs, and automatic scoring to evaluate medical students' performance, showing high reliability but limitations in complex reasoning [85]. MedQA-CS [163] introduced MedStuLLM for procedural skill evaluation, with MedExamLLM functioning as an AI-driven examiner, assessing clinical information collection, physical exams, and differential diagnoses. These frameworks strengthen AI-driven medical education and provide scalable and structured assessments.

2.3 Medical AI Teamwork Agents

Unlike the previous two categories, which simulate real-world roles, AI Teamwork Agents focus on tackling complex tasks in an AI hospital that a single agent cannot efficiently complete. These agents are responsible for information extraction, logical reasoning, and critical decision-making processes in the medical workflow, including disease analysis, diagnosis evaluation, patient triage, medical planning, and final decision-making. By appropriately configuring different types of teamwork agents, AI hospitals can enhance overall collaboration efficiency, making the diagnostic process more precise and standardized. Below, we explore the functions of various agents and their implementations in existing research [3].

Goal-Driven Reasoning Agent is vital in AI hospital systems, guiding tasks toward predefined goals by coordinating multi-step reasoning processes. In the AIPatient [164], the reasoning agent consists of a retrieval agent, an abstraction agent, a knowledge-graph-query-generation agent, and a rewriting agent, forming a continuous information processing pipeline. The retrieval agent extracts relevant nodes and edges from the AIPatient knowledge graph based on natural language queries to build an initial knowledge base. The abstraction agent refines and promotes the query to a higher-level concept representation to help subsequent reasoning. Then, the knowledge-graph-query-generation agent converts the abstract query into a precise Cypher query to retrieve structured information, while the rewriting agent converts the query results into natural language responses tailored to patient needs. This integration ensures a seamless transition from information extraction to response generation, converting user queries into actionable outputs. Similarly, the generator-reasoner framework in [55] adopts a dual-agent system, where the generator generates arguments, and the reasoner is equipped with a symbolic solver to evaluate their validity by tracking semantic relations between arguments (e.g., attack or support). The iterative process produces a structured abstract argument framework that enhances decision support through symbolic reasoning. Other goal-driven reasoning agents facilitate planning, execution, and interactive feedback. The interactive coding and execution feedback agent [115] dynamically optimizes task performance by iteratively

[3] For further details, see Appendix Table 6.

refining the execution plan by interacting with the LLM agent with the code executor.

Clinical Judge Agent evaluates AI-driven diagnoses and reasoning for compliance with several criteria. For example, the CRAFT-MD framework [63] uses the Grader AI Agent to assess LLMs' accuracy in multi-turn diagnostic conversations automatically. The efficacy and safety agents in ClinicalAgent [165] analyze the effectiveness and risks of drugs. The supervisor agent in [114] ensures compliance with instructions. The fact locator in [166] improves knowledge verification. The chief physician agent in [40] reviews the AI-generated summary after diagnosis. The moderator in [128] oversees the AI-human conversation to decide when to terminate.

Critic Agent improving AI performance through different types of feedback. JR2 and SD agents in [65] counteract bias by critiquing diagnoses and refining differentials. The verifier in [55] uses self-questioning to improve logic. The self-playing critic in [115, 128] enhances model reliability and medical dialogue by reviewing diagnostic interactions. The supervisor agent in [80] challenges reasoning assumptions and promotes robust decision-making.

Planning Agent structuring workflows by breaking problems into subtasks. The planning agent in [165] employs least-to-most reasoning to improve the accuracy of clinical trials. The navigation agent in [114] optimizes task assignment. NoteChat's planning module [135] structures conversations using keywords to reduce hallucinations. Some planning agents act as triage responsibility to optimize patient prioritization. While this role could also belong to a nurse agent, some studies do not explicitly involve real-world role-playing but rather assign tasks to the next functional agent in the team. For example, the moderators and recruiters in the [68] agents assess query complexity and assign cases to the appropriate team. The hierarchical classification system [94] improves triage by integrating RAG and confidence scores. Navigation agent [156] will forward complaints directly to relevant specialists to ensure efficient consultation.

Decision Agent facilitate consensus-based clinical decision-making. The Decision Agent in [124] synthesizes findings from multiple sources to derive final clinical conclusions. The ColaCare framework [142] uses MetaAgents to coordinate different DoctorAgents, mediate conflicting assessments, and refine reports to produce coherent, scientifically valid diagnoses.

Recording Agent records key insights from medical interactions. The Summarization Agent in [164] combines rewritten content with conversation history to ensure the continuity of multi-turn conversations. The Recording Agent in [65] summarizes key findings into comprehensive differential diagnoses and highlights learning points in collaborative discussions. The Supervision Agent [11] monitors overall conversation quality and information completeness and suggests further data collection when necessary. The Polishing Agent in [135] improves the synthesized conversation for logical consistency, ensuring it follows the correct medical reasoning framework.

2.4 AI-Assisted Research Agents

In AI hospitals, AI-assisted research agents optimize new knowledge discovery, research support, and scientific review. Below, we summarize the key roles of research planning agent, research executor agent, scientific critic agent, and database agents [4]. **Research Planning Agent** plays a crucial role in structuring research tasks and ensuring efficient problem decomposition in complex domains. The Principal Investigator Agent in virtual laboratories [122] strategically guides research projects by optimizing scientific impact, ensuring systematic exploration of research directions. Similarly, CellAgent [148] introduces a hierarchical decision-making framework to decompose scRNA-seq data analysis tasks, allowing the planner to intelligently identify key preprocessing and analytical steps based on user input. Unlike general-purpose planners, the LLM Planner in DrugAgent [91] focuses on optimizing the "idea space" for machine learning tasks, iteratively refining potential solutions by evaluating feasibility and performance. By dynamically adjusting ideas based on experimental results, the LLM Planner minimizes redundant failures and enhances system-wide efficiency. **Research Executor Agent** facilitates clinical research by assisting in hypothesis testing, statistical analysis, and experiment interpretation. For example, researcher agents in [141] reduce programming overhead while optimizing research workflows by iteratively self-refinement. [122] designed Scientist Agents with domain-specific expertise (e.g., biologists, computer scientists, etc.) to facilitate clinical research and system evalua-

[4] For further details, see Appendix Table 4.

tion, ensuring that AI-driven insights are aligned with scientific rigor. CellAgent's Executors [148] are designed for systematically executing decomposed tasks in scRNA-seq analysis, ensuring minimal execution failures through adaptive learning mechanisms. DrugAgent's LLM Instructor [91] employs a structured methodology for problem decomposition, knowledge requirement identification, tool construction and reuse, mitigating common domain-specific knowledge gaps that often lead to errors in ML-driven drug discovery. **Scientific Critic Agent** is responsible for assessing the quality and validity of AI-generated solutions, ensuring reliable decision-making in research and clinical settings. For example, the Evaluator in CellAgent [148] functions as an expert reviewer, verifying the effectiveness of various methodologies and optimizing results through self-iterative refinements. **Database Agent** is designed to retrieve, manage, and integrate medical information for improved decision-making. For example, HospInfoAssistant in [11] facilitates interactions with Hospital Information Systems (HIS), efficiently handling patient records and administrative queries. EHRAgent [115] identifies and retrieves necessary medical records and structured data to answer clinical questions effectively.

3 Core Components: Interaction Patterns

AI Hospital employs different interaction patterns to enhance efficiency, reliability, and decision-making. These include task-focused collaboration, expert-guided decision-making, iterative problem optimization, automated knowledge integration, role-based coordination, and interactive debate. Each simplifies workflow and helps different agents move toward a common goal.

Task-Focused Collaboration in AI hospital systems is achieved by decomposing complex medical tasks into specialized sub-tasks, ensuring efficiency and consistency through structured execution. Many works have adopted this paradigm by designing modular architectures where different components focus on specific functionalities. For example, [85] leverages an ERRG workflow (e.g., Extract, Retrieve, Rewrite, Generate), which decomposes response generation into extraction, retrieval, rewriting, and final generation, ensuring precise and context-aware interactions. AIPatient [164] incorporating multiple agents define distinct roles, such as retrieval, abstraction, query generation, and response rewriting, ensuring that each agent specializes in a particular sub-task while maintaining an integrated workflow to roleplay different patients better. Similarly, [88] follows a similar strategy for patient simulation, with dedicated components handling state tracking, memory retention, and response generation to ensure effective multi-turn interactions. ClinicalAgent [165] decomposes complex diagnostic problems into smaller, solvable sub-questions and assigns them to specialized reasoning agents, ensuring systematic decision-making. Similarly, ArgMed-Agents [55] allocate clinical reasoning tasks across three key agents—generators, verifiers, and reasoners—allowing systematic argument generation and evaluation. In addition, EHRAgent [115] decomposes complex tabular reasoning in EHR workflows into different stages, including medical data integration, long-term memory-based example optimization, interactive feedback processing, and debugging. Each stage is assigned to a dedicated module for structured execution to improve the efficiency and consistency of task processing [5].

Expert-Guided Decision-Making in the AI hospital framework ensures that AI-driven medical decisions are reliable and clinically useful. Several studies have explored different approaches to integrating expert supervision into decision-making processes. For example, [93] leverages experts' feedback to refine LLM-patients-generated responses, where experts provide natural language critiques that are then transformed into behavioral rules guiding future responses. Several works [37, 28, 68, 124] collectively emphasize the role of experts in AI-driven medical decision-making, ensuring that complex cases receive domain-specific expertise and consensus-driven evaluation. Similarly, ColaCare [142] framework enhances decision reliability by linking AI doctors to specialized expert models, with a higher-level MetaAgent synthesizing and validating judgments. Additionally, expert oversight is crucial in mitigating diagnostic biases, as senior physicians provide corrective guidance to refine junior doctors' assessments [65]. Furthermore, expert involvement is not limited to direct patient care but also extends to research-oriented AI hospital applications, where principal investigators (PIs) and scientific reviewers proxy in structured discussions that mirror the real-world peer review process to improve scien-

[5] For further details, see Appendix Table 7.

tific conclusions [122]. Finally, AI-driven medical education benefits from expert-guided decision-making, where virtual medical specialists assess and refine medical students' diagnostic reasoning by providing feedback, corrective insights, and domain knowledge summaries [143] [6].

Iterative Problem Optimization (IPO) continuously refines problem-solving strategies through multi-step feedback loops. For example, [164] uses feedback-driven loops to refine queries dynamically based on initial retrieval consistency, ensuring robust and accurate results through multiple rounds of rewriting. [37] enhances physician agents' interactions by extracting previous high-quality conversational trajectories, allowing iterative refinement of questioning techniques for more precise and professional diagnoses. Similarly, nurse agents in a medical reception scenario [11] perform a three-step process of "action decision, reflection with feedback, response generation" in each round of conversation, that is, after the initial decision, they make corrections based on the feedback of the supervisory agent, thereby continuously optimizing the final response. Moreover, doctor agents in [128] refine diagnostic interactions through iterative self-play, using criticism-based feedback loops to enhance response accuracy. Furthermore, many collaborative frameworks [55, 124, 142, 40] allow their AI agents to criticize and modify each other's reasoning processes iteratively, leveraging self-consistency and confidence-based adjustments to reach refined conclusions. Finally, in many scenarios related to programming implementation [141, 115], feedback loops are often used to drive iterative code refinement, gradually improving accuracy through multiple self-assessment cycles [7].

Automated Knowledge Integration (AKI) plays a crucial role in AI hospital systems by seamlessly merging diverse sources of medical knowledge and patient-specific data to facilitate accurate, context-aware decision-making. For example, knowledge-enhanced retrieval agents in [114] integrate externally sourced medical knowledge with internally synthesized agent outputs, enriching decision-making with a combination of learned and retrieved information. Some methods involve memory-based knowledge integration, where patient information in [88] is segmented into short-term, long-term, and working memory. These structured memories are dynamically fused to generate responses tailored to evolving clinical interactions, ensuring contextual coherence. Similarly, the attention and trajectory Libraries in [37] automatically collect and verify medical dialogue information, allowing agents to share and utilize filtered knowledge. Here, the directed acyclic graph (DAG) is used to prevent information redundancy and enhance agent collaboration in multi-disciplinary decision-making. In addition, multi-modal integration is another key component of AKI. [142] combines structured data with unstructured clinical text in EHRs. Drhouse [157] leverages multi-source data, such as sensor inputs and expert knowledge, to refine diagnostic reasoning and generate structured outputs that align with medical best practices. ClinicalLab [156] merges diverse forms of medical evidence (e.g., lab results, imaging information, and expert analyses) into holistic diagnostic and treatment plans. Finally, many AI hospital systems employ team-based models that integrate diagnostic insights from multiple agents through a structured decision-making process, ensuring a comprehensive and rigorous evaluation [28, 68, 122, 166]. Several key approaches include adaptive probability distribution fusion (APDF) techniques, which reconcile diagnostic discrepancies among experts to generate consensus-based recommendations [132]; confidence reports for validation [94]; structured reasoning and symbolic argumentation methods to enhance diagnostic transparency and interpretability [55]; and recording mechanisms [124, 65] for summarizing expert analyses into comprehensive reports [8].

Role-based Coordination in multi-agent AI hospitals involves AI agents assigned specific roles in medical training, diagnosis, and patient interactions. These roles simulate healthcare settings, including patient consultations, multi-disciplinary team (MDT) collaborations, and structured decision-making, enhancing medical simulations' realism and effectiveness. For instance, AI-simulated patients in [37] present symptoms and engage with multiple AI-simulated physicians to assess diagnostic reasoning and decision-making. Similarly, the interaction between thera-

[6] For further details, see Appendix Table 8.

[7] For further details, see Appendix Table 9.

[8] For further details, see Appendix Table 10.

pists and clients in psychological counseling simulation [133, 107] allows AI-generated treatment to be aligned with clinical best practices. In addition, MDTs in real-world healthcare involve specialists collaborating for optimal treatment plans. Therefore, AI-driven frameworks replicate this by assigning different agents to distinct roles, such as PCP, radiologists, and other specialists. For example, different physician agents provide diagnostic insights, which are synthesized by higher-level agents (e.g., MetaAgent) to form a comprehensive diagnosis [142, 28, 80]. Another good example is AgentClinic [113], which distributes tasks between agents that simulate patients, physicians, and technicians, thereby improving the realism of medical training and AI evaluation. Finally, some other works enhance AI engagement in clinical encounters, benefiting decision-making and patient interactions beyond diagnosis. For example, virtual hospital framework [81] includes AI agents as triage nurses, receptionists, and follow-up providers, ensuring role-based interaction models cover the entire patient journey [9].

Multi-Round Interactive Debate enables agents to reconcile differing viewpoints and reach well-supported consensus decisions. For example, in some collaborative diagnostic settings, multiple physician agents often engage in structured discussions under the guidance of the central coordinating agent [40], supervisory agent [80], or moderator agent [68] to critically evaluate and discuss each other's diagnostic reports, identify key points of disagreement, and facilitate targeted discussions, ensuring that conflicting viewpoints are thoroughly reviewed before consensus is reached. Similarly, [124] integrates voting mechanisms where agents express approval or disapproval of a proposed summary report, iteratively refining it through multiple cycles of critique and revision until all participants reach an agreement. In addition, [117] adopts diverse debate strategies (e.g., Society of Minds, Multi-Persona, ChatEval) and utilizes multiple agents to refine their positions and reach a final consensus iteratively. Another research [94] employs collaborative discussion where agents propose classification results, reasoning, and confidence scores, iteratively refining their judgments through structured deliberation. Inspired by Byzantine consensus principles, this approach incorporates early stopping mechanisms to conclude discussions when all agents reach a high-confidence agreement efficiently. Finally, the multi-agent debate has also been explored in research team meetings [122], where research agents sequentially present their perspectives based on predefined agenda questions, while a designated critic agent evaluates and challenges these responses before a principal investigator agent synthesizes the discussion into a final decision [10].

[9] For further details, see Appendix Table 11.

4 Core Components: Tool Integration

In AI hospitals, agents use diverse tools to enhance efficiency and accuracy. For example, **Retrieval systems** ensure rapid access to medical knowledge by dynamically retrieving patient records and evidence-based guidelines, aiding both patient and physician agents in contextual reasoning [37, 68, 142]. **Knowledge graphs** structure medical knowledge into interconnected networks, enabling AI systems to navigate relationships between symptoms, treatments, and medical histories for informed decision support [85, 164, 28]. **Medical decision trees** provide structured diagnostic pathways, ensuring AI-driven recommendations align with established clinical guidelines and expert knowledge [157, 77]. **LLM-as-KB** transforms LLMs into dynamic knowledge repositories, allowing AI to synthesize medical insights beyond static databases [165, 42]. **Smart devices and sensor data** integration facilitate real-time health monitoring, merging wearable data with EHR insights to enhance predictive analytics and personalized care [157, 1]. **Multi-modality processing tools** enable AI hospitals to integrate textual, visual, and sensor data, improving tasks such as radiology interpretation and decision tree-based diagnostics [85, 157, 78]. **Computational reasoning tools** equip AI with logical inference and code execution capabilities, supporting automated clinical research and data-driven modeling [141, 55]. Finally, some **other clinical decision support tools** optimize diagnostic accuracy by leveraging external APIs, existing predictive models/systems, and structured reporting systems [28, 132, 78]. And some **other biomedical research tools** accelerate drug discovery and genomic analysis, enabling AI-powered advancements in computational biology and molecular medicine [122, 62, 91] [11].

[10] For further details, see Appendix Table 12.

[11] For further details, see Appendix Table 13 and 14.

5 Core Components: Memory Management

AI Hospital leverages structured memory management for adaptive learning and decision-making. Long-Term Memory retains knowledge across sessions, integrating internal model updates and external databases for enhanced reasoning. Short-Term Memory maintains contextual continuity within interactions, refining responses dynamically. Multi-Agent Shared Working Memory synchronizes knowledge across agents, enabling iterative optimization and collaborative decision-making [12].

5.1 Long-Term Memory (LTM)

Internal Memory embedded in the model parameters serves as a foundational knowledge repository for the agent to support zero-shot and few-shot tasks. For example, [85] leverages the inherent common-sense knowledge within LLMs to supplement missing information in clinical case graphs, ensuring the generation of plausible attributes based on pre-existing knowledge. [140] integrates internal memory by fine-tuning ChatGPT with real patient clinical records, resulting in more accurate adverse event and drug predictions.

External Memory is stored in databases, knowledge graphs, retrieval systems, and external tools to dynamically supplement and expand the agents' internal memory. It integrates two functions: **Static Storage** serves as a long-term, cross-session repository for stable knowledge, ensuring continuity in decision-making processes through databases and structured knowledge graphs. For instance, disease-specific expert agents integrate professional knowledge from NIH, enriching their domain expertise with static external storage [132]. [136] uses a cognitive conceptualization diagram (CCD) to structure and store patient-related beliefs and historical data, ensuring continuity in simulated patient interactions. Similarly, structured ESI manuals in [94] provide stable domain references, ensuring uniform decision-making across different agents. Additionally, expert medical knowledge databases built from public medical dialogues, textbooks, and diagnostic guidelines, along with some pre-indexed medical databases and external knowledge retrievers that access resources like Wikipedia, are typical examples of static storage in AI hospitals for retrieving stable and structured knowledge [157, 114, 166]. Similarly, Drug databases, knowledge graphs, and clinical trial registries are also used by many works [28, 165, 91] as external repositories for evaluating drug efficacy and safety, reinforcing static storage's role in evidence-based decision-making. **Dynamic Updating** enables AI hospital systems to integrate real-time information through external tools such as retrieval systems, sensors, and APIs, allowing models to adapt to changing tasks and environments. For example, [93] employs an interactive pipeline where expert feedback continuously refines behavior guidelines, dynamically updating AI patient responses based on human input. DrHouse [157] continuously syncs the latest clinical guidelines, ensuring medical recommendations remain up-to-date. Similarly, [141] dynamically updates information from sources like PubMed, GitHub, and Wikipedia, ensuring the latest evidence informs AI-generated content. In addition, several works [115, 113, 11, 117, 37, 81] leveraging long-term memory optimize few-shot prompting by dynamically storing and retrieving past successful cases to refine task execution. Notably, LTM stores sometimes accessed knowledge transferred from Short-Term Memory (in § 5.2) and retains user-specific information across multiple sessions, such as recurring health preferences and medical history. By maintaining long-term health records, LTM ensures consistency in AI-generated responses over time. For example, if a user prefers concise medical advice, this preference is stored and automatically influences future interactions.

5.2 Short-Term Memory (STM)

STM temporarily retains relevant details within a given task, ensuring continuity in interactions and reasoning while automatically clearing stored information once the task concludes. Several works related to medical conversation systems [92, 85, 164, 37, 133, 88, 63, 157, 27, 84, 128, 169, 55] leverage STM to maintain context and coherence across multi-turn exchanges. Some approaches integrate dialogue history or extract key entities to align with structured case graphs, ensuring context-aware responses dynamically. Others employ immediate and summary memory mechanisms to regulate context visibility, mitigate LLM forgetfulness with reminder phrases, or track reasoning steps and patient records for consistent and accurate decision-making throughout consultations. Similarly, Other works on multi-agent medical diag-

[12] For further details, see Appendix Table 15.

nostic systems [68, 124, 65, 11, 122, 117, 148] utilize dynamic context updating, shared task-specific memory, and local memory to track dialogue interactions, reasoning sequences, and subtask details, ensuring coherent multi-round discussions and collaborative decision-making.

5.3 Multi-Agent Shared Working Memory

Some works in the AI hospital incorporate Working Memory (WM), a hierarchical memory management structure designed to synchronize and share knowledge among multiple agents, facilitating collaborative decision-making. It is important to clarify the relationship between WM and STM in the AI hospital. These two types of memory often coexist and interact, facilitating information conversion between them. However, their key distinction lies in their scope and function: WM serves as a specialized, external temporary memory independent of individual agents. It retains intermediate results, integrates feedback, and enables structured reasoning across multiple steps. STM, on the other hand, is an internal temporary memory unique to each agent, primarily used for processing information within the agent itself. While both memory types contribute to multi-agent coordination, WM plays a more prominent role in global information integration and structured decision-making across agents.

In the AI hospital, some works employ a dynamic inference buffer to retain key data points to support continuous decision-making. For example, [94] integrates classification results, confidence scores, and reasoning from multiple agents into a summary report (Repo), which is iteratively referenced and updated across discussions to ensure seamless information flow. Another system [55] employs an argumentation framework where a reasoning agent records and structures the entire iterative process, preserving logical steps for symbolic inference. Additionally, a decision-making system [166] allows each agent to access prior execution traces as contextual input, ensuring continuity in multi-step decision-making.

For some other works, WM also retains intermediate results in multi-stage reasoning tasks to assist in complex decision execution. In [68], past dialogue records are accessed at each new iteration, preserving previous decisions to inform future steps. [148] maintains a global memory that stores finalized code snippets across sub-tasks, allowing shared retrieval among multiple expert agents, thereby reducing redundancy and improving efficiency in multi-agent collaboration.

Finally, feedback integration is another essential function of WM, allowing iterative optimization of learning and decision-making. In a medical team discussion system [68], a moderator agent reviews and synthesizes discussion content, generating feedback for each agent to refine decision-making. Expert agents in [124] participate in iterative voting and revision, ensuring consensus is reached through stepwise report updates. In collaborative consultations [142], a meta-doctor agent consolidates feedback from individual doctor agents to determine if further discussion is required. Another approach introduces a recorder agent that extracts key learning points from discussions, compiling a differential diagnosis list that evolves with each interaction [65]. Scientific team meetings follow a structured criticism and improvement cycle, where feedback from critique agents leads to continuous refinement of answers until an optimized response is produced [122]. A planning system records tool construction failures and dynamically adjusts future idea generation based on past failures, ensuring iterative improvement through working memory feedback loops [91].

6 Core Components: Reasoning Mechanisms

Different reasoning mechanisms in the AI hospital enhance decision-making. Direct reasoning follows a logical sequence, with Single-path reasoning ensuring stepwise inference and Multi-path reasoning exploring multiple trajectories for comprehensive conclusions. Feedback-based reasoning refines decisions dynamically via either External Feedback from experts or self-feedback, where AI iteratively evaluates and corrects its logic [13].

6.1 Direct Reasoning

Direct reasoning is a mechanism that relies on a sequential and explicit logical chain to derive conclusions step by step from the problem to the answer. This approach emphasizes a continuous logical process from the starting to the end without relying on external feedback or self-feedback.

Single-path Reasoning follows a sequential, structured approach, where each inference step builds upon the previous one to arrive at a conclusion in a linear manner. Several works have

[13] For further details, see Appendix Table 16.

demonstrated the effectiveness of this approach in different contexts. For example, ERRG [85] follows a structured pipeline of extracting entities, retrieving relevant subgraphs, rewriting information into natural language, and generating responses. [136] constructs cognitive conceptualization maps (CCDs) through stepwise information extraction, integrating psychological reasoning to simulate patient communication patterns. ClientCAST [133] builds psychological profiles based on structured records and follows a predefined sequence for interaction simulation and post-interaction evaluation, ensuring a fixed procedural flow. Many works related to diagnostic reasoning and medical consultations follow a structured approach, as seen in MDAgents [68], which handle low-complexity cases with a simple stepwise few-shot prompting method, and in expert systems and clinical diagnostic frameworks that emphasize single-path reasoning by breaking down decision-making into stepwise sub-tasks such as preliminary assessment, differential diagnosis, and treatment planning [156]. In addition, several works explicitly adopt CoT reasoning, following a linear inference process, including structured prompts for reasoning through diagnostic cases [113], AI-driven nurse simulators operating based on predefined clinical interaction sequences [11], coding generation agents leveraging CoT prompting for structured decision-making [141], least-to-most reasoning that structures problem-solving from simpler subcomponents to more complex ones [165], and Chain-of-Diagnosis, which sequentially assesses and gathers symptom-disease information to reach a diagnosis [24].

Multi-path Reasoning differs from single-path reasoning by allowing multiple parallel inference trajectories to unfold simultaneously. Instead of committing to a single logical chain, this approach explores different reasoning paths in parallel and integrates the results to derive a more comprehensive and flexible conclusion. Several works leverage multi-path reasoning to improve diagnostic accuracy and decision-making by enabling multiple specialists or agents to contribute independently before reaching a consensus. For instance, EvoPatient [37], RareAgents [28], MDAgents [68], and MedAgents [124] employ directed acyclic graphs, multidisciplinary teams, or iterative discussions to integrate diverse perspectives, while other approaches such as multi-agent collaboration [142], diagnostic interaction models [40], and expert systems leveraging self-consistency [84] ensure robustness by aggregating independently generated reasoning trajectories. Other works related to probabilistic and symbolic decision-making have also adopted similar strategies [132, 55]. In addition, LLM planners in [91] generate multiple candidate ideas in parallel before selecting the best one through experimental validation. Finally, parallel medical research meeting simulations in [122] further illustrate this mechanism, where multiple identical discussions are conducted independently, and their results are later synthesized into a single optimal response.

6.2 Feedback-Based Reasoning

Feedback-based reasoning is a mechanism that dynamically adjusts the reasoning process by continuously incorporating feedback information to refine and improve the thought path. The agent can refine its initial answers by continuously integrating new feedback, ensuring the reasoning results are more accurate and better suited to real-world situations.

External Feedback enhances reasoning by integrating real-time data, expert input, and structured external resources, allowing AI systems to adjust reasoning dynamically for improved accuracy and relevance. A typical scenario is that when information is incomplete, agents interact to gather additional details until the requirements are met. For example, when diagnostic information is insufficient, the doctor agent must ask the patient for new symptoms. In works like MediQ [84], CoD [24], and CRAFT-MD [63], dynamic medical consultation systems progressively refine their understanding through patient interactions, where each round of questioning integrates new responses to update the doctor agent's reasoning for diagnosis. Similarly, agents can collect information from tools as evidence to support their decision-making. RareAgents enables doctor agents to query external diagnostic and treatment tools, such as Phenomizer and DrugBank, integrating real-time clinical knowledge into the decision-making process [28]. The interactive programming platform in [141] incorporates expert & system environment feedback and external coding repositories, enabling the data scientist agents to adapt their coding outputs based on structured human guidance and research databases.

Self Feedback enables AI agents to iteratively refine their own rationale and outputs by internally

evaluating their logic, identifying inconsistencies, and making necessary corrections. Some works leverage self-feedback to improve the authenticity of role play. The principle adherence module in [93] ensures responses meet predefined standards by generating a checklist of "yes/no" evaluation questions and iteratively rewriting answers until all criteria are satisfied. Similarly, AIPatient incorporates an internal validation agent that assesses retrieved information for consistency and, if necessary, rewrites queries and regenerates responses through multiple iterations [164]. In addition, self-feedback plays a crucial role in automated diagnosis and clinical decision-making, where multi-agent frameworks iteratively refine assessments through internal critique and supervision, ensuring more reliable outcomes [68, 124, 142, 65]. Reflection-based reasoning techniques, such as Reflection CoT and self-play mechanisms, enable AI models to evaluate and refine their outputs through structured error analysis and collaborative discussions [113, 94, 11, 81, 40, 128, 80]. Other applications include code generation [141], drug discovery [91], medical research [122], and medical exam question generation [161].

7 Applications: Simulating Specific Scenarios

7.1 Clinical Workflow Simulation

Clinical Workflow Simulation has leveraged multi-agent systems to recreate comprehensive medical scenarios, spanning the entire continuum of patient care.

One strand of research focuses on simulating the complete medical consultation and diagnosis workflow, typically with different patient agents and one or more doctor agents, as well as some other agents evaluating and providing feedback, to assess how different consultation strategies affect diagnostic accuracy. For example, [92] introduced a multi-agent AI hospital environment that simulates the environment from "consultation" to "diagnosis." By combining real doctor-patient interaction strategies with synthetic medical records, the researchers trained a highly realistic patient simulator that can exhibit human-like behaviors such as emotional responses and proactive questioning. Their study divided the consultation process into four key stages: chief complaint inquiry, symptom refinement, related symptom query, and medical history collection. A noteworthy finding highlights

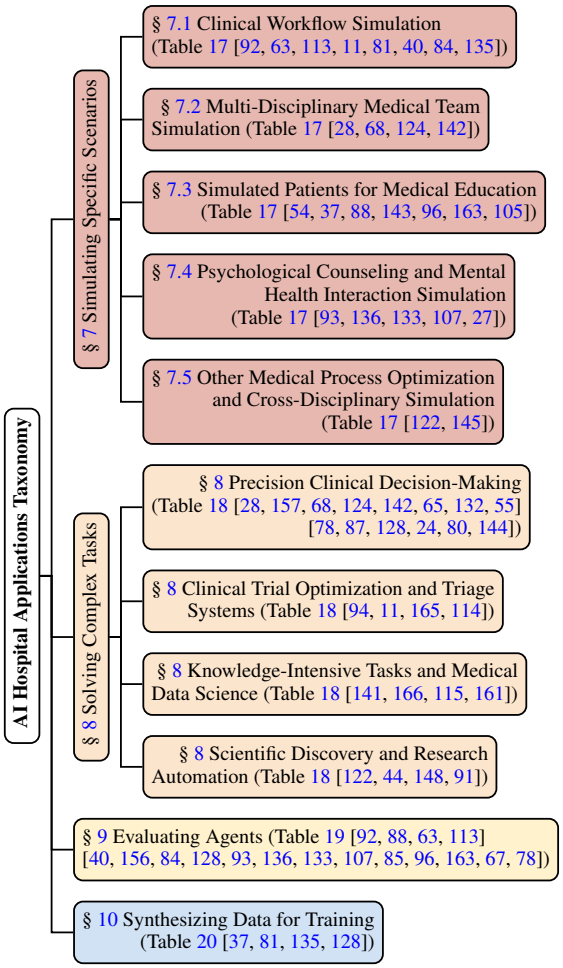

Figure 3: Taxonomy of AI hospital applications.

a phenomenon similar to Liebig's law, that is, the weakest stage in the consultation process will ultimately limit the accuracy of the diagnosis. [63] proposed the CRAFT-MD framework, in which the clinical LLM acts as a doctor and interacts with a patient-AI agent built according to structured medical cases. The framework can realize realistic consultation and medical history collection simulations, in which the grader-AI agent can automatically evaluate the diagnostic accuracy. In addition, the study used dialogue summaries to discuss the ability of the agent to integrate information and reasoning in complex clinical processes. [84] introduces the MEDIQ framework, which simulates multiple rounds of doctor-patient interactions to iteratively collect missing information before a confident diagnosis can be reached. A key innovation of this study is the integration of multiple abstention strategies—including numerical scores, binary decision, and scaling ratings—along with rationale generation and self-consistency mechanisms. These enhancements enable the expert system to decide when further questioning is needed

rather than making a diagnosis prematurely. Similarly, [40] proposed AI Hospital, a multi-agent framework replicating real-world clinical workflows. Their system includes player-controlled doctor agents and non-player characters (NPCs) representing patients, examiners, and attending physicians, enabling dynamic multi-round interactions. In addition, they discussed some collaborative dispute resolution mechanisms, where multiple doctor agents iteratively discuss cases while a Central Agent mediates disagreements to improve diagnostic accuracy by effectively simulating real-world collaborative diagnosis. Finally, [113] introduced AgentClinic, a multimodal agent benchmark featuring patient, doctor, measurement, and mediator agents to simulate the entire clinical workflow, from initial consultation to final diagnosis. A distinguishing feature of AgentClinic is its ability to model cognitive and implicit biases, challenging AI models to navigate scenarios with incomplete information and subjective decision-making. Finally, [113] introduces AgentClinic, a multimodal agent benchmark featuring Patient, Measurement, Doctor, and Moderator agents, simulating the entire clinical diagnosis workflow. A notable feature of AgentClinic is its ability to model cognitive and implicit biases, challenging different LLM-based doctor agents to navigate scenarios with incomplete information and subjective decision-making.

Another research direction is to unlock more scenario simulations throughout the patient journey. For example, [11] developed PIORS (Personalized Intelligent Outpatient Reception System), which includes a patient, a nurse, an information assistant, and the clinical doctor, and is integrated with the hospital information system (HIS). A key innovation of PIORS is its SFMSS (Service Flow aware Medical Scenario Simulation) framework, which combines predefined action spaces, dynamic role-playing, and supervisory feedback mechanisms to generate high-fidelity simulated conversation data. This approach improves the accuracy of Department Recommendations and simplifies information collection, ensuring a seamless transition to diagnosis and treatment. Similarly, [81] built an AI Hospital named "Agent Hospital," which fully reproduced the entire clinical process of patients from Disease Onset, Triage, Registration, Consultation, Medical Examination, Diagnosis, Medicine Dispensary to Convalescence through multiple agents (including patients, doctors, and nurse agents). Its core method, MedAgent-Zero, makes full use of LLMs-driven autonomous agents, allowing doctor agents to achieve self-evolution by accumulating correct and incorrect medical cases in the process of continuous interaction with patient agents. Specifically, the system builds a medical case base and experience base, records the diagnosis and treatment decisions and their reasoning process during each visit, and provides references for subsequent similar cases through RAG, thereby significantly improving the accuracy of doctor agents in the future examination, diagnosis, and treatment tasks without any manual annotation of data, achieving leading performance on real-world medical evaluation datasets, and observing a scaling law similar to that of real-world doctors improving their capabilities by accumulating correct and incorrect medical cases.

7.2 Multi-Disciplinary Medical Team Simulation

Multi-Disciplinary medical Team (MDT) simulation aim to replicate real-world medical teams' collaborative processes, including diagnostic discussions, treatment planning, and cross-disciplinary information integration. These simulations not only focus on the synergistic application of specialized medical knowledge but also emphasize internal team communication, information sharing, and decision-making optimization to enhance efficiency and accuracy in complex clinical scenarios.

In the context of rare disease diagnosis and treatment, [28] introduced the RareAgents framework, which utilizes a multi-agent system to simulate real-world multidisciplinary collaboration. In this system, a patient agent first conveys symptom information, after which an attending physician agent dynamically assembles a MDT based on the patient's condition. The specialist agents engage in multiple rounds of interaction, sharing information while leveraging a dynamic long-term memory mechanism and medical toolkits (such as diagnostic databases and drug information systems) to achieve precise rare disease identification and personalized treatment recommendations. This method effectively replicates the collaborative decision-making processes within medical teams and enhances reasoning capabilities and diagnostic accuracy in complex clinical settings through robust memory structures and tool utilization. Additionally, [68] proposed the MDAgents framework, employing a hierarchical agent collaboration strategy, dynamically adjusting the decision-making team

size based on the complexity of medical inquiries. For low-complexity cases, a single general doctor agent handles the decision independently, while for medium- and high-complexity cases, an MDT or Integrated Care Team (ICT) is assembled by recruiter agent to simulate expert discussions, information exchange, and consensus formation. MDAgents also incorporates a moderator agent review mechanism and external medical knowledge bases to refine the decision-making process, ensuring optimal diagnosis and treatment recommendations after multiple rounds of interaction. This adaptive collaborative strategy accurately reflects the complexity and efficiency of cross-disciplinary teamwork in real-world medical environments. Similarly, [124] introduced the MEDAGENTS framework, which simulates the collaborative discussion process among MDT medical experts through expert role-playing. The framework follows a four-phase collaboration process: 1) Expert Recruitment – Automatically selects and assembles specialists from various fields based on case characteristics. 2) Independent Analysis – Each expert agent evaluates the patient's condition based on their specialized knowledge. 3) Collaborative Consultation – Experts engage in multiple rounds of discussion, exchanging insights and iteratively refining their conclusions. 4) Decision Making – The insights from all expert agents are synthesized to finalize the diagnosis and treatment plan. This multi-agent collaboration mechanism not only enhances agents' accuracy in medical reasoning tasks but also demonstrates their potential to simulate real-world MDT cooperation without requiring additional training.

In the context of EHR modeling and clinical data analysis, [142] proposed the ColaCare framework, which integrates MDT simulation to improve clinical data modeling. In this framework, DoctorAgent processes structured EHR data and expert models to generate an initial assessment, incorporating authoritative medical guidelines such as the Merck Manual of Diagnosis and Therapy through a retrieval module. Meanwhile, MetaAgent consolidates assessments from multiple DoctorAgents, iteratively refining the clinical decision report through collaborative discussions. ColaCare effectively replicates expert discussions and evidence integration seen in real-world MDT meetings and enhances predictive modeling for patient mortality risk by combining numerical predictions with textual reasoning through a multimodal fusion network.

7.3 Simulated Patients for Medical Education & Training

In medical education, simulated patients are crucial in providing a controlled, risk-free environment for training medical students and clinicians. These simulations enhance skills in patient communication, clinical reasoning, and diagnostic decision-making while mitigating real-world risks. Advances in LLM-agent-driven simulated patients have significantly improved the fidelity and interactivity of these training systems. For example, [37] introduced EvoPatient, a multi-agent evolutionary framework that simulates real-world diagnostic processes for medical training. EvoPatient models the diagnostic process into a series of phases (i.e., complaint generation, triage, interrogation, and conclusion), which are integrated into a simulated flow. EvoPatient features patient and doctor agents that engage in multi-turn conversations and use RAG to ensure that patient agents generate responses based on historical medical records and the Big Five personality traits. A key innovation here is its unsupervised co-evolution mechanism, which constructs attention and trajectory libraries to validate and refine conversation examples iteratively. Over time, patient agents evolve into standardized simulated patients, while doctor agents continuously improve their questioning strategies. Experimental results highlight significant improvements in alignment with training needs, response quality, and robustness, making EvoPatient an effective tool for adaptive medical education. Similarly, [143] proposed MEDCO, a multi-agent collaborative platform designed for medical education, consisting of patients, medical students, radiologists, and medical experts, each playing distinct roles. Its key innovations include: 1) Multi-stage design – Training is divided into initialization, learning, and practice phases, allowing students to progress from initial consultations to final diagnoses. 2) Cross-disciplinary collaboration—Simulated interactions include patient symptom expression, radiological image interpretations from radiologists, and expert feedback, fostering an interdisciplinary learning experience. 3) Multi-modal support – The system integrates text-based dialogue, medical imaging, and clinical reports, enhancing realism and diagnostic accuracy. 4) Memory and peer discussion mechanisms – A memory system helps students track and recall learned knowledge, while peer discussion modules encourage collaborative rea-

soning and multi-perspective thinking, simulating real-world clinical learning curves. By combining structured learning phases, interdisciplinary collaboration, and multimodal integration, MEDCO provides a highly interactive, safe, and realistic training environment, demonstrating how multi-agent systems can enhance medical communication and diagnostic skills. Finally, [96] proposed introducing the AI Structured Clinical Examination (AI-SCE) framework to improve the fidelity of medical examination simulation, which borrows multi-agent coordination technology similar to autonomous driving simulation to assist clinical training. AI-SCE emphasizes process transparency and explainable reasoning, as well as a tool-assisted workflow, rather than focusing solely on the final diagnostic outcome. Based on this idea, [163] developed an AI-SCE framework, MedQA-CS, inspired by USMLE Step 2 CS, a real-world Objective Structured Clinical Examination (OSCE). In this framework, MedStuLLM simulates medical students in history-taking and diagnosis, while MedExamLLM acts as an AI examiner, providing structured evaluations. This approach offers a more realistic and dynamic alternative to traditional multiple-choice question exams.

7.4 Psychological Counseling and Mental Health Interaction Simulation

In the field of psychological counseling and mental health interaction simulation, AI Hospital provides a realistic, interactive, and scalable environment for training mental health professionals, evaluating therapeutic interactions, and enhancing personalized mental health support. These systems simulate therapist-client interactions, simulate emotional dynamics, and support research on intervention strategies and treatment effects.

ClientCAST [133] follows a structured simulation process: 1) Psychological Profile Construction—Each simulated client is assigned a detailed psychological profile, including problem descriptions, symptoms, personality traits, and emotional characteristics, to ensure that the simulated client exhibits expected psychological behaviors. 2) Multi-Agent Interactive Simulation – Using LLMs, one agent assumes the role of the client while another plays the therapist's role. The client agent follows its predefined psychological profile, producing emotionally consistent and contextually appropriate responses in a dynamic therapy session. 3) Client-Centered Evaluation—After each session, the simulated client completes a standardized questionnaire assessing factors such as therapeutic alliance, emotional response, and perceived consultation quality. This evaluation distinguishes between high- and low-quality therapy interactions and highlights LLM-based therapists' strengths and weaknesses in empathetic responses and communication strategies. 4) Comparative Model Analysis – The study systematically compares multiple LLMs (e.g., Claude-3, GPT-3.5, LLaMA3-70B, and Mixtral 8×7B) to evaluate their consistency in patient simulation and therapeutic effectiveness. By constructing an AI-driven multi-agent counseling environment, ClientCAST provides a new perspective on LLM applications in mental health and introduces novel evaluation methods for AI-driven psychological interventions. Similarly, [107] builds on interactive role-playing simulations to reconstruct realistic psychotherapy conversations. This study employs two GPT-4 models—one acting as the therapist, the other as the client—to generate long-form, multi-turn dialogues (up to 50 rounds). The system utilizes a structured integrative therapy framework that divides the conversation into three phases: exploration, insight, and action. The study also includes automated evaluation methods to compare LLM-generated interactions with human therapist dialogues, assessing differences in therapeutic language quality, logical flow, and client engagement. This multi-agent approach addresses real-world psychotherapy's high costs, privacy risks, and scalability challenges while generating high-quality synthetic datasets for benchmarking AI-based mental health support systems. Furthermore, [27] developed an AI psychiatric consultation system featuring doctor and patient agents, where the doctor agent follows DSM-5 diagnostic criteria, and the patient agent exhibits natural language interaction, emotional fluctuations, and resistance behaviors, enhancing AI's applicability in mental health care.

Finally, the authenticity of the patient agent is a bottleneck in psychological counseling and mental health interaction simulation. Some work has explored this direction. [93] introduced Roleplay-doh, which converts expert feedback into behavioral rules, optimizing AI-simulated patient interactions to provide a more realistic training environment for novice therapists. [136] proposed PATIENT-Ψ, integrating Cognitive Behavioral Therapy (CBT) with diverse cognitive models, allowing AI patients to simulate various psychological states better and help trainees practice cognitive modeling and ther-

apeutic strategies.

7.5 Other Medical Process Optimization and Cross-Disciplinary Simulation

Various studies have explored AI-driven methodologies to enhance scenarios like research collaboration and public health modeling in the broader domain of medical process optimization and cross-disciplinary simulation. For example, to explore the cross-disciplinary research collaboration, [122] developed a multi-agent "Virtual Lab" platform, where LLM-powered agents assume specialized roles, such as principal investigator, biologist, computer scientist, and scientific critic. These agents interact, debate, and iterate through structured team meetings and private consultations, collectively designing novel nanobody treatments for SARS-CoV-2 variants. The system integrates advanced biomedical tools like ESM, AlphaFold-Multimer, and Rosetta, demonstrating how AI can facilitate interdisciplinary research and accelerate scientific discovery. By enforcing clear role assignments and domain expertise, this AI-human collaborative framework offers a scalable model for problem-solving across complex scientific domains and medical workflow innovation. In the realm of public health and epidemiological modeling, [145] proposed a multi-agent platform that merges generative AI (e.g., ChatGPT) with agent-based epidemic modeling (ABM). Unlike traditional ABM approaches, which rely on predefined behavioral rules, this system allows agents to make autonomous decisions using LLM-driven real-time reasoning. Each agent evaluates health status, personality traits, and public health data (e.g., daily case reports) to determine mobility and quarantine behaviors. The system successfully recreates pandemic dynamics, including infection waves, self-isolation patterns, and flattened epidemic curves, by tightly integrating GAI-driven agent decision-making with ABM disease propagation models. This approach highlights generative AI's potential in modeling large-scale, multi-disciplinary simulations, bridging epidemiology, behavioral science, and AI-based decision support.

8 Applications: Solving Complex Tasks

Precision Clinical Decision-Making AI-driven precision clinical decision-making in AI hospitals has seen significant advancements that enhance diagnostic accuracy, reduce biases, and improve interoperability.

Firstly, many works focus on improving the diagnosis of complex and rare diseases. For example, RareAgents [28] introduces an MDT framework where PCP and Specialist agents collaborate to diagnose rare diseases, integrating long-term memory and medical tools for more consistent and accurate recommendations. Similarly, [80] applies a multi-agent conversation framework to rare disease diagnosis, coordinating multiple "doctor agents" under supervisory control to ensure deeper diagnostic reasoning and more effective testing recommendations. Meanwhile, MDAgents [68] dynamically adjusts its decision-making framework based on case complexity, shifting from single-agent to MDT or ICT, effectively optimizing resource allocation in clinical settings. The AMSC framework [132] coordinates multiple domain-specific LLMs, integrating probabilistic distribution fusion to enhance diagnostic reliability while optimizing implicit symptom modeling. AMIE [128] refines decision-making through self-play simulations, optimizing LLM capabilities in diagnostic reasoning, patient interaction, and communication skills.

In some other complex situations, Dr-House [157] leverages sensor data and real-time literature retrieval to refine its decision-making, utilizing adaptive reasoning and dynamic knowledge updates to enhance diagnostic reliability. MEDAIDE [144] leverages RAG, intent recognition, and agent collaboration to enhance strategic reasoning and multi-agent coordination in complex medical decision-making. By integrating structured query rewriting and intent-driven agent activation, MEDAIDE improves LLM performance in handling composite clinical queries and facilitates more accurate and context-aware decision analysis. MMedAgent integrates multimodal LLMs with specialized clinical tools, dynamically selecting and applying relevant resources through instruction tuning, demonstrating superior accuracy in medical imaging and diagnostic tasks [78]. Finally, [87] implements a multi-agent AI system specifically for traumatic brain injury (TBI) rehabilitation, employing structured guideline retrieval and role-based AI agents to enhance reliability and empathy in patient interactions.

Other works focus on improving decision transparency quality and mitigating biases. For example, DiagnosisGPT introduces a Chain-of-Diagnosis (CoD) approach, decomposing clinical reasoning into specialized sub-modules to enhance de-

cision transparency and reduce uncertainty [24]. ArgMed-Agents employs a recursive argumentation approach where generator, validator, and reasoning agents iteratively refine clinical reasoning, improving interpretability and decision robustness [55]. MedAgents [124] enhances medical LLM reasoning through multi-agent collaboration, where domain-specific LLMs independently analyze cases, iteratively refine conclusions, and reach a consensus, reducing errors and biases. [65] demonstrates a multi-agent AI system that effectively identifies and corrects cognitive biases such as confirmation, anchoring, and premature closure biases, improving diagnostic accuracy in complex medical scenarios. Through structured agent interactions, the system enhances decision-making, supporting clinicians in making more accurate diagnoses.

Clinical Trial Optimization and Triage Systems Multi-agent AI frameworks that enhance efficiency, accuracy, and transparency have significantly benefited clinical trial optimization and triage systems. TRIAGEAGENT [94] employs a heterogeneous multi-agent approach, integrating RAG to incorporate authoritative guidelines such as the Emergency Severity Index (ESI) manual, ensuring structured decision-making in emergency triage. PIORS [11] optimizes outpatient triage through LLM-driven nurse and assistant agents, dynamically adjusting recommendations based on patient history and real-time hospital information system (HIS) data, improving workflow efficiency and patient experience. Similarly, ClinicalAgent [165] advances clinical trial optimization by leveraging multiple specialized agents, including efficacy and safety evaluators, enrollment predictors, and planning agents that apply ReAct and Least-to-Most reasoning to improve trial result prediction and participant recruitment. Finally, MAKA [114] enhances patient-trial matching by employing multi-agent collaboration to detect and compensate for LLM knowledge gaps, integrating external medical databases and internal retrieval mechanisms, and increasing trial eligibility predictions' accuracy.

Knowledge-Intensive Tasks and Medical Data Science Multi-agent AI hospitals are also crucial in advancing knowledge-intensive tasks and medical data science by integrating agents with structured workflows for clinical research and data analysis. One study [141] demonstrates how AI agents assist in clinical data science by automating Python and R code generation for medical data analysis, using CoT prompting and self-reflection to enhance accuracy and efficiency. The SMART framework [166] refines knowledge retrieval and fact-checking through dedicated agents for intent parsing, structured retrieval, fact extraction, and response generation, reducing hallucinations and improving clinical decision support. EHRAgent [115] enhances EHR analysis by employing interactive code generation and execution feedback, allowing AI to autonomously query multi-table datasets and optimize query accuracy with minimal human intervention. Finally, MCQG-SRefine [161] employs a multi-agent framework where retrieval, generation, critique, and correction agents collaborate iteratively to refine USMLE-style question generation. By integrating retrieval-augmented prompting, structured critique, and self-improving correction loops, this approach enhances the quality and difficulty of generated medical exam questions, demonstrating how a multi-agent AI team can be leveraged to automate complex educational content creation in medical domains.

Scientific Discovery and Research Automation AI hospitals also accelerate scientific discovery and automate biomedical research by leveraging multi-agent systems to handle complex interdisciplinary tasks. The Virtual Lab [122] introduces an AI-driven principal investigator (PI) agent leading domain-specific AI agents, enabling cross-disciplinary research such as SARS-CoV-2 nanobody design by integrating protein language models (ESM), protein structure prediction tools (AlphaFold-Multimer), and molecular simulation software (Rosetta) to generate and validate new antibodies. Similarly, the AI Scientist [44] employs multi-agent collaboration through LLMs, machine learning tools, and experimental platforms to automate hypothesis generation, experimental validation, and adaptive knowledge integration, significantly accelerating biomedical discovery and reducing research costs. Additionally, CellAgent [148] enhances single-cell RNA sequencing analysis by using specialized AI agents (planner, executor, evaluator) to automate data processing, optimize parameters, and iteratively assess quality, thereby improving reproducibility and reduce human intervention in bioinformatics research. Finally, DrugAgent [91] automates machine learning-driven drug discovery by dynamically selecting appropriate tools for data processing, model training, and

evaluation, enabling tasks such as ADMET prediction with minimal manual intervention, significantly streamlining pharmaceutical research and accelerating drug development workflows.

9 Applications: Evaluating Agents

Evaluating AI agents in clinical environments has evolved from static benchmarks to dynamic, multi-agent simulations, where real-time interactions and role-playing serve as crucial components of assessment frameworks.

Firstly, several studies have demonstrated that multi-agent environments provide a more comprehensive evaluation of AI models in clinical decision-making and patient engagement. For instance, CRAFT-MD [63], AgentClinic [113], and MEDIQ [84] assess not only diagnostic accuracy but also information-gathering strategies and patient engagement, ensuring that AI models are evaluated in a way that closely mirrors real-world clinical dynamics. Similarly, a major advancement in AI hospital agent evaluation lies in integrating automated interaction systems and dynamic state tracking, which enable AI-driven assessments to reflect the complexities of clinical workflows. The AIE framework [88] combined with the State-Aware Patient Simulator (SAPS) exemplifies this approach, where AI agents are assessed based on their ability to dynamically track patient status, adjust questioning strategies, and refine diagnoses through multi-turn conversations. By leveraging state tracking and memory mechanisms, SAPS ensures coherence in patient responses, making it a more reliable evaluation tool than traditional static medical knowledge benchmarks. This is also reflected in the evaluation framework of several studies about psychological counseling agents [93, 133, 133].

Additionally, many studies have evaluated how LLM agents develop key skills such as integrating medical expertise, communication, and collaborative decision-making in multi-agent settings. The AI Hospital framework [40] employs a dispute-resolution mechanism where multiple AI doctors engage in discussions, with a central agent synthesizing their opinions to reach a consensus. This enhances diagnostic reliability while assessing interdisciplinary communication and teamwork, which is critical in real-world hospital settings. Similarly, ClinicalLab [156] designs an agent-driven system where AI specialists contribute insights across medical domains, simulating cross-departmental collaboration. By mirroring multidisciplinary workflows, it serves as a testbed for evaluating AI coordination and medical knowledge exchange in complex decision-making environments.

Another critical dimension in AI hospital evaluation is multi-modal assessment and the evaluation frameworks for tools, For example, MMedAgent [78] integrates diagnostic reasoning with medical imaging tasks such as MRI, CT, and X-ray interpretation, providing a benchmark for assessing how well AI agents can bridge textual knowledge with visual clinical decision-making. Similarly, MEDCALC-BENCH [67] evaluates AI agents' ability to leverage assistive tools like medical calculators for quantitative clinical assessments.

Lastly, an important evaluation scenario in AI hospital research is the simulation of the Objective Structured Clinical Examination (OSCE), which medical students take to assess their clinical skills. Several studies, such as MedQA-CS [163], OSCE-Bot [105], and AI-SCE [96], replicate the OSCE framework by simulating comprehensive clinical skill assessments. These include information gathering, diagnostic reasoning, and patient interaction, providing a more nuanced evaluation of AI agents' capabilities than conventional multiple-choice (MCQ) tests.

10 Applications: Synthesizing Data for Training

Synthesizing Data for Training plays a vital role in AI hospitals, allowing researchers to circumvent privacy issues associated with real patient data while generating sufficiently diverse and clinically realistic training datasets to improve the clinical capabilities of AI systems. For example, EvoPatient [37] introduces a multi-agent co-evolution framework, where patient and doctor agents engage in multi-turn conversation to complete diagnostic processes. EvoPatient ensures contextual coherence and medical realism by structuring the simulation into distinct stages (e.g., chief complaint generation, triage, consultation, and diagnosis). Additionally, it incorporates multidisciplinary doctor agents to enhance the diversity of medical scenarios. The key innovation lies in its co-evolution mechanism: the Attention Library automatically filters and stores high-quality question-answer pairs, refining patient agents' ability to express medical conditions in a standardized manner. Meanwhile, the Trajectory Library records the interaction se-

quences between doctor and patient agents, capturing diagnostic reasoning patterns and optimizing patient agent responses. This automated data synthesis mechanism significantly reduces reliance on manual annotations and ensures data standardization and credibility, making it a robust foundation for training medical LLMs.

Similarly, [81] introduces MedAgent-Zero with two key components: patient agent generation and doctor agent evolution. Patient agents are generated using LLMs and medical knowledge bases, ensuring that their medical histories, symptoms, and diagnostic reports align with real-world medical standards. A quality control agent verifies the accuracy and consistency of the generated data. The doctor agent evolution process incorporates a Medical Case Base and an Experience Base to enhance diagnostic performance. The Medical Case Base records successful diagnostic cases, enabling retrieval-based improvements, while the Experience Base helps doctor agents learn from misdiagnoses by comparing errors with correct diagnoses and refining their decision-making rules. Experimental results show that MedAgent-Zero significantly improves diagnostic accuracy after training on thousands of synthetic patients and achieves state-of-the-art performance on real-world medical benchmarks such as MedQA. Moreover, the study explores the "Scaling Law in Evolution," demonstrating that as the number of processed patient cases increases, the diagnostic performance of doctor agents continues to improve. The alignment between virtual training and real-world medical tasks indicates that skills learned in synthetic environments can be generalized effectively to real clinical applications. The design of MedAgent-Zero enables doctor agents to grow progressively stronger without requiring actual patient data, highlighting the potential of synthetic data in medical LLMs training.

In addition, [135] presents NoteChat to generate high-quality doctor-patient dialogue synthetic data conditioned on clinical notes. NoteChat employs a three-stage collaborative generation approach to help overcome privacy limitations in real medical dialogue data collection: (1) The Planning Module utilizes external medical knowledge (e.g., MedSpaCy and QuickUMLS) to structure dialogue flows, ensuring completeness and coherence in medical conversations; (2) The Role-Playing Module assigns different LLMs to simulate doctors and patients, generating dialogues through multi-turn interactions that capture realistic doctor-patient communication styles; (3) The Polish Module integrates self-reflection and expert feedback to iteratively refine dialogue content iteratively, enhancing medical accuracy and logical consistency. Experimental results show NoteChat provides a high-quality synthetic medical conversation data source for training medical LLMs.

Finally, [128] introduces AMIE (Articulate Medical Intelligence Explorer), which utilizes a self-play simulation environment and automatic feedback mechanisms to improve doctor agents' ability to handle diverse medical conditions while enhancing their history-taking, diagnostic reasoning, communication skills, and empathy. By performing self-play-style training on a wide range of medical cases, AMIE helps overcome the coverage limitations and noise issues in traditional datasets.

11 Key Challenges and Future Directions

11.1 Agent Roles

In AI hospitals, different agents should exhibit behavioral patterns consistent with their designated roles to enhance the realism and practicality of medical simulations in different situations.

Some work has mentioned and tried to improve this in their scenarios, but discussion and evaluation of this in more scenarios is necessary and needs to be more unified. For example, Doctor agents should exhibit variations in diagnostic styles, communication methods, and decision-making processes, even when based on the same underlying model [68]. Patient agents must dynamically adjust their responses across different stages, ensuring that they gradually reveal medical history during consultations rather than disclosing everything at once [135]. In after-visit and patient education scenarios, they should provide responses that are comprehensible to laypeople rather than excessively technical explanations [51, 17, 160, 89, 71]. Subsequent work may consider better integrating STM/LTM/WM modules to maintain contextual coherence [169]. At the same time, recent advancements in role-playing [21] and personalization [21, 23, 170] methods in the general NLP domain, such as interview-driven persona modeling [103] and expert feedback-based refinements [93], can be leveraged to improve agent behavior.

Another key aspect is managing information asymmetry, a fundamental characteristic of real-

world medical conversations [6, 51]. Doctor agents seek comprehensive patient information, whereas patient agents may selectively withhold certain details due to privacy concerns or psychological barriers [47]. Modeling patient responses using hedging language can better reflect real-world uncertainty, and employing utility functions can capture how patients weigh different trade-offs, such as balancing disclosure of medical history versus preserving personal comfort [76]. Additionally, patients tend to avoid negative diagnoses and adjust responses based on perceived risk, behaving more conservatively when severe illnesses are a concern. These behavioral tendencies should be embedded into AI agents to enhance realism.

Inverse reinforcement learning (IRL) [18] is another promising approach for improving the decision-making of patient and doctor agents. Some work uses a small predefined action space to better control agents' behavior and facilitate optimization. However, since patients in the real world do not follow a predefined reward function, IRL can be used to infer their underlying decision-making processes and other unconsidered actions. This enables AI agents to learn patterns, such as when patients decide to seek medical attention, comply with prescribed treatments, or respond to doctor recommendations [119]. Training doctor and patient agents to align with observed human decision-making trajectories will significantly improve their realism in medical simulations and further improve the generalizability of these methods in the real world.

Finally, ensuring the diversity of patient agents is another key challenge, as homogeneous behaviors among agents can limit the robustness of evaluation and data synthesis [164, 9]. To address this issue, demographic attributes should be supplemented with other factors, such as social determinants of health (SDOH) [101], including housing stability, economic status, and educational background. Additionally, some studies have attempted to extract information from actual clinical notes to construct agent profiles or memories, which to some extent increases diversity. However, in the real world, a patient's information is much more extensive, whereas clinical notes only capture a small portion. This makes it more challenging to reconstruct a patient agent with sufficient informational depth based on the compressed representation in clinical notes. While some approaches have leveraged LLMs' commonsense reasoning capabilities and knowledge graphs to alleviate this problem, more in-depth exploration is needed to effectively reconstruct patient agents with sufficient informational depth based on clinical notes. These enhancements will enable the AI hospital to reflect diverse patient populations more accurately, thereby improving the generalizability and fairness of AI applications in healthcare.

11.2 Interaction Patterns

The interaction patterns within multi-agent AI hospital systems remain largely undefined, particularly regarding the roles, behaviors, and interactions of humans within these systems. Currently, most existing studies do not explore the scenario where humans are directly embedded in the system, but rather humans (whether experts or ordinary people) are just observers, evaluators, or provide some external feedback. A fundamental question is whether humans should only act as observers or actively participate as participants, replacing or supplementing certain AI agents. If participation is required, how can a unified framework to guide human identity and participation patterns in different scenarios be more conveniently and appropriately defined? In addition, integrating real human interactions into AI hospital systems could open new research directions, such as evaluating whether humans can accurately distinguish between AI agents and other human participants during collaboration. This approach aligns with the Turing test concept and may redefine how AI Hospital is assessed and applied in medical contexts.

Additionally, incorporating strategic decision-making and modeling uncertainty into the AI hospital can enhance system intelligence [10, 34, 64, 104, 57]. Current approaches rely on end-to-end LLM predictions without explicit mathematical modeling. By leveraging some methodologies like game theory [14, 120, 35, 36, 49], we can better model asymmetric information challenges in medical interactions [32]. For instance, doctor-patient interactions can be framed as zero-sum games (e.g., patients withholding symptoms to test diagnostic ability) or cooperative games (e.g., Nash Bargaining for optimized questioning). Multi-agent systems can employ Stackelberg games [45] to optimize information exchange, with doctor agents guiding patient agents toward informative disclosures. Evolutionary game theory [15] may enable agents to refine their strategies over time. Bayesian games may model medical uncertainty,

allowing doctor agents to use Bayesian inference to refine questioning strategies while patient agents adjust responses based on perceived health status [130, 19].

11.3 Tool Integration

In the current AI hospital, tools are often treated as static utilities. Most works directly integrate them without systematically evaluating and adapting their effectiveness within different scenarios [134]. A key challenge for the future is how to leverage the AI Hospital—an environment that closely resembles the real world—to better evaluate and validate new tools and determine whether they are truly effective rather than relying on traditional static benchmarks and flawed evaluation metrics. For example, while new tools such as domain-specific RAG [150, 152, 151, 167, 139, 87], GraphRAG [90, 146, 70, 95, 66, 46], and LLM-as-KB [42] always demonstrated their advantages over previous methods on certain benchmark datasets, it remains unclear whether these advantages translate effectively into AI hospital agents or real-world users. In a real clinical setting, the success of a tool is not solely measured by standard benchmark performance but also by its ability to support different agents in providing more reliable and interpretable assistance to both clinicians and patients. Therefore, a crucial future direction is establishing the AI Hospital as a more unified and robust evaluation framework that goes beyond traditional quantitative metrics and instead assesses tools based on their real-world impact on agent interactions and patient outcomes.

11.4 Memory Management

Existing research largely relies on static EHRs as memory to represent patients, often utilizing RAG or GraphRAG-based methods [164] to retrieve relevant background information to support patient agents to generate appropriate and factual responses. While this approach enables a certain level of personalization, it still faces significant challenges, particularly in comprehensively and dynamically representing a patient's longitudinal EHR and optimizing memory access mechanisms [149]. One million-dollar question is how to accurately represent a patient's long-term health information. Current methods often treat EHRs as static knowledge bases, overlooking the temporal dependencies in disease progression and medical decision-making. Future research can explore constructing temporal graphs [110] to encode a patient's medical history, medication usage, and visit records in a time-series format, allowing LLM agents to identify critical transition points in disease progression and adjust their interaction strategies accordingly [22]. For example, in chronic disease management, patient agents may not immediately adhere to a doctor's recommendations but instead undergo a habit-forming process where they gradually adjust their health behaviors. Therefore, the memory module must be able to model a patient's evolving adherence to long-term medical advice and dynamically adapt the way and frequency in which doctor agents provide information. Similarly, LLM agents can leverage habit formation models to simulate how long-term patients gradually adapt and modify their health behaviors [116, 168]. For instance, some patients may rely heavily on doctor agents for guidance in the early stages of a disease, but as they become more familiar with disease management, they may transition toward making more autonomous decisions.

Another critical issue is designing more sophisticated trigger mechanisms to optimize memory access and retrieval. A typical scenario is patient education [17], where even if a doctor agent provides relevant information, a patient may fail to comprehend it if the readability level does not align with their health literacy. As a result, the memory module must incorporate more fine-grained access control mechanisms. For instance, if a patient agent exhibits comprehension difficulties (e.g., asking repeated questions or giving incoherent responses), the system should automatically adjust how information is stored and retrieved in the future, ensuring that when information is recalled, it is presented in a manner that better matches the patient's health literacy level. This mechanism can be further refined by using reading comprehension difficulty models to shape how patient agents interpret and respond to doctor queries, making their behavior more aligned with that of individuals with low health literacy.

11.5 Reasoning Mechanisms

Most research is still limited to direct reasoning with a single path. However, this design struggles to generalize to the complex and dynamic real-world medical environment. Therefore, establishing a more adaptive reasoning framework that enables AI agents to make more reasonable

decisions in uncertain environments is a key direction for future research. Note that this does not mean that direct reasoning of a single path should be discarded; instead, it should be used only as part of the agent reasoning mechanism to handle appropriate scenarios. In recent years, significant progress has been made in clinical reasoning, particularly in some multi-hop reasoning medical QA benchmarks [155, 59, 41, 127, 56]. These methods exhibit stronger adaptability in simulating clinical reasoning, allowing LLMs to handle complex medical reasoning tasks more effectively. However, these methods usually require a lot of computation during training or testing and have not been proven to be more efficient and flexible for agents in dynamic environments. The future challenge is further integrating these reasoning capabilities into AI hospital agents.

A core issue is that medical AI agents must be able to handle uncertainty and base their actions on reasoning [10, 5]. For example, in an AI hospital system, a patient agent may change its mind during a conversation, while a doctor agent may lack complete information about the patient's health status. In such cases, AI must understand and infer "Why did the patient agent change their mind?" to adjust its decision-making process accordingly. This involves not only general knowledge reasoning but also uncertainty modeling to improve AI agents' judgment and reduce hallucinations. For example, to better model uncertainty in AI hospital systems, Bayesian Inference and Markov Decision Processes (MDP) offer promising approaches [12, 106]. Bayesian Networks enable AI to probabilistically reason over patient symptoms, history, and socioeconomic status, dynamically adjusting decisions via Bayesian updates. MDPs further support decision-making in dynamic interactions, optimizing actions based on state transitions and rewards. Given the inherent uncertainty in medical reasoning, Partially Observable MDPs (POMDPs) may provide a more realistic framework, allowing AI to infer missing patient information and adopt strategies like information gathering or abstaining from uncertain decisions.

11.6 Simulating Specific Scenario & Solving Complex Tasks

One of the primary challenges in AI Hospital applications lies in achieving more precise and comprehensive medical simulations, particularly in integrating time-sensitive and event-driven information. Currently, most simulations are confined to patient visits, with limited consideration of pre-visit preparations, and even fewer studies focusing on after-visit follow-ups or daily patient care. However, in real-world healthcare settings, many critical factors occur beyond the visit itself, such as chronic disease management, post-surgical recovery, and long-term health interventions. Additionally, public health events like COVID-19 impact hospital operations and patient behaviors, necessitating adaptive multi-agent AI systems [14]. However, current systems lack flexibility to model such disruptions, limiting realism [53]. For example, social cognitive theory may offer a framework in such context for simulating patient decision-making, as individuals often rely on social dynamics over medical advice [158, 4]. Integrating observational learning and social adaptation into AI agents can enhance patient behavior modeling, improving simulation fidelity and AI-driven health solutions.

Moreover, the robustness and reliability of AI Hospital remain major concerns. While multi-agent architectures showcase promising potential, they also introduce inherent challenges [13], such as LLMs hallucination generation [58, 173, 82], alignment issues, and limitations in long-text processing, which hinder their effectiveness in complex medical tasks. These problems are further exacerbated by the high frequency of interactions between agents, leading to computational bottlenecks and error accumulation, degrading the entire system's performance. For example, patient agents may incorrectly attribute their symptoms to severe illnesses (such as cancer) based on incomplete or incorrect information, while doctor agents may develop biases influenced by recent diagnostic cases [108]. If left unchecked, these biases can not only reduce the reliability of individual agents but also propagate errors throughout the system, amplifying their negative impact.

Finally, risk management in the AI Hospital is crucial [10]. Risks like the cumulative effect of error and the inability to handle long-tail cases or rare scenarios all underscore the importance of implementing safeguard mechanisms. For instance, in long-tail medical cases, the system may struggle to adapt effectively, leading to false positives or negatives, compromising diagnostic accuracy and wasting healthcare resources. To mitigate these risks,

[14] https://en.wikipedia.org/wiki/Impact_of_the_COVID-19_pandemic_on_hospitals

future work should integrate uncertainty quantification, allowing agents to trigger safety protocols when encountering ambiguous cases. Additionally, extreme scenario simulations should be employed to strengthen testing environments, ensuring system reliability under complex conditions. Designing error isolation mechanisms can prevent a single agent's mistake from cascading through the entire system. Finally, human expert intervention remains a critical safeguard, ensuring that AI-generated decisions align with ethical and medical standards through expert oversight and real-time monitoring.

11.7 Evaluating Agents

Compared to general-domain evaluation methods, the unique characteristics of the medical setting—such as the roles of doctors and patients and the complexity of tasks—make human evaluation particularly challenging [123]. As a result, most existing approaches still focus on task accuracy, traditional generation metrics, or naive LLM-as-Judge evaluation methods, with limited consideration of efficiency and cost factors. Future research should explore more effective evaluation methods that align more closely with real-world medical practice. For instance, in actual healthcare environments, doctors are typically assessed through patient feedback, peer reviews, and survey-based evaluations [8]. These social evaluation mechanisms have not yet been fully integrated into AI hospital system assessments [98]. Additionally, drawing inspiration from the Turing test [100], researchers could investigate systematic methods to measure the "intelligence" and "usability" of AI agents during medical interactions.

Another overlooked aspect is cost and efficiency. In the general NLP domain, Scaling Test Time Compute (TTC) has become a crucial factor in assessing system performance improvements [118]. However, in AI hospital research, little attention has been given to how computational resource consumption impacts the practical value of a system [40, 117]. Many AI hospital designs (e.g., Iterative Problem Optimization or Multi-Round Interactive Debate) achieve superior performance partially due to increased inference computational power rather than genuine intelligent collaboration. Therefore, future evaluation frameworks should consider how to standardize the cost of AI agents and establish reasonable value metrics. For example, an agent's computational resource demands, inference time, and performance gains could be factored into a weighted cost model to analyze the trade-offs between efficiency, cost, and performance across different strategies. Furthermore, in medical tasks, how different agents (e.g., expert-level AI vs. smaller-scale medical AI) collaborate to minimize costs—such as reducing reliance on high-cost models—remains an open question. One potential direction of exploration may be to simulate expert-medical student task delegation and collaboration. Here, experts are often more expensive in real-world tasks (such as medical annotation and evaluation), so strong LLMs that require more computational cost can be used, while corresponding medical students can use LLMs with weaker capabilities but more cost-effective. It is an interesting topic to study how to maintain high-quality results in tasks such as medical annotation and evaluation while reducing the reliance on strong LLMs (i.e., reducing the computational cost of the entire system).

Additionally, most AI hospital research predominantly relies on general-purpose LLMs such as GPT-4 [2] and LLaMA [38], with limited exploration of medical-specific LLMs like DoctorGLM [153], HuatuoGPT [25], BianQue [29], BioLLaMA [126], BioMistral [72], and Baichuan-M1 [131]. Some studies, such as MedQA-CS [163], have noted that while medical LLMs achieve higher exam scores, they often lose emergent abilities—which are crucial for agentic behavior in AI hospital settings. As a result, many approaches merely use these medical models as "tools" [42] rather than active agents. Future work should focus on preserving these agentic capabilities in medical LLMs, given their clear advantage in medical knowledge. Moreover, this challenge aligns with the previously mentioned evaluation metric deficiencies—new benchmarks beyond medical exams must be developed to assess these models comprehensively. Without such advancements, it will be difficult to ensure simultaneous progress in both medical knowledge and real-world medical problem-solving capabilities.

11.8 Synthesizing Data for Training

Efficiently synthesizing high-quality data for training in AI hospital systems remains a core challenge. Although existing studies, such as DeepSeek-r1 [33], have demonstrated that models can continuously improve through reinforcement learning (RL) [60] in specific environments without supervised data, AI hospitals, as complex medical envi-

ronments, have not yet been fully utilized as RL environments to support the training of medical LLMs and intelligent agents while providing high-quality synthetic data. In traditional RL frameworks, agents optimize their policies by interacting with the environment and receiving reward signals. However, in medical scenarios, the scarcity of real-world data and ethical constraints pose challenges in designing appropriate environments and reward mechanisms. AI hospitals offer a controlled simulation environment that can construct different types of feedback signals based on patient simulations, physician decision-making processes, and the success rate of medical tasks [81, 102, 109, 162, 97]. For example, the AI hospital can simulate different patient recovery processes in training a postoperative care assistant. The agent's decisions—such as adjusting care plans, recommending follow-ups, or modifying medication regimens—can receive rewards based on changes in the patient's virtual health status. If the agent's decision accelerates patient recovery (e.g., an improvement in the virtual patient's health score), it receives a positive reward; if it leads to adverse events (e.g., a decline in the health score or the occurrence of complications), it receives a negative reward. This interactive feedback mechanism not only reduces reliance on manually labeled datasets but also enables agents to learn optimal medical decision-making strategies through trial and error.

However, ensuring that AI hospitals generate sufficiently diverse and fair data remains a critical challenge. Current synthetic data mechanisms primarily rely on manually designed rules, making it difficult to accurately reflect the complexity of real-world medical scenarios [48]. For instance, existing datasets often lack simulations of postoperative care and other longitudinal medical tasks, as well as sufficiently rich medical annotations, limiting the adaptability and generalization capabilities of intelligent agents. Additionally, with the introduction of self-training techniques, if agents are continuously trained on self-generated data, the homogenization of data distribution could lead to mode collapse or extreme biases, ultimately degrading model performance in real-world applications [7]. Therefore, future research should focus on developing more dynamic data synthesis mechanisms, leveraging multi-agent collaboration to generate data that better reflects real-world medical scenarios. Additionally, integrating multimodal information—such as text, images, and speech—can enhance the expressiveness of these datasets [3]. Simultaneously, robust data evaluation and bias detection mechanisms must be established to ensure that synthetic data not only improves agent capabilities but also avoids reinforcing existing errors, safeguarding fairness and reliability [129, 112].

11.9 Governance, Ethics, and the Roles of AI Researchers and Medical Practitioners

The governance of AI hospital systems faces numerous technical and ethical challenges, particularly in terms of transparency and responsibility allocation [69, 111]. As these systems play an increasingly significant role in medical decision-making, defining accountability when intelligent agents make erroneous decisions has become a complex and urgent issue. Since medical AI systems may contain implicit biases, especially in resource allocation, such biases could exacerbate social inequality. One of the core issues is establishing a practical governance framework that ensures these systems uphold ethical standards, security, and privacy protection while continuously improving through interdisciplinary collaboration [16]. Furthermore, the lack of transparent version control and change tracking makes it difficult for researchers to monitor the evolution of these systems, hindering in-depth assessments of their real-world impact. To address these issues, a comprehensive governance mechanism is required. This includes introducing human oversight at critical decision-making stages, implementing transparent version management strategies, and promoting international cooperation to develop unified AI medical governance standards, ensuring responsible development.

From the perspective of AI researchers, the key challenge moving forward is how to fully leverage this simulation system to address the various technical challenges discussed above in § 11.1 to § 11.8. For medical practitioners, the challenge lies in truly integrating AI hospital systems into clinical practice while improving the fairness and accessibility of medical services without disrupting existing workflows. The introduction of AI should not become an additional burden for physicians but should be designed as a seamless integration into current workflows, allowing medical teams to utilize these systems naturally to enhance decision-making. At the same time, AI applications in clinical practice must be designed to assist rather than replace professional medical judgment. Therefore,

deep involvement from medical professionals in developing and testing these systems is crucial. By engaging doctors, nurses, and other healthcare personnel in the development process, AI hospitals can be designed to meet real clinical needs effectively and address practical challenges in medical practice, bridging the gap between AI research and real-world applications.

12 Conclusion

AI hospitals represent a transformative approach to AI-driven healthcare, leveraging multi-agent systems to enhance complex tasks such as clinical decision-making, automate scenario simulations, evaluate LLM-based AI agents, and generate synthetic data to facilitate AI agent evolution. This paper systematically categorizes AI hospital research in terms of core components, applications, existing challenges, and future directions, providing a structured framework for understanding its development.

13 Limitations and Societal Impacts

Despite our best efforts, some limitations remain. Due to space constraints, we can only provide a concise summary of each method rather than an exhaustive technical discussion. Even though we have included a more detailed discussion in the appendix, readers may still need to refer to the original papers and code repositories for full implementation details. The appendix lists key AI hospital resources that provide open-source code and data. Additionally, our literature review primarily focuses on studies from *ACL, NeurIPS, ICLR, ICML, AAAI, some medical journals, and select preprint servers (arXiv, medRxiv, bioRxiv). As a result, there is a possibility that we may have overlooked relevant work published in other venues. We acknowledge the evolving nature of this field and remain committed to staying engaged with ongoing research discussions, updating our perspectives, and incorporating relevant advancements in future iterations.

Beyond these technical and methodological limitations, AI hospitals also introduce broader societal implications that must be carefully considered. One of the most pressing concerns is ensuring equitable access to AI-driven healthcare solutions. While AI hospitals have the potential to enhance medical accessibility, their effectiveness is contingent on data diversity and infrastructure availability. Many existing AI models are trained on datasets predominantly sourced from high-resource settings, which may not generalize well to underserved populations. If not carefully designed, AI hospitals risk reinforcing healthcare disparities rather than alleviating them. Addressing this challenge requires the integration of diverse, representative datasets and the development of AI systems that prioritize inclusivity, particularly in low-resource environments.

Ethical and regulatory challenges also pose significant concerns for AI hospitals. The reliance on vast amounts of patient data raises critical questions about privacy, security, and informed consent. Ensuring compliance with regulations such as HIPAA and GDPR is essential, yet existing legal frameworks often struggle to keep pace with rapid AI advancements. Furthermore, AI-driven medical interactions must maintain transparency—patients should be clearly informed when they are engaging with AI rather than human clinicians. Establishing robust guidelines for AI deployment in healthcare settings will be crucial in mitigating risks related to misinformation, data misuse, and unintended biases.

Another major consideration is the evolving relationship between AI and healthcare professionals. AI hospitals have the potential to significantly enhance medical workflows, reducing administrative burdens and providing decision support. However, concerns about workforce displacement and overreliance on AI-generated insights remain. To maximize benefits, AI hospitals should be designed to complement rather than replace human expertise, ensuring that medical professionals retain agency in critical decision-making processes. Furthermore, AI systems must be interpretable and explainable, allowing clinicians to understand and trust AI-generated recommendations rather than blindly accepting automated outputs.

Liability and accountability in AI-driven healthcare also present complex challenges. In cases where AI hospital systems make erroneous diagnoses or treatment recommendations, assigning responsibility becomes a critical issue. Unlike traditional medical practice, where liability typically falls on healthcare providers, AI-driven decisions may involve multiple stakeholders, including AI developers, model trainers, and deploying institutions. The absence of clear legal frameworks to address these issues could hinder AI hospital adoption in real-world clinical settings. Future work should explore regulatory mechanisms that ensure

accountability while encouraging innovation in AI-driven healthcare.

Finally, AI hospitals represent an inflection point in the future of medical education and research. Simulated patient interactions and AI-assisted training offer unprecedented opportunities for personalized and scalable medical learning. However, excessive reliance on AI-generated feedback could weaken traditional medical training paradigms, necessitating careful integration of AI within existing curricula. Additionally, biases in training data may propagate into AI-assisted education, potentially influencing diagnostic patterns and medical decision-making. Addressing these risks requires ongoing validation of AI-generated educational content and collaborative efforts between AI researchers and medical educators to ensure that AI hospitals enhance, rather than distort, clinical reasoning skills.

As AI hospitals continue to evolve, balancing innovation with societal responsibility is imperative. While these systems offer transformative potential in healthcare, their widespread adoption must be guided by principles of fairness, transparency, and accountability. Moving forward, interdisciplinary collaboration between AI researchers, healthcare professionals, ethicists, and policymakers will be essential to ensuring that AI hospitals serve as equitable, reliable, and beneficial tools in the medical landscape.

References

[1] Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, and Ramesh C. Jain. 2023. Conversational health agents: A personalized llm-powered agent framework. *ArXiv*, abs/2310.02374.

[2] OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Made laine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Jo hannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo,

Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

[3] Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. 2022. Multimodal biomedical ai. *Nature Medicine*, 28:1773 – 1784.

[4] Mohammed Al Owayyed, Myrthe Tielman, Arno Hartholt, Marcus Specht, and Willem-Paul Brinkman. 2024. Agent-based social skills training systems: the artes architecture, interaction characteristics, learning theories and future outlooks. *Behaviour & Information Technology*, pages 1–28.

[5] Sauliha Rabia Alli, Soaad Qahhār Hossain, Sunit Das, and Ross Upshur. 2024. The potential of artificial intelligence tools for reducing uncertainty in medicine and directions for medical education. *JMIR Medical Education*, 10(1):e51446.

[6] Steven M Ariss. 2009. Asymmetrical knowledge claims in general practice consultations with frequently attending patients: Limitations and opportunities for patient participation. *Social Science & Medicine*, 69(6):908–919.

[7] Anmol Arora, Joseph E Alderman, Joanne Palmer, Shaswath Ganapathi, Elinor Laws, Melissa Mccradden, Lauren Oakden-Rayner, Stephen R. Pfohl, Marzyeh Ghassemi, Francis Mckay, Darren Treanor, Negar Rostamzadeh, Bilal A. Mateen, Jacqui Gath, Adewole O Adebajo, Stephanie Kuku, Rubeta N Matin, Katherine Heller, Elizabeth Sapey, Neil J. Sebire, Heather Cole-Lewis, Melanie J. Calvert, Alastair Keith Denniston, and Xiaoxuan Liu. 2023. The value of standards for health datasets in artificial intelligence-based applications. *Nature Medicine*, 29:2929 – 2938.

[8] Rebecca Baines, Sam Regan de Bere, Sebastian Stevens, James Morley Read, Martin Marshall, Mirza Lalani, Marie Bryce, and Julian Archer. 2018. The impact of patient feedback on the medical performance of qualified doctors: a systematic review. *BMC Medical Education*, 18.

[9] Michiel J Bakkum, Mariëlle G Hartjes, Joost D Piët, Erik M Donker, Robert Likic, Emilio Sanz, Fabrizio de Ponti, Petra Verdonk, Milan C Richir, Michiel A van Agtmael, et al. 2024. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. *British Journal of Clinical Pharmacology*, 90(3):640–648.

[10] Erin P Balogh, Bryan T Miller, John R Ball, et al. 2015. Committee on diagnostic error in health care; board on health care services; institute of medicine; the national academies of sciences, engineering, and medicine. improving diagnosis in health care. *Improving diagnosis in health care*.

[11] Zhijie Bao, Qingyun Liu, Ying Guo, Zhengqiang Ye, Jun Shen, Shirong Xie, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2024. Piors: Personalized intelligent outpatient reception based on large language model with multi-agents medical scenario simulation. *arXiv preprint arXiv:2411.13902*.

[12] Casey C Bennett and Kris Hauser. 2013. Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach. *Artificial intelligence in medicine*, 57(1):9–19.

[13] Markus Bertl, Yngve Lamo, Martin Leucker, Tiziana Margaria, Esfandiar Mohammadi, Suresh Kumar Mukhiya, Ludwig Pechmann, Gunnar Piho, and Fazle Rabbi. 2023. Challenges for ai in healthcare systems. In *International Conference on Bridging the Gap between AI and Reality*, pages 165–186. Springer Nature Switzerland Cham.

[14] Amy Blake and Bryan T Carroll. 2016. Game theory and strategy in medical training. *Medical education*, 50(11):1094–1106.

[15] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697.

[16] Rabai Bouderhem. 2024. Shaping the future of ai in healthcare through ethics and governance. *Humanities and Social Sciences Communications*, 11:1–12.

[17] Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

[18] Mohamed-Amine Chadi and Hajar Mousannif. 2022. Inverse reinforcement learning for healthcare applications: A survey. In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning*, volume 1, pages 97–102.

[19] Theodora Chatzimichail and Aristides T Hatjimihail. 2023. A bayesian inference based computational tool for parametric and nonparametric medical diagnosis. *Diagnostics*, 13(19):3135.

[20] Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Toby Jia-Jun Li, and Dakuo Wang. 2025. Towards a design guideline for rpa evaluation: A survey of large language model-based role-playing agents. *arXiv preprint arXiv:2502.13012*.

[21] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

[22] Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang. 2024. Predictive modeling with temporal graphical representation on electronic health records. In *IJCAI: proceedings of the conference*, volume 2024, page 5763.

[23] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

[24] 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*, abs/2407.13301.

[25] Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023. Huatuogpt-ii, one-stage training for medical adaption of llms. *ArXiv*, abs/2311.09774.

[26] Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. 2024. A survey on multi-generative agent system: Recent advances and new frontiers. *arXiv preprint arXiv:2412.17481*.

[27] Siyuan Chen, Mengyue Wu, Ke Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *ArXiv*, abs/2305.13614.

[28] Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. 2024. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. *arXiv preprint arXiv:2412.12475*.

[29] Yirong Chen, Zhenyu Wang, Xiaofen Xing, Huimin Zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *ArXiv*, abs/2310.15896.

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

[31] Jirapun Daengdej, Kitikorn Dowpiset, Kitti Phothikitti, and Vechayan Choychoowong. 2024. Multi-agent model for clinical decision support system. In *Bioethics of Cognitive Ergonomics and Digital Transition*, pages 171–184. IGI Global.

[32] Kris De Jaegher and Marc Jegers. 2001. The physician–patient relationship as a game of strategic information transmission. *Health Economics*, 10(7):651–668.

[33] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jiong Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaokang Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

[34] Tejaswini Dhawale, Lotte M Steuten, and H Joachim Deeg. 2017. Uncertainty of physicians and patients in medical decision making.

[35] George A Diamond, Alan Rozanski, and Michael Steuer. 1986. Playing doctor: application of game theory to medical decision-making. *Journal of chronic diseases*, 39(9):669–677.

[36] Benjamin Djulbegovic, Iztok Hozo, and John PA Ioannidis. 2015. Modern health care as a game theory

problem. *European journal of clinical investigation*, 45(1):1–12.

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

[38] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Bap tiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Laurens Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Babu Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Ro main Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Ben Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva,

Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

[39] Rahma Elkamouchi, Abdelaziz Daaif, and Kamal Elguemmat. 2024. Multi-agents system in healthcare: A systematic literature review. In *International Conference on Smart Applications and Data Analysis*, pages 200–214. Springer.

[40] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *International Conference on Computational Linguistics*.

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

[42] Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Annual Meeting of the Association for Computational Linguistics*.

[43] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

[44] 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):6125–6151.

[45] Matthias Gerstgrasser and David C Parkes. 2023. Oracles & followers: Stackelberg equilibria in deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 11213–11236. PMLR.

[46] S. Gilbert, J. N. Kather, and A. Hogan. 2024. Augmented non-hallucinating large language models as medical information curators. *npj Digital Medicine*, 7:100.

[47] Virginia Teas Gill and Douglas W Maynard. 2006. Explaining illness: patients' proposals and physicians' responses. *Studies in Interactional Sociolinguistics*, 20:115.

[48] Mauro Giuffré and Dennis L. Shung. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine*, 6.

[49] Dionysius Glycopantis and Charitini Stavropoulou. 2018. An agency relationship under general conditions of uncertainty: a game theory application to the doctor–patient interaction. *Economic Theory Bulletin*, 6:15–28.

[50] Alex J. Goodell, MD MS Simon N Chu, Dara Rouholiman, and MD Larry F Chu. 2023. Augmentation of chatgpt with clinician-informed tools improves performance on medical calculation tasks. In *medRxiv*.

[51] Cinzia Greco. 2020. Too much information, too little power: the persistence of asymmetries in doctor-patient relationships. *Anthropology now*, 12(2):53–60.

[52] 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*.

[53] Önder Gürcan. 2024. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 134–144.

[54] Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling, et al. 2024. A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR medical education*, 10(1):e53961.

[55] Shengxin Hong, Liang Xiao, Xin Zhang, and Jianxia Chen. 2024. Argmed-agents: Explainable clinical decision reasoning with llm disscusion via argumentation schemes. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5486–5493. IEEE.

[56] Minda Hu, Licheng Zong, Hongru Wang, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, and Irwin King. 2024. Serts: Self-rewarding tree search for biomedical retrieval-augmented generation. *arXiv preprint arXiv:2406.11258*.

[57] Yebowen Hu, Xiaoyang Wang, Wenlin Yao, Yiming Lu, Daoan Zhang, Hassan Foroosh, Dong Yu, and Fei Liu. 2024. Define: Enhancing llm decision-making with factor profiles and analogical reasoning. *arXiv preprint arXiv:2410.01772*.

[58] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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

[60] Pushkala Jayaraman, Jacob Desman, Moein Sabounchi, Girish N. Nadkarni, and Ankit Sakhuja. 2024. A primer on reinforcement learning in medicine for clinicians. *NPJ Digital Medicine*, 7.

[61] Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W. John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and Zhiyong Lu. 2024. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *ArXiv*, abs/2402.13225.

[62] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *ArXiv*.

[63] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2023. Guidelines for rigorous evaluation of clinical llms for conversational reasoning. *medRxiv*, pages 2023–09.

[64] Jerome P Kassirer. 1989. Our stubborn quest for diagnostic certainty.

[65] 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:e59439.

[66] Mutahira Khalid, Raihana Rahman, Asim Abbas, Sushama Kumari, Iram Wajahat, and Syed Ahmad Chan Bukhari. 2024. Accelerating medical knowledge discovery through automated knowledge graph generation and enrichment. In *International Knowledge Graph and Semantic Web Conference*, pages 62–77. Springer.

[67] Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid A. Anwar, Andrew Zhang, Aidan Gilson, Maxwell Singer, Amisha D. Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, and Zhiyong Lu. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *ArXiv*, abs/2406.12036.

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

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

[70] Prerana Sanjay Kulkarni, Muskaan Jain, Disha Sheshanarayana, and Srinivasan Parthiban. 2024. Hecix: Integrating knowledge graphs and large language models for biomedical research. *arXiv preprint arXiv:2407.14030*.

[71] Sunjae Kwon, Zonghai Yao, Harmon S Jordan, David A Levy, Brian Corner, and Hong Yu. 2022. Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 11733.

[72] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of

open-source pretrained large language models for medical domains. In *Annual Meeting of the Association for Computational Linguistics*.

[73] Moustafa Laymouna, Yuanchao Ma, David Lessard, Tibor Schuster, Kim Engler, and Bertrand Lebouché. 2024. Roles, users, benefits, and limitations of chatbots in health care: rapid review. *Journal of medical Internet research*, 26:e56930.

[74] Tyler Alise Le, Arpi Jivalagian, Tasneem Hiba, Joshua Franz, Shahab Ahmadzadeh, Treniece Eubanks, Leisa Oglesby, Sahar Shekoohi, Elyse M Cornett, and Alan D Kaye. 2023. Multi-agent systems and cancer pain management. *Current Pain and Headache Reports*, 27(9):379–386.

[75] Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large language models in finance (finllms). *Neural Computing and Applications*, pages 1–15.

[76] Esa Lehtinen. 2013. Hedging, knowledge and interaction: Doctors' and clients' talk about medical information and client experiences in genetic counseling. *Patient education and counseling*, 92(1):31–37.

[77] Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

[78] Binxu Li, Tiankai Yan, Yuanting Pan, Zhe Xu, Jie Luo, Ruiyang Ji, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024. Mmedagent: Learning to use medical tools with multi-modal agent. In *Conference on Empirical Methods in Natural Language Processing*.

[79] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

[80] J. Li, X. Chen, W. Liu, L. Wang, Y. Guo, M. You, others, and K. Li. 2023. One is not enough: Multi-agent conversation framework enhances rare disease diagnostic capabilities of large language models.

[81] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *ArXiv*, abs/2405.02957.

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

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

[84] Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Neural Information Processing Systems*.

[85] Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*.

[86] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

[87] Zhenzhu Li, Jingfeng Zhang, Zhou Wei, Jianjun Zheng, and Yinshui Xia. 2024. Gpt-agents based on medical guidelines can improve the responsiveness and explainability of outcomes for traumatic brain injury rehabilitation. *Scientific Reports*, 14.

[88] Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.

[89] Jung Hoon Lim, Sunjae Kwon, Zonghai Yao, John P Lalor, and Hong Yu. 2024. Large language model-based role-playing for personalized medical jargon extraction. *arXiv preprint arXiv:2408.05555*.

[90] Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z Li, and Kaicheng Yu Biokgbench. 2024. A knowledge graph checking benchmark of ai agent for biomedical science. *arXiv preprint arXiv*, 2407.

[91] Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Tianfan Fu, and Yue Zhao. 2024. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *ArXiv*, abs/2411.15692.

[92] Zhaocheng Liu, Quan Tu, Wen Ye, Yu Xiao, Zhishou Zhang, Hengfu Cui, Yalun Zhu, Qiang Ju, Shizheng Li, and Jian Xie. 2025. Exploring the inquiry-diagnosis relationship with advanced patient simulators. *arXiv preprint arXiv:2501.09484*.

[93] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *arXiv preprint arXiv:2407.00870*.

[94] Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764.

[95] Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. 2024. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6).

[96] Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul Janardhan Butte, and Ahmed Alaa. 2024. Evaluating large language models as agents in the clinic. *NPJ Digital Medicine*, 7.

[97] Prakamya Mishra, Zonghai Yao, Parth Vashisht, Feiyun Ouyang, Beining Wang, Vidhi Dhaval Mody, and Hong Yu. 2024. Synfac-edit: Synthetic imitation edit feedback for factual alignment in clinical summarization. *arXiv preprint arXiv:2402.13919*.

[98] Sally Moy, Mona Irannejad, Stephanie Jeanneret Manning, Mehrdad Farahani, Yomna Ahmed, Ellis Gao, Radhika Prabhune, Suzan Lorenz, Raza Mirza, and Christopher Klinger. 2024. Patient perspectives on the use of artificial intelligence in health care: a scoping review. *Journal of Patient-Centered Research and Reviews*, 11(1):51.

[99] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

[100] Oded Nov, Nina Singh, and Devin M. Mann. 2023. Putting chatgpt's medical advice to the (turing) test: Survey study. *JMIR Medical Education*, 9.

[101] Jasmine Chiat Ling Ong, Benjamin Jun Jie Seng, Jeren Zheng Feng Law, Lian Leng Low, Andrea Lay Hoon Kwa, Kathleen M Giacomini, and Daniel Shu Wei Ting. 2024. Artificial intelligence, chatgpt, and other large language models for social determinants of health: Current state and future directions. *Cell Reports Medicine*, 5(1).

[102] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

[103] Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

[104] Stephen G Pauker and Jerome P Kassirer. 1980. The threshold approach to clinical decision making. *The New England journal of medicine*, 302(20):1109–1117.

[105] Daniela SM Pereira, Filipe Falcão, Andreia Nunes, Nuno Santos, Patrício Costa, and José Miguel Pêgo. 2023. Designing and building oscebot® for virtual osce–performance evaluation. *Medical Education Online*, 28(1):2228550.

[106] Kristina Polotskaya, Carlos S Muñoz-Valencia, Alejandro Rabasa, Jose A Quesada-Rico, Domingo Orozco-Beltrán, and Xavier Barber. 2024. Bayesian networks for the diagnosis and prognosis of diseases: A scoping review. *Machine Learning and Knowledge Extraction*, 6(2):1243–1262.

[107] Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.

[108] Thomas P Quinn, Manisha Senadeera, Stephan Jacobs, Simon Coghlan, and Vuong Le. 2021. Trust and medical ai: the challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association*, 28(4):890–894.

[109] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

[110] Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.

[111] Jeannette A Rogowski and Lynn A. Karoly. 1998. Health insurance portability and accountability act of 1996.

[112] Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter M Abadir, and Rama Chellappa. 2024. Addressing cognitive bias in medical language models. *ArXiv*, abs/2402.08113.

[113] 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*.

[114] Hanwen Shi, Jin Zhang, and Kunpeng Zhang. 2024. Enhancing clinical trial patient matching through knowledge augmentation with multi-agents. *arXiv preprint arXiv:2411.14637*.

[115] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. 2024. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Conference on Empirical Methods in Natural Language Processing*.

[116] Ben Singh, Andrew Murphy, Carol Maher, and Ashleigh E Smith. 2024. Time to form a habit: A systematic review and meta-analysis of health behaviour

habit formation and its determinants. In *Healthcare*, volume 12, page 2488. Multidisciplinary Digital Publishing Institute.

[117] Andries P. Smit, Paul Duckworth, Nathan Grinsztajn, Kale ab Tessera, Thomas D. Barrett, and Arnu Pretorius. 2023. Should we be going mad? a look at multi-agent debate strategies for llms. In *International Conference on Machine Learning*.

[118] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *ArXiv*, abs/2408.03314.

[119] Aaron J Snoswell, Centaine L Snoswell, and Nan Ye. 2024. Eliciting patient preferences and predicting behaviour using inverse reinforcement learning for telehealth use in outpatient clinics. *Frontiers in Digital Health*, 6:1384248.

[120] Haoran Sun, Yusen Wu, Yukun Cheng, and Xu Chu. 2025. Game theory meets large language models: A systematic survey. *arXiv preprint arXiv:2502.09053*.

[121] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Zhongyi Shui, Xiaoxuan Yu, Yizhi Zhao, Honglin Li, Yunlong Zhang, Ruojia Zhao, Xinheng Lyu, and Lin Yang. 2023. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *AAAI Conference on Artificial Intelligence*.

[122] Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2024. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pages 2024–11.

[123] Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V Stolyar, Katelyn Polanska, Karleigh R McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, et al. 2024. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ digital medicine*, 7(1):258.

[124] 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*.

[125] Muhammad Usman Tariq. 2024. Multi-agent models in healthcare system design. In *Bioethics of Cognitive Ergonomics and Digital Transition*, pages 143–170. IGI Global.

[126] Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. 2023. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *Journal of the American Medical Informatics Association : JAMIA*.

[127] Hieu Tran, Zonghai Yao, Junda Wang, Yifan Zhang, Zhichao Yang, and Hong Yu. 2024. Rare: Retrieval-augmented reasoning enhancement for large language models. *arXiv preprint arXiv:2412.02830*.

[128] T. Tu, A. Palepu, M. Schaekermann, K. Saab, J. Freyberg, R. Tanno, others, and V. Natarajan. 2024. Towards conversational diagnostic ai. *ArXiv*, abs/2401.05654.

[129] Daiju Ueda, Taichi Kakinuma, Shohei Fujita, Koji Kamagata, Yasutaka Fushimi, Rintaro Ito, Yusuke Matsui, Taiki Nozaki, Takeshi Nakaura, Noriyuki Fujima, Fuminari Tatsugami, Masahiro Yanagawa, Kenji Hirata, Akira Yamada, Takahiro Tsuboyama, Mariko Kawamura, Tomoyuki Fujioka, and Shinji Naganawa. 2023. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42:3 – 15.

[130] Vivek Verma, Ashwani Kumar Mishra, and Rajiv Narang. 2019. Application of bayesian analysis in medical diagnosis. *Journal of the Practice of Cardiovascular Sciences*, 5(3):136–141.

[131] Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, Zhengyun Zhao, Da Pan, Fan Yang, Fei Kou, Fei Li, Fuzhong Chen, Guosheng Dong, Han Liu, Hongda Zhang, Jin He, Jinjie Yang, Kangxi Wu, Ke-Ye Wu, Lei Su, Linlin Niu, Lin-Lin Sun, Mang Wang, Peng Fan, Qi Shen, Rihui Xin, Shunya Dang, Song Zhou, Weipeng Chen, Wenjing Luo, Xin Chen, Xin Men, Xionghai Lin, Xu Dong, Yan Zhang, Yifei Duan, Yuyan Zhou, Zhi-Xing Ma, and Zhi-Yan Wu. 2025. Baichuan-m1: Pushing the medical capability of large language models.

[132] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024. Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107*.

[133] Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024. Towards a client-centered assessment of llm therapists by client simulation. *arXiv preprint arXiv:2406.12266*.

[134] Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

[135] Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2023. Notechat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Annual Meeting of the Association for Computational Linguistics*.

[136] Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024. Patient-psi: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797.

[137] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

[138] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, 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*.

[139] Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023. Augmenting black-box llms with medical textbooks for clinical question answering. *ArXiv*, abs/2309.02233.

[140] Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. 2024. Twin-gpt: digital twins for clinical trials via large language model. *ACM Transactions on Multimedia Computing, Communications and Applications*.

[141] Zifeng Wang, Benjamin Danek, Ziwei Yang, Zheng Chen, and Jimeng Sun. 2024. Can large language models replace data scientists in clinical research? *arXiv preprint arXiv:2410.21591*.

[142] Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Tianlong Wang, Wen Tang, Yasha Wang, Chengwei Pan, Ewen M Harrison, Junyi Gao, et al. 2024. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. *arXiv preprint arXiv:2410.02551*.

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

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

[145] Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. *ArXiv*, abs/2307.04986.

[146] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

[147] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

[148] Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, MingZhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, Shaoqing Jiao, and Jiajie Peng. 2024. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *bioRxiv*.

[149] Feng Xie, Han Yuan, Yilin Ning, Marcus Eng Hock Ong, Mengling Feng, Wynne Hsu, Bibhas Chakraborty, and Nan Liu. 2022. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of biomedical informatics*, 126:103980.

[150] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *ArXiv*, abs/2402.13178.

[151] Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, et al. 2025. Rag-gym: Optimizing reasoning and search agents with process supervision. *arXiv preprint arXiv:2502.13957*.

[152] Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 30:199–214.

[153] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *ArXiv*, abs/2304.01097.

[154] Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S Yu. 2024. Large language models for education: A survey. *arXiv preprint arXiv:2405.13001*.

[155] Shaochen Xu, Yifan Zhou, Zhengliang Liu, Zihao Wu, Tianyang Zhong, Huaqin Zhao, Yiwei Li, Hanqi Jiang, Yi Pan, Junhao Chen, et al. 2024. Towards next-generation medical agent: How o1 is reshaping decision-making in medical scenarios. *arXiv preprint arXiv:2411.14461*.

[156] Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, Li Zhu, and Xuandong Zhao. 2024. Clinicallab: Aligning agents

for multi-departmental clinical diagnostics in the real world. *ArXiv*, abs/2406.13890.

[157] Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29.

[158] Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

[159] Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Feiyun Ouyang, Beining Wang, David McManus, Dan Berlowitz, and Hong Yu. 2025. Unveiling gpt-4v's hidden challenges behind high accuracy on usmle questions: Observational study. *Journal of Medical Internet Research*, 27:e65146.

[160] Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, Hong Yu, et al. 2023. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. *arXiv preprint arXiv:2312.15561*.

[161] Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. 2024. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. *arXiv preprint arXiv:2410.13191*.

[162] Zonghai Yao, Benjamin J Schloss, and Sai P Selvaraj. 2023. Improving summarization with human edits. *arXiv preprint arXiv:2310.05857*.

[163] Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and Hong Yu. 2024. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *ArXiv*, abs/2410.01553.

[164] Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Wenyue Hua, Mingyu Jin, Guang Chen, et al. 2024. Aipatient: Simulating patients with ehrs and llm powered agentic workflow. *arXiv preprint arXiv:2409.18924*.

[165] 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*, pages 1–10.

[166] Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2024. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. *arXiv preprint arXiv:2407.09893*.

[167] Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan A. Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curtis P. Langlotz, Rita Lee, Joanna Melia, Joanna Nelson, Karim Sallam, Stacey Tullis, Melissa Ann Vogelsong, John Patrick Cunningham, and William Hiesinger. 2024. Almanac - retrieval-augmented language models for clinical medicine. *NEJM AI*, 1 2.

[168] Chao Zhang, Joaquin Vanschoren, Arlette Van Wissen, Daniël Lakens, Boris De Ruyter, and Wijnand A IJsselsteijn. 2022. Theory-based habit modeling for enhancing behavior prediction in behavior change support systems. *User Modeling and User-Adapted Interaction*, 32(3):389–415.

[169] Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. 2023. Llm-based medical assistant personalization with short- and long-term memory coordination. In *North American Chapter of the Association for Computational Linguistics*.

[170] Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

[171] 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*.

[172] Wei Zhu, Wenfeng Li, Xing Tian, Pengfei Wang, Xiaoling Wang, Jin Chen, Yuanbin Wu, Yuan Ni, and Guotong Xie. 2024. Text2mdt: extracting medical decision trees from medical texts. *arXiv preprint arXiv:2401.02034*.

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

ID	Paper Title	Venue	Code/Data	Study
1	Exploring the Inquiry-Diagnosis Relationship with Advanced Patient Simulators	arXiv	link	[92]
2	Leveraging Large Language Model as Simulated Patients for Clinical Education	arXiv	No	[85]
3	A GPT-Powered Chatbot as a Simulated Patient to Practice History Taking	JMIR Med Edu	No	[54]
4	Designing and building OSCEBot ® for virtual OSCE – Performance evaluation	Med Edu Online	No	[105]
5	Roleplay-doh: Enabling domain-experts to create LLM-simulated patients	EMNLP24	link	[93]
6	AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow	arXiv	link	[164]
7	LLMs Can Simulate Standardized Patients via Agent Coevolution	arXiv	link	[37]
8	PATIENT-Ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals	EMNLP24	link	[136]
9	Towards a Client-Centered Assessment of LLM Therapists by Client Simulation	arXiv	link	[133]
10	Automatic Interactive Evaluation for LLMs with State Aware Patient Simulator	arXiv	link	[88]
11	Guidelines For Rigorous Evaluation of Clinical LLMs For Conversational Reasoning	medRxiv	link	[63]
12	RAREAGENTS: Autonomous Multi-disciplinary Team for Rare Disease Diagnosis	arXiv	No	[28]
13	TWIN-GPT: Digital Twins for Clinical Trials via Large Language Model	TOMM24	No	[140]
14	Interactive Agents: Simulating Counselor-Client Psychological Counseling	arXiv	link	[107]
15	DrHouse: An LLM-empowered Diagnostic Reasoning System	IMWUT	No	[157]
16	MDAgents: Adaptive Collaboration of LLMs for Medical Decision-Making	NeurIPS 2024	link	[68]
17	MEDAGENTS: Large Language Models as Collaborators for Medical Reasoning	ACL Findings 2024	link	[124]
18	ColaCare: Enhancing EHR Modeling through LLM Multi-Agent Collaboration	arXiv	link	[142]
19	Mitigating Cognitive Biases in Clinical Decision-Making via Multi-Agent LLMs	JMIR	No	[65]
20	AgentClinic: A Multimodal Agent Benchmark for Simulated Clinical Environments	arXiv	link	[113]
21	TRIAGEAGENT: Multi-Agents for LLM-Based Clinical Triage	EMNLP Findings 2024	link	[94]
22	PIORS: Personalized Intelligent Outpatient Reception Using Multi-Agents	arXiv	link	[11]
23	Can Large Language Models Replace Data Scientists in Clinical Research?	arXiv	No	[141]
24	The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies	BioRxiv	link	[122]
25	MEDCO: Medical Education Copilots Using Multi-Agent Framework	arXiv	No	[143]
26	Should We Be Going MAD? Multi-Agent Debate Strategies for LLMs	ICML2024	link	[117]
27	Beyond Direct Diagnosis: Multi-Specialist Agent Consultation for Diagnosis	arXiv	No	[132]
28	ClinicalAgent: Clinical Trial Multi-Agent System with LLM Reasoning	BCB '24	link	[165]
29	Enhancing Clinical Trial Patient Matching via Multi-Agent Knowledge Augmentation	arXiv	No	[114]
30	ArgMed-Agents: Explainable Clinical Decision Reasoning via Argumentation	BIBM2024	No	[55]
31	Synergistic Multi-Agent Framework with Trajectory Learning	AAAI25	link	[166]
32	Empowering Biomedical Discovery with AI Agents	Cell	No	[44]
33	MedDM: LLM-executable Clinical Guidance Tree for Decision-Making	arXiv	No	[77]
34	Text2MDT: Extracting Medical Decision Trees from Texts	arXiv	No	[172]
35	BioKGBench: A Knowledge Graph Benchmark	arXiv	link	[90]
36	Medical Graph RAG: Safe LLMs via Graph Retrieval-Augmented Generation	arXiv	link	[146]
37	HeCiX: Integrating Knowledge Graphs and LLMs for Biomedical Research	arXiv	No	[70]
38	KRAGEN: Knowledge Graph-Enhanced RAG for Biomedical Problem Solving	Bioinformatics	link	[95]
39	Accelerating Medical Knowledge Discovery via Automated Knowledge Graphs	KGSWC 2024	No	[66]
40	Augmented Non-Hallucinating LLMs as Medical Information Curators	npj Digital Medicine	No	[46]
41	Benchmarking Retrieval-Augmented Generation for Medicine	ACL Findings 2024	link	[150]
42	Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions	arXiv	link	[152]
43	Almanac — Retrieval-Augmented Language Models for Clinical Medicine	NEJM AI	link	[167]
44	Augmenting Black-box LLMs with Medical Textbooks for Biomedical QA	EMNLP Findings 2024	link	[139]
45	To Generate or Retrieve? Effectiveness of Artificial Contexts in Medical QA	ACL 2024	link	[42]
46	AgentMD: Empowering Language Agents for Risk Prediction	arXiv	link	[61]
47	MedCalc-Bench: Evaluating LLMs for Medical Calculations	NeurIPS 2024	link	[67]
48	Augmenting ChatGPT with Clinician-Informed Tools for Medical Calculations	medRxiv	No	[50]
49	GeneGPT: Augmenting LLMs with Domain Tools for Biomedical Information	Bioinformatics	link	[62]
50	EHRAgent: Code-Empowered LLMs for Few-shot Complex Tabular Reasoning	EMNLP 2024	link	[115]
51	MMedAgent: Learning to Use Medical Tools with Multi-modal Agent	EMNLP Findings 2024	link	[78]
52	Conversational Health Agents: A Personalized LLM-Powered Agent Framework	arXiv	link	[1]
53	PathAsst: A Generative AI Assistant for Pathology Analysis	AAAI Technical Track	link	[121]
54	GPT-agents Based on Medical Guidelines for Traumatic Brain Injury Rehabilitation	Scientific Reports	No	[87]
55	CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis	arXiv	No	[148]
56	DrugAgent: Automating AI-Aided Drug Discovery via LLM Multi-Agent Collaboration	arXiv	No	[91]
57	Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents	arXiv	No	[81]
58	AI Hospital: Benchmarking LLMs in a Multi-agent Medical Interaction Simulator	COLING 2025	link	[40]
59	ClinicalLab: Aligning Agents for Multi-Departmental Clinical Diagnostics	arXiv	link	[156]
60	LLM-empowered Chatbots for Psychiatrist and Patient Simulation	arXiv	No	[27]
61	MediQ: Question-Asking LLMs and a Benchmark for Interactive Clinical Reasoning	NeurIPS 2024	link	[84]
62	Epidemic Modeling with Generative Agents	arXiv	link	[145]
63	NoteChat: A Dataset of Synthetic Patient-Physician Conversations	ACL 2024 Findings	link	[135]
64	Evaluating Large Language Models as Agents in the Clinic	npj Digital Medicine	No	[96]
65	MedQA-CS: Benchmarking LLMs Clinical Skills Using an AI-SCE Framework	arXiv	link	[163]
66	Towards Conversational Diagnostic AI	arXiv	No	[128]
67	LLM-based Medical Assistant Personalization with Short- and Long-Term Memory	NAACL 2024	link	[169]
68	CoD: Towards an Interpretable Medical Agent Using Chain of Diagnosis	arXiv	link	[24]
69	Multi-Agent Conversation Framework Enhances Rare Disease Diagnosis in LLMs	Preprint	link	[80]
70	MEDAIDE: Towards an Omni Medical Aide via Specialized LLM-based Multi-Agent Collaboration	Preprint	No	[144]
71	RAG-Gym: Optimizing Reasoning and Search Agents with Process Supervision	Preprint	link	[151]
72	MCQG-SRefine: Multiple Choice Question Generation and Evaluation with Iterative Self-Critique, Correction, and Comparison Feedback	NAACL 2025	link	[161]

Key Challenges	Future Research Directions
§ 11.1 Profile/Roles	
1. Role Consistency: Ensuring that doctor and patient agents exhibit behavior consistent with their roles across different contexts, such as doctors demonstrating diverse diagnostic styles and decision-making approaches, while patients dynamically adjust their information disclosure strategies based on consultation stages. **2. Modeling Information Asymmetry**: Simulating real-world information asymmetry, where patients may selectively disclose information due to privacy or psychological factors, while doctors must make decisions with limited information. **3. Inverse Reinforcement Learning (IRL) for Patient Decision-Making**: Real patients' behaviors are not driven by fixed reward functions. Using IRL to learn patient decision patterns (e.g., healthcare-seeking timing, treatment adherence) can enhance patient agent realism. **4. Patient Population Diversity**: Current patient agents may be overly homogeneous. Integrating social determinants of health (SDOH), such as housing, economic status, and educational background, can enhance diversity, ensuring system fairness and generalizability.	**1. Enhancing Role-Playing and Personalization Techniques**: Utilizing short-term memory modules, interview-driven personality modeling, and expert feedback optimization to make agent behavior more aligned with real-world medical scenarios. **2. Modeling Patient Behavior with Uncertainty**: Introducing behavior patterns like avoidance of negative diagnoses and risk perception adjustments to better simulate patient decision-making using heuristic methods and utility functions. **3. Using IRL to Improve Patient Agent Realism**: Learning real patients' decision trajectories to enable AI agents to better simulate patient decision-making across different contexts, thereby improving medical simulations. **4. Building More Representative Patient Agents**: Incorporating factors such as SDOH to ensure AI hospital systems accurately reflect the healthcare behaviors of diverse populations, improving applicability in medical training and patient education.
§ 11.2 Interaction Patterns	
1. Defining Human Roles: Current AI hospital systems primarily view AI agents as assistive tools, without clarifying whether humans should act as observers, active participants, or even replace certain AI agent functions. **2. Strategic Decision-Making and Information Uncertainty Modeling**: Existing interaction models rely mainly on end-to-end LLM predictions, lacking explicit mathematical modeling, making it difficult to capture inherent information asymmetry in medical scenarios. **3. Collaboration and Competition Among Multi-Agent Systems**: Long-term interactions between LLM agents remain underexplored. Doctor and patient agents may have competitive relationships in certain tasks (zero-sum games) but are mostly cooperative (cooperative games). **4. Modeling Medical Uncertainty**: Both patient and doctor agents may lack complete information during diagnosis. Optimizing interaction strategies in highly uncertain environments remains a challenge.	**1. Incorporating Human Interaction for Evaluation and Enhancement**: Embedding real humans in AI hospital systems to explore the differentiation between AI agents and humans (Turing-like tests) and assess optimal human-AI collaboration models. **2. Optimizing AI Agent Interaction via Game Theory**: Using methods such as Stackelberg games, Bayesian games, and informational games to model AI hospital systems, improving decision-making under information asymmetry. **3. Enhancing Long-Term Evolution Mechanisms Among AI Agents**: Applying evolutionary game theory to optimize strategies over time, such as patient agents learning effective symptom disclosure and doctor agents refining diagnostic questioning techniques. **4. Using Bayesian Inference to Improve Medical Decision-Making**: Developing Bayesian game-based diagnostic strategies that allow doctor agents to optimize questioning methods under uncertainty, while patient agents dynamically adjust responses based on perception, improving realism and medical education value.
§ 11.3 Tools	
1. Static Integration of Tools: Current AI hospital systems treat tools as static components, lacking systematic evaluation methods to assess their actual effectiveness in medical environments. **2. Uncertainty in Tool Effectiveness**: For instance, LLM-as-KB has shown superiority over traditional RAG in specific benchmarks, but its advantages in real-world medical applications remain unclear. **3. Lack of Real-World Impact-Based Evaluation Frameworks**: Existing tool evaluations rely primarily on standardized quantitative metrics, whereas clinical applications should assess tools based on their impact on agent interactions and patient health outcomes.	**1. Dynamic Tool Integration and Adaptive Optimization**: Exploring how AI hospital system tools can dynamically adapt to different tasks and contexts rather than being statically invoked, enhancing applicability in complex medical decision-making. **2. Validating Tool Performance in Real Medical Tasks**: Moving beyond traditional benchmarks to establish evaluation frameworks specific to AI hospital systems, measuring tool effectiveness in supporting doctor decision-making and improving patient education. **3. Analyzing the Impact of Tools on Agent Interactions and Medical Outcomes**: Developing novel evaluation metrics to assess how tools influence doctor-patient agent collaboration efficiency, information accuracy, and overall decision-making quality.
§ 11.4 Memory	
1. Limitations of Static EHR: Current methods treat EHRs as static knowledge bases, neglecting the temporal dependencies of disease progression, making it difficult to reflect patients' long-term health conditions comprehensively. **2. Insufficient Dynamic Memory Access Mechanisms**: Existing memory modules lack effective triggering mechanisms, making it difficult to dynamically adjust information storage and retrieval based on patient health literacy or behavioral feedback. **3. Lack of Patient Behavior Modeling**: Current systems fail to simulate long-term patient health behavior changes, such as how treatment adherence evolves in chronic disease management, making it challenging for doctor agents to adapt their interaction strategies.	**1. Time-Series Health Data Modeling**: Constructing temporal graphs to encode patient history, medication usage, and consultation records, enabling LLM agents to identify key disease progression points and optimize medical interactions. **2. Intelligent Memory Access Optimization**: Introducing adjustable access control mechanisms, such as health literacy-based reading difficulty detection, ensuring that patient agents receive medical information at an appropriate comprehension level. **3. Behavioral Adaptive Memory Modules**: Leveraging habit-forming models to simulate patients transitioning from doctor dependence to autonomous health management, allowing AI agents to provide personalized medical support at different stages.
§ 11.5 Reasoning Patterns	
1. Limitations of Single Reasoning Paths: CExisting methods primarily rely on direct step-by-step reasoning, which struggles to handle the complexity and dynamic nature of real-world medical environments. **2. Insufficient Handling of Uncertainty**: Doctor and patient agents often lack complete information during interactions, and current AI reasoning frameworks struggle to flexibly adjust decisions, increasing the likelihood of errors or hallucinations. **3. Lack of Dynamic Reasoning Mechanisms**: AI agents in multi-agent interactions still operate with independent reasoning, lacking the ability to dynamically adjust decisions based on ongoing interactions, limiting their performance in complex medical tasks.	**1. Expanding Uncertainty Modeling Methods**: Incorporating Bayesian inference to allow AI agents to adjust reasoning paths through probabilistic updates rather than relying solely on deterministic reasoning. **2. Introducing Time-Series Decision Models**: Utilizing Markov Decision Processes (MDP) to optimize AI agents' decision-making in patient interactions, enabling dynamic diagnostic strategies based on state changes. **3. Using POMDPs for Partially Observable Environments**: Applying Partially Observable Markov Decision Processes (POMDPs) to help AI agents make more reasonable inferences when full patient history is unavailable, such as prompting clarifying questions instead of making premature conclusions. **4. Integrating Multi-Agent Collaborative Reasoning**: Developing new reasoning mechanisms that enable different AI agents to dynamically adjust their decisions based on shared information, improving the intelligence and adaptability of the overall medical system.
§ 11.6 Simulating Specific Scenarios & Solving Complex Tasks	
1. Limitations in Medical Simulations: Current systems focus primarily on patient consultation stages, lacking comprehensive simulations of preoperative preparation, postoperative recovery, and chronic disease management, reducing real-world applicability. **2. Influence of External Environmental Factors**: Public health events (e.g., COVID-19) can alter hospital operations and patient behaviors, but existing systems lack adaptability to unexpected events, limiting their generalization capabilities. **3. Insufficient Social Cognition Modeling**: Patient decision-making is often influenced by social dynamics, peer influence, and observational learning, yet current AI agents lack the ability to simulate these behaviors, reducing their effectiveness in health education and disease management. **4. System Robustness Issues**: Multi-agent architectures may lead to hallucination generation, bias accumulation, and difficulties in handling long-form interactions, where frequent interactions amplify errors, decreasing overall system reliability. **5. Inadequate Risk Management**: Existing systems struggle to handle long-tail cases, rare diseases, or adversarial attacks, where error accumulation may lead to misdiagnosis or resource waste, requiring improved safety mechanisms.	**1. Expanding Coverage of Medical Scenarios**: Incorporating long-term health management, postoperative recovery, and chronic disease monitoring modules into AI hospital systems to improve simulation comprehensiveness and real-world adaptability. **2. Enhancing Adaptability to External Events**: Developing dynamic behavior adjustment and memory mechanisms to enable AI agents to respond effectively to public health crises or emergency medical situations, improving system robustness. **3. Incorporating Social Cognition Theories**: Designing patient agents with observational learning mechanisms to simulate the impact of social influences on medical decision-making and optimizing AI interaction in online patient communities and medical forums. **4. Optimizing Multi-Agent Collaboration Frameworks**: Reducing error propagation by developing fair benchmarking tests and optimization algorithms to ensure multi-agent systems outperform single-agent or standalone LLMs in complex tasks. **5. Introducing Uncertainty Quantification and Safety Protocols**: Implementing safety triggers (e.g., expert intervention, anomaly detection) in high-risk scenarios and using extreme-case simulations to enhance system reliability in rare disease cases.
§ 11.7 Evaluating Agents	
1. Limitations of Existing Evaluation Methods: Current evaluation frameworks focus primarily on task accuracy, traditional generation metrics, or LLM-as-Judge assessments, lacking alignment with real-world medical environments where doctors rely on patient feedback and peer reviews. **2. Insufficient Consideration of Computational Costs and Efficiency**: High performance in multi-agent AI hospital systems may partially depend on increased computational resources, but no standardized cost-performance trade-off analysis framework currently exists, making evaluations unrealistic. **3. Lack of Fair Benchmarking Tests**: Inconsistent test datasets, varying computational resource allocation, and vague task definitions hinder cross-system comparisons, reducing the reliability of evaluation results. **4. Limitations of Medical LLMs in Agent-Based Tasks**: While medical-specific LLMs (e.g., Med-PaLM2, DoctorGLM) possess superior medical knowledge, their intelligent behavior in AI hospital environments remains weak, often relegating them to tools rather than autonomous agents.	**1. Developing More Realistic Agent Evaluation Frameworks**: Incorporating social evaluation mechanisms (e.g., patient feedback, peer ratings, interaction quality analysis) to simulate how doctors are assessed in real-world environments, making evaluations more aligned with medical practice. **2. Optimizing Computational Cost Assessment**: Creating weighted cost models that analyze trade-offs between computational resource consumption, inference time, and performance gains, reducing over-reliance on large models in multi-agent AI hospital systems. **3. Establishing Fair Multi-Agent Benchmark Tests**: Standardizing test datasets, computational resources, and task definitions to ensure fair and reliable evaluations between multi-agent and single-agent systems, improving reproducibility in research. **4. Enhancing Medical LLMs' Agent Capabilities**: Investigating how to retain intelligent agent capabilities in medical-specific LLMs, such as optimizing autonomous decision-making and interaction strategies to enable them to perform complex tasks in multi-agent environments. **5. Developing Evaluation Standards Beyond Medical Exams**: Moving beyond medical exam-based evaluations to build broader clinical task benchmarks covering medical reasoning, interaction ability, and real-world applications for a more comprehensive performance assessment.
§ 11.8 Synthesizing Data for Training	
1. Limitations of RL in Medical Environments: AI hospitals have not been fully utilized as reinforcement learning (RL) environments, and real-world medical data scarcity and ethical constraints make it difficult to design appropriate training environments and reward mechanisms. **2. Lack of Diversity and Fairness in Synthetic Data**: Current synthetic data generation heavily relies on manual rules, failing to comprehensively simulate real-world medical scenarios. Long-term self-training may lead to data homogeneity and mode collapse, reducing model generalizability. **3. Absence of Standardized and Shareable Training Data**: Existing training environments are relatively isolated, making it difficult for different AI hospital systems to share synthetic data, limiting model portability and cross-system applicability.	**1. Utilizing AI Hospitals as RL Training Environments**: Designing reward mechanisms based on patient simulation and doctor decision-making, enabling AI agents to optimize medical decision-making through interactive learning, such as improving post-surgery care interventions. **2. Enhancing the Dynamism and Multimodal Nature of Synthetic Data**: Incorporating multi-agent collaboration to generate synthetic data that more closely mirrors real-world conditions while integrating text, images, and speech to improve data expressiveness. **3. Developing Data Quality Assessment and Bias Detection Mechanisms**: Creating automated data evaluation tools to detect and correct biases and errors in synthetic data, ensuring that it enhances AI agent capabilities without introducing unfairness. **4. Establishing Standardized and Shareable Synthetic Data Frameworks**: Developing unified data standards and benchmarks to facilitate synthetic data sharing across AI hospital systems, improving model stability and portability.
§ 11.9 Governance, Ethics, and the Roles of AI Researchers and Medical Practitioners	
1. Accountability and Transparency: As AI hospital systems play a growing role in medical decision-making, a major ethical concern is how to define accountability for errors made by AI agents while ensuring system transparency and traceability. **2. Bias and Its Impact on Healthcare Equity**: Medical AI systems may introduce implicit biases in resource allocation, exacerbating social inequalities. A unified governance framework is lacking to regulate fairness, safety, and privacy protection. **3. Challenges in Clinical Integration of AI Hospital Systems**: AI is still difficult to seamlessly integrate into doctors' workflows. Healthcare professionals may perceive AI as an additional burden rather than a genuinely useful clinical support tool. **4. Lack of Interdisciplinary Collaboration**: There remains a gap between AI research and medical practice. Limited involvement of physicians and healthcare professionals in AI development results in systems that fail to effectively address real-world medical needs.	**1. Establishing Governance Frameworks for AI Hospital Systems**: Implementing human oversight mechanisms to monitor critical decisions, introducing transparent version management to ensure system updates are traceable, and promoting international collaboration to develop unified AI governance standards in healthcare. **2. Enhancing Fairness and Explainability in AI Hospital Systems**: Developing fairness evaluation and bias correction mechanisms to ensure equitable resource allocation and prevent AI from reinforcing biases in medical decision-making. **3. Seamless Integration of AI into Clinical Workflows**: Designing AI systems that align with doctors' workflows, ensuring they serve as assistive tools rather than additional burdens, and developing user interfaces that meet clinical needs. **4. Bridging AI Research and Medical Practice**: Encouraging active participation of physicians, nurses, and other healthcare professionals in AI development and evaluation to ensure AI hospital systems effectively address clinical challenges and improve the synergy between AI research and real-world healthcare applications. **5. Exploring High-Fidelity Clinical Simulation Environments**: Utilizing AI hospital systems to create realistic medical training environments that enhance AI agents' autonomous learning capabilities, optimizing their performance in medical education, patient education, and long-term self-learning.

Table 2: Key Challenges and Future Directions for different core components and applications in AI Hospital.

Study	Methodology	Key Contribution
	Patient Agents	
[92]	Constructs patient agents using MedDialog-based structured dialogue strategies and fine-tunes Qwen2.5-72B-Instruct via LoRA.	Enhances realism in patient-agent interactions by reducing hallucination rates and improving adaptation to medical contexts.
[85]	Develops the CureFun framework with prompt design, chain-of-thought reasoning, and RAG to optimize LLM-based virtual patient simulations.	Improves role consistency and stability in patient-agent simulations, supporting cost-effective medical training.
[54]	Implements prompt-based optimization to enhance GPT's ability to role-play as a simulated patient for history-taking training.	Demonstrates that LLMs can serve as interactive patient agents, improving medical students' communication skills.
[105]	LLM-powered patient agents and their integration with NLP, OSCEBot, and SOCRATES-based symptom simulation.	Highlights the role of LLM-based patient agents in scalable, interactive medical education and standardized clinical assessment.
[164]	Introduces AIPatient, integrating a knowledge graph and RAG to improve the accuracy and personalization of patient interactions.	Achieves high accuracy in EHR-based medical Q&A while improving adaptability to diverse patient needs.
[37]	Proposes EvoPatient, using evolutionary multi-turn dialogues for unsupervised learning to optimize patient-agent behavior.	Enhances patient-agent realism by improving standardized medical communication through iterative self-improvement.
[88]	Develops the AIE framework with a state-aware patient simulator (SAPS) to dynamically assess LLMs in clinical dialogues.	Establishes a structured, scalable method to evaluate LLMs' clinical decision-making through realistic multi-turn interactions.
[63]	Introduces the CRAFT-MD framework, combining controlled patient simulations with AI and expert-driven assessments.	Enhances LLM evaluation in patient-centered clinical reasoning and diagnostic tasks.
[28]	Designs a personalized MDT framework leveraging LLM-driven patient agents to simulate rare disease cases.	Enables AI-assisted rare disease diagnosis and interdisciplinary medical discussions.
[140]	Implements TWIN-GPT digital twins to model individual patient disease progression and treatment response.	Advances patient-specific simulation for personalized clinical research and trials.
[113]	Uses symptom-based patient agents that interact with doctor agents by revealing symptoms without knowing the diagnosis.	Supports AI-driven medical assessments by simulating realistic patient uncertainty in clinical encounters.
[11]	Generates patient personas using GPT-4o based on personality traits and demographic factors to enhance communication realism.	Creates diverse, behaviorally accurate patient agents to improve medical interaction training.
[40]	Develops patient agents with predefined communication traits (cooperativeness, curiosity, personalization) to enhance realism.	Ensures authentic patient behaviors in AI-driven clinical interactions.
[156]	Implements ClinicalAgent, where patient agents engage in full diagnostic workflows with multi-agent interaction.	Facilitates AI-driven end-to-end medical consultation modeling.
[84]	Designs MediQ, a patient system interacting with expert systems in structured medical settings.	Supports AI-driven clinical reasoning by providing accurate, structured patient responses.
[135]	Uses the NoteChat framework to model LLM-based patient agents with realistic conversational styles.	Improves AI-assisted patient communication by simulating varied linguistic and interaction patterns.
[128]	Develops patient agents that roleplay specific medical conditions and actively engage with doctor agents.	Enhances realism in AI-driven medical scenario training by incorporating patient-initiated interactions.
	Psychological Patient Agents	
[93]	Develops a human-in-the-loop pipeline where domain experts provide qualitative feedback to guide LLM-generated psychological patient responses.	Enhances LLM-based psychological patient realism by improving adherence to expert-defined behavioral guidelines, making them more suitable for therapist training.
[136]	Introduces PATIENT-Ψ, a psychological patient agent based on cognitive behavioral therapy (CBT), integrating cognitive models with LLMs to simulate realistic patient interactions.	Provides a structured CBT training tool that improves trainees' clinical skills and confidence through interactive, lifelike patient simulations.
[133]	Proposes the ClientCAST framework, where LLMs simulate therapy clients and provide self-reported assessments of therapy sessions to evaluate LLM therapists.	Establishes a patient-centered evaluation method for AI therapists, enabling cost-effective, scalable assessments while maintaining ethical safety.
[107]	Constructs psychologically diverse client personas from PsyQA datasets, enabling multi-turn interactions with AI therapists for mental health training.	Creates a standardized pool of simulated therapy clients, supporting AI-driven evaluations and therapist training.
[143]	Implements MEDCO patient agents that simulate a variety of symptoms and mental health conditions to engage with medical trainees.	Enhances medical education by providing realistic, interactive psychiatric training scenarios.
[27]	Designs a patient chatbot that exhibits realistic psychiatric behaviors, including symptom honesty, colloquial expression, emotional fluctuations, and resistance.	AI-driven psychiatric simulations by capturing key psychological traits to enhance mental health training realism.
	Resident Agents	
[81]	Develops a dynamic agent-based simulation where Resident Agents start as healthy individuals and transition into Patient Agents upon illness, following a structured hospital navigation process.	Provides a comprehensive, adaptive simulation of real-world healthcare-seeking behaviors, enabling realistic assessments of disease management and hospital workflow.
[145]	Uses generative agents with distinct personalities and behavioral patterns to simulate decision-making in disease outbreaks, considering personal health conditions and social health information.	Models the impact of individual and societal health awareness on medical-seeking behaviors, supporting AI-driven public health simulations.

Table 3: Summary of Studies on Patient, Psychological Patient, and Resident Agents

Study	Methodology	Key Contribution
	Research Planning Agent	
[122]	Utilizes a Principal Investigator Agent (PI Agent) to guide research projects, maximize scientific impact, and coordinate other agents within the team.	Proposes a multi-agent collaboration framework that enables LLMs to autonomously direct and optimize research workflows, enhancing intelligent research management.
[148]	Employs a hierarchical decision-making framework to analyze user requirements and plan the full scRNA-seq data analysis workflow, including preprocessing, quality control, and analytical steps.	Develops an intelligent planning system that allows AI to understand and construct bioinformatics workflows like human experts, improving automation in data-driven research.
[91]	Uses a two-stage approach (idea generation + idea optimization) to iteratively refine and optimize candidate solutions, ensuring optimal performance for ML tasks.	Systematically manages and optimizes the "idea space" to enhance AI decision-making in drug discovery, reducing infeasible solutions and improving experimental efficiency and success rates.
	Research Executor Agents	
[141]	Integrates LLMs with code generation, Chain-of-Thought (CoT), and Self-Reflection techniques to assist with data analysis, hypothesis testing, and result interpretation.	Reduces the programming burden for medical researchers by optimizing research workflows, improving analysis accuracy, and enabling interactive refinement of AI-generated code.
[122]	Develops specialized scientist agents for different expertise areas (e.g., biology, computer science) to support clinical research, hypothesis testing, and system performance evaluation.	Facilitates interdisciplinary collaboration by incorporating domain-specific AI agents that enhance clinical study execution and analysis.
[148]	Implements a hierarchical execution system where low-level executors systematically perform decomposed sub-tasks, ensuring accurate execution of scRNA-seq data analysis.	IEnhances reliability in biomedical data processing by leveraging AI-driven execution agents capable of understanding and troubleshooting domain-specific computational tools.
[91]	Uses a structured process involving task decomposition, knowledge requirement identification, tool construction, and reuse to support domain-specific ML tasks in drug development.	Improves tool reliability and reduces coding errors by incorporating structured reasoning, automated tool verification, and reusable knowledge resources for AI-assisted research execution.
	Scientific Critic Agents	
[148]	Implements an iterative self-optimization mechanism where an Evaluator assesses the quality of data processing results and refines solutions through hyperparameter tuning and tool selection.	Enhances the reliability of AI-driven research by enabling automated evaluation and optimization, ensuring high-quality solutions comparable to expert assessments.
	Database Agents	
[11]	Uses natural language processing (NLP) to extract structured parameters from user queries and interact with the Hospital Information System (HIS) via API calls for patient record management and administrative information retrieval.	Enhances hospital database accessibility by enabling AI-driven interaction with HIS, automating patient record management, and supporting administrative queries through NLP-based structured retrieval.
[114]	Integrates information from multiple sources, including indexed medical databases and online resources, to enrich clinical trial criteria with relevant domain-specific knowledge.	Improves the accuracy and relevance of clinical trial data retrieval by augmenting queries with external and internal medical knowledge sources.
[166]	Reformulates user intent and retrieves relevant medical information from external knowledge bases to support clinical decision-making.	Enhances the precision of medical information retrieval by dynamically adapting search strategies based on reconstructed user queries.
[115]	Identifies and retrieves structured medical data, such as tables and records, to support accurate clinical reasoning and decision-making.	Streamlines medical data integration by facilitating structured retrieval and reasoning, improving the efficiency of AI-assisted healthcare decision-making.
[87]	Clusters, classifies, and stores medical guidelines as structured data for efficient retrieval and relevance evaluation.	Enhances guideline-based decision support by structuring and organizing medical knowledge for fast and relevant information retrieval in clinical settings.

Table 4: Summary of Studies on AI-Assisted Research Agents

Study	Methodology	Key Contribution
	General Doctor Agents	
[92]	Analyzes different LLM-based doctor models' inquiry strategies in five-round patient interviews.	Identifies how balanced inquiry strategies improve diagnostic accuracy, with GPT-4o performing best in comprehensive patient assessments.
[37]	Introduces the EvoPatient framework, where PCP agents evolve inquiry strategies through multi-turn dialogue learning.	Enhances patient interaction quality and diagnostic accuracy, particularly in multi-disciplinary consultations.
[63]	Develops the CRAFT-MD framework for evaluating LLM-based medical reasoning in interactive clinical dialogues.	Establishes a structured assessment for LLM-based PCPs, revealing limitations in multi-turn inquiry and diagnostic reasoning.
[28]	Proposes RareAgents, a multi-agent framework where PCPs conduct initial assessments and coordinate with specialists for rare disease diagnosis.	Improves diagnostic accuracy for rare diseases by integrating PCPs with expert-driven multi-agent decision-making.
[28]	Implements DrHouse, an LLM-based virtual PCP that actively collects patient symptoms and sensor data for diagnosis.	Enhances AI-driven medical consultations by integrating real-time patient data with iterative diagnostic refinement.
[68]	Introduces MDAgents, an adaptive framework where PCP agents handle routine cases while escalating complex ones to specialists.	Optimizes workload distribution between generalist and specialist AI doctors, improving system efficiency.
[113]	Simulates doctor agents with constrained inquiry capabilities to assess their ability to gather information for diagnosis.	Models real-world diagnostic constraints to evaluate LLM-based doctors' effectiveness under limited questioning scenarios.
[132]	Uses open-source LLMs as PCPs in an MCQA framework, generating probabilistic disease diagnoses from symptom inputs.	Demonstrates that PCP LLMs provide broad but shallow diagnostic capabilities, highlighting the need for specialist integration.
[40]	Designs AI-driven doctor agents that actively collect subjective and objective patient data for informed diagnosis.	Improves AI doctors' realism by simulating experienced clinicians' structured diagnostic workflows.
	Specialist Agents	
[37]	Uses the EvoPatient framework, where specialist agents dynamically recruit multi-disciplinary experts to optimize diagnostic precision and treatment strategies.	Enhances LLM-driven specialist agents' ability to handle complex medical cases and improve diagnostic reasoning in multi-disciplinary settings.
[28]	Implements an MDT framework where pre-defined specialist agents collaborate through structured discussions to reach consensus on diagnosis and treatment.	Demonstrates that expert-defined role structures improve diagnostic accuracy and decision reliability in multi-specialty medical consultations.
[68]	Introduces a recruitment-based system where specialist agents are selectively engaged based on case complexity, collaborating under a structured decision-making process.	Optimizes the adaptive deployment of medical specialists, ensuring expert-level care for complex cases while maintaining efficiency for routine diagnoses.
[124]	Develops the MEDAGENTS framework, where specialist agents independently assess patient conditions and collaborate via an MDT agent to consolidate medical decisions.	Establishes a structured, multi-expert system that enhances diagnostic robustness and consensus-driven decision-making.
[142]	Implements DoctorAgent within the ColaCare framework, leveraging specialist models trained on EHR data and medical literature for in-depth clinical analysis.	Strengthens the reliability of multi-specialist consultations by integrating real-world clinical knowledge and evidence-based recommendations.
[132]	Deploys specialist agents trained on NIH medical knowledge, each focused on a specific disease domain, to improve targeted diagnostics.	Demonstrates that specialist-trained LLMs outperform generalist models in disease-specific diagnostic tasks, particularly in recall and precision.
[81]	Simulates 14 internal medicine specialists, each responsible for diagnosing diseases and designing treatment plans in a hospital setting.	Provides a structured, multi-specialty simulation for realistic disease management and treatment strategy development.
[156]	Uses multi-specialist consultation, where agents (e.g., pediatricians, orthopedic specialists) collaborate through iterative diagnostic and testing processes.	Enhances multi-specialty coordination by systematically integrating lab and imaging results into decision-making.
[84]	Implements an expert system where specialist agents engage in structured patient interviews, hypothesis testing, and iterative decision-making.	Improves clinical reasoning accuracy by modeling expert physician decision processes.
[135]	Designs specialist doctor agents with predefined linguistic and professional styles for realistic medical role-play in the NoteChat framework.	Ensures high-quality, structured medical dialogues that enhance interactive patient-agent training.
[128]	Develops an empathetic specialist doctor agent for online consultations, focusing on history-taking, diagnosis, and treatment planning.	Enhances AI-driven telemedicine by providing realistic, patient-centered diagnostic support.
	Therapist Agents	
[133]	Develops the ClientCAST framework, where LLM-simulated clients interact with LLM therapists, and therapy effectiveness is evaluated via standardized questionnaires.	Establishes a structured, cost-effective evaluation method for AI therapists, highlighting their strengths and limitations in therapeutic alliance and emotional response.
[107]	Implements a therapist agent based on an integrative therapy model, guiding psychological counseling through three structured phases: exploration, insight, and action.	Provides a structured, process-driven AI therapist framework that mimics real-world psychotherapy workflows.
[27]	Designs a psychiatrist chatbot that conducts structured diagnostic interviews, emphasizing symptom comprehensiveness, deep inquiry, and emotional support.	Enhances AI-driven psychiatric support by combining diagnostic precision with empathetic conversational strategies.
	Nurse Agents	
[11]	Develops a nurse agent that follows a three-step process (action decision-making, reflective feedback, response generation) to assist with triage, administrative tasks, and patient support.	Enhances outpatient service efficiency and patient experience by standardizing nursing workflows and providing personalized guidance.
[81]	Simulates four nurse agents responsible for triage and daily treatment interventions, ensuring patients follow appropriate medical pathways.	Supports AI-driven clinical workflow optimization by efficiently directing patients into the correct care process.
	Medical Technician Agents	
[113]	Implements a measurement agent that generates predefined medical test results based on patient conditions and scenario templates.	Ensures accurate and context-aware diagnostic testing, supporting realistic clinical decision-making for AI doctors.
[143]	Designs an LLM-based radiologist agent capable of interpreting multiple imaging modalities (X-ray, CT, MRI, ultrasound) and generating structured radiology reports.	Enhances AI-driven diagnostic imaging interpretation by integrating professional radiology knowledge with automated reporting tools.
[81]	Simulates a radiologist agent specializing in medical imaging interpretation within a hospital environment.	Provides AI-driven radiology expertise, improving diagnostic precision in clinical simulations.
[40]	Develops an examiner agent that systematically processes medical test requests, validates them, and provides structured diagnostic results.	Standardizes medical testing workflows to enhance the realism and reliability of AI-driven clinical diagnostics.
[156]	Integrates specialized laboratory and radiology technician agents into the ClinicalAgent framework, providing laboratory diagnostic results (LDR) and imaging diagnostic results (IDR).	Supports multi-specialty medical decision-making by enabling seamless integration of laboratory and imaging data.
	Medical Student and Examiner Agents	
[85]	Develops the CureFun framework, integrating structured case graphs, retrieval-augmented generation (RAG), and automated scoring to train medical student agents and assess their clinical reasoning.	Enhances objective, scalable medical training by simulating realistic patient interactions and providing automated expert-level evaluations.
[105]	Implements medical student and examiner agents using NLP techniques (e.g., SBERT) and the SOCRATES framework, enabling structured training and evaluation in history-taking and clinical reasoning.	Improves cost-effective, personalized medical education by offering structured guidance and AI-driven performance assessments.
[143]	Uses advanced prompt strategies to simulate expert evaluators who assess medical students' diagnostic skills, patient interactions, and treatment planning.	Enhances AI-assisted medical education by enabling interactive, case-based clinical training with structured AI feedback.
[163]	Develops MedQA-CS, where LLM-as-Medical-Student (MedStuLLM) interacts with simulated patients, while LLM-as-Examiner (MedExamLLM) evaluates clinical reasoning.	Establishes a benchmark for AI-driven medical training and evaluation, advancing LLM-based assessment frameworks for clinical education.

Table 5: Summary of Studies on Medical Professional Agents

Study	Methodology	Key Contribution
	Goal-Driven Reasoning Agent	
[164]	Establishes a sequential and collaborative information processing pipeline with retrieval, abstraction, query generation, and response rewriting agents.	Ensures seamless transformation of user queries into structured knowledge retrieval and patient-friendly responses, optimizing information flow in AI-patient interactions.
[65]	Applies logical reasoning to generate and refine differential diagnoses based on initial assessments and team discussions.	Enhances diagnostic decision-making by ensuring a structured and iterative refinement of hypotheses in clinical scenarios.
[114]	Uses self-reflection techniques to identify missing information in trial standards and assess LLMs' understanding of clinical trials and patient characteristics.	Improves clinical reasoning by detecting knowledge gaps and ensuring detailed and coherent trial standard interpretation.
[55]	Constructs an abstract argumentation framework through iterative reasoning, identifying semantic relations (attack/support) among arguments.	Strengthens AI decision-making by systematically validating medical arguments using a symbolic solver for robust reasoning.
[166]	Synthesizes retrieved knowledge and factual evidence to generate responses, similar to expert agents aiding in diagnosis, treatment, and evaluation.	Ensures AI-driven medical reasoning aligns with expert-level clinical support, improving accuracy and reliability in patient interactions.
[115]	Uses LLMs and code execution for multi-turn interactive feedback to refine task planning and execution.	Enhances AI-driven reasoning through continuous execution feedback, improving adaptability in medical simulations.
[87]	Integrates medical guidelines and retrieval-based LLM reasoning to generate expert-level diagnostic and treatment responses.	Provides a knowledge-driven medical QA system that mirrors expert decision-making in clinical support tasks.
[80]	Employs continuous reasoning, integrating patient information and medical knowledge to iteratively refine diagnostic hypotheses.	Facilitates AI-driven medical diagnosis by maintaining a goal-oriented reasoning process for precise rare disease identification.
	Clinical Judge Agent	
[63]	Uses interactive AI simulations and automated grading agents to evaluate LLMs' diagnostic reasoning and history-taking abilities in controlled medical dialogues.	Establishes a structured evaluation framework that highlights LLMs' limitations in medical reasoning, emphasizing the importance of multi-turn interaction, information synthesis, and expert-guided assessment.
[113]	Parses unstructured diagnostic statements from doctor agents and determines their correctness against ground truth labels.	Enhances diagnostic evaluation by resolving inconsistencies in free-text diagnoses, ensuring accurate assessment of AI-generated clinical reasoning.
[165]	Retrieves biomedical knowledge from DrugBank and HetioNet to evaluate drug effectiveness based on biological interactions and clinical evidence while analyzing pharmacological data and clinical trial history to assess potential risks and adverse effects.	Supports AI-based clinical decision-making by providing structured pharmacological insights on drug efficacy while enhancing patient safety through statistical risk modeling and historical failure analysis.
[114]	Ensures consistency between AI-augmented clinical standards and original medical guidelines, maintaining adherence to predefined instructions.	Strengthens AI-driven clinical standardization by enforcing alignment with medical protocols, reducing deviations in automated clinical workflows.
[166]	Identifies and extracts relevant factual evidence from knowledge sources at both document and sentence levels.	Improves AI-assisted clinical validation by filtering out irrelevant information, enhancing trustworthiness and interpretability in medical reasoning.
[40]	Evaluates doctor agents' diagnostic performance by comparing their summarized reports against gold-standard medical records.	Ensures AI-assisted diagnoses align with expert medical standards, refining clinical judgment in AI-driven patient interactions.
[128]	Continuously monitors and evaluates doctor-patient interactions, determining the natural conclusion of the dialogue.	Enhances AI-facilitated medical consultations by assessing the coherence and completeness of AI-driven diagnostic discussions.
	Critic Agent	
[65]	Conducts critical evaluations of initial diagnoses, identifying inconsistencies, proposing alternative differential diagnoses, and mitigating cognitive biases (e.g., confirmation bias, anchoring bias).	Enhances diagnostic accuracy by iteratively refining clinical reasoning, reducing premature closure and knowledge biases in AI-assisted medical decision-making.
[122]	Critically reviews and improves scientific reasoning within a multi-agent virtual laboratory setting.	Elevates research quality by ensuring robust scientific critique and continuous refinement of AI-generated hypotheses.
[55]	Uses self-questioning techniques to validate newly generated arguments, iterating corrections until logically sound justifications are reached.	Improves the reliability of AI-driven reasoning by enforcing rigorous verification and iterative argument refinement.
[115]	Iteratively analyzes errors using an AI feedback system inspired by rubber-duck debugging and self-play techniques.	Optimizes AI reasoning performance through automated self-critique and iterative debugging, ensuring error reduction over successive iterations.
[87]	Assesses response accuracy and cross-references medical guidelines to validate information credibility.	Strengthens trust in AI medical responses by enforcing evidence-backed assessments and ensuring adherence to authoritative guidelines.
[128]	Implements a self-play framework where an LLM critic iteratively evaluates and refines diagnostic dialogue responses based on real-case context.	Enhances conversational AI's diagnostic reasoning by improving response quality, patient engagement, and clinical accuracy through iterative self-improvement.
[80]	Challenges diagnostic reasoning through active questioning and critique, prompting AI doctor agents to reassess assumptions and refine conclusions.	Strengthens AI-driven clinical discussions by ensuring diagnoses undergo multi-perspective validation, reducing logical gaps and improving diagnostic precision.
	Planner Agent	
[165]	Uses a LEAST-TO-MOST reasoning approach to systematically break down complex medical problems into smaller, manageable subproblems, assigning them to the most suitable expert agents; applies Few-Shot learning to refine task allocation in clinical settings.	Enhances workflow efficiency by ensuring precise problem decomposition and targeted task distribution, optimizing clinical trial execution and medical decision-making.
[114]	Selects the most appropriate enhancement agents to enrich trial standards based on domain-specific knowledge, dynamically routing tasks accordingly.	Improves workflow adaptability by intelligently coordinating specialized agents for task augmentation, ensuring comprehensive and efficient trial standard enhancement.
[166]	Clarifies user queries by integrating contextual cues, identifying key questions, unifying task formats, and decomposing multi-step queries for structured information retrieval.	Enhances medical AI comprehension by refining ambiguous user instructions into well-defined queries, enabling accurate and efficient access to relevant knowledge.
[135]	Structures doctor-patient interactions to ensure logical and efficient knowledge organization while minimizing hallucinations.	Optimizes medical conversations by improving the coherence and accuracy of information exchange, leading to better decision-making in AI-assisted patient interactions.
[68]	The Moderator Agent acts as a GP or emergency physician, evaluating the complexity of medical queries and directing cases to the appropriate handling pathway (single agent, MDT, or ICT), while the Recruiter Agent assembles specialized expert teams accordingly.	Ensures efficient triage by dynamically assigning medical queries to the appropriate expertise level, optimizing resource allocation and clinical decision-making.
[124]	Assembles multidisciplinary expert teams based on the clinical nature of the problem.	Enhances triage efficiency by ensuring complex cases receive specialized expertise from diverse medical disciplines.
[94]	Uses a hierarchical classification system where Doctor A performs coarse-grained triage (high vs. low urgency), Doctor B refines the classification, and Doctor C optimizes triage levels through multi-round discussions, integrating RAG-based dynamic confidence scoring.	Improves triage accuracy by leveraging multi-stage classification, retrieval-enhanced decision-making, and cost-effective prioritization strategies.
[156]	Uses patient complaints and prior LLM ranking knowledge to assign cases to the most relevant clinical departments, ensuring timely and accurate specialist consultations.	Enhances initial triage efficiency by rapidly identifying patient needs and directing them to the appropriate medical teams for further diagnosis and treatment.
	Decision Agent	
[124]	Uses consensus-based reports to derive final clinical decisions, ensuring alignment with validated medical assessments.	Strengthens AI-assisted medical decision-making by enforcing consistency and reliability in deriving conclusive clinical answers.
[142]	Integrates and reconciles multiple doctor agents' clinical evaluations, analyzing conflicting information, refining reports, and determining whether to continue or finalize discussions.	Enhances collaborative AI decision-making by mediating diverse expert opinions, ensuring a scientifically robust and comprehensive final diagnosis.
[165]	Uses a hierarchical Transformer-based model to predict enrollment success in clinical trials based on predefined inclusion criteria.	Optimizes clinical trial decision-making by ensuring adequate participant recruitment, improving statistical power and intervention evaluation reliability.
	Recorder Agent	
[164]	Integrates rewritten responses with dialogue content and continuously updates conversation history for multi-turn interactions.	Ensures accurate documentation of patient-provider interactions, maintaining coherence and traceability in medical dialogues.
[65]	Extracts key discussion points, compiles the final differential diagnosis list, and highlights critical learning insights from collaborative efforts.	Enhances diagnostic traceability and knowledge retention by systematically summarizing and structuring medical discussions.
[11]	Monitors conversation quality and information completeness, using memory storage, extraction, and comparison to ensure comprehensive data collection.	Improves clinical documentation by identifying missing patient information and guiding further data collection for accurate decision-making.
[135]	Refines generated dialogue data to enhance logical consistency and adherence to medical reasoning principles.	Ensures high-quality, medically coherent documentation by improving the clarity and accuracy of AI-generated medical dialogues.

Table 6: Summary of Studies on Medical AI Teamwork Agents

Study	Methodology	Key Contribution
[92]	Decomposes the consultation process into diagnosis and inquiry handled by separate models.	Demonstrates that splitting inquiry and diagnosis into specialized LLMs improves medical consultation efficiency.
[85]	Uses entity recognition, relation extraction, and attribute extraction to build structured case graphs.	Ensures structured and context-aware data processing for AI-driven medical dialogue systems.
[105]	Implements modular chatbot architecture with a knowledge base, conversation module, and API-driven interactions.	Facilitates efficient and structured information retrieval and response generation.
[93]	Transforms expert feedback into structured guidelines for automated question-answering.	Enhances medical chatbot accuracy through structured, rule-based decision-making.
[164]	Assigns specialized agents for knowledge retrieval, query generation, and response adaptation.	Ensures context-aware patient interaction by leveraging structured multi-agent collaboration.
[88]	Splits patient simulation into state tracking, memory modules, and response generation.	Enables realistic and structured AI-driven patient simulation for diagnostic training.
[157]	Divides diagnostic workflows into knowledge retrieval, sensor data integration, and decision-making modules.	Improves diagnostic accuracy through specialized processing pipelines.
[94]	Implements a multi-agent triage workflow for document classification, summarization, and consensus-building.	Ensures structured and explainable triage decisions via role-specific agents.
[165]	Uses hierarchical problem decomposition and assigns specialized agents for different clinical evaluations.	Enhances diagnostic accuracy by breaking complex queries into manageable subproblems.
[114]	Splits knowledge discovery, fact verification, and patient matching into specialized modules.	Improves decision accuracy by structuring complex medical knowledge workflows.
[55]	Uses distinct agents for argument generation, verification, and symbolic reasoning.	Enables structured, evidence-based clinical decision-making via multi-agent collaboration.
[166]	Implements intent recognition, knowledge retrieval, fact validation, and response generation agents.	Ensures contextual accuracy and reliability in AI-driven medical question-answering.
[115]	Uses modular agents for knowledge integration, interactive coding, and debugging.	Enhances AI-driven clinical tool reliability through structured execution and feedback loops.
[87]	Assigns agents for medical guideline classification, retrieval, evaluation, and response generation.	Improves adherence to medical guidelines through structured, multi-agent query handling.
[135]	Decomposes dialogue planning into structured phases: high-level planning, role-based dialogue, and iterative refinement.	Enhances doctor-patient AI interactions by ensuring structured and contextually relevant dialogue generation.

Table 7: Summary of Studies on Task-Focused Collaboration in AI Hospitals

Study	Methodology	Key Contribution
[93]	Experts provide qualitative feedback, which is transformed into behavioral rules for response generation.	Ensures AI responses align with expert-defined behavioral guidelines, improving reliability.
[37]	Doctor agents receive structured records and multidisciplinary expert inputs to refine their questions.	Enhances diagnostic accuracy by integrating expert-designed guidance in question formulation.
[28]	A pool of 41 clinical specialists engage in multi-round discussions to provide expert insights.	Ensures accurate and consensus-driven medical decisions through expert-guided discussions.
[68]	A moderator agent assigns cases to appropriate expert teams based on complexity	Optimizes resource allocation by dynamically involving experts in decision-making.
[124]	Domain-specific expert agents analyze and discuss medical problems before making a final decision.	Improves decision robustness by leveraging expert consensus across disciplines.
[142]	Individual DoctorAgents provide initial assessments, while a MetaAgent consolidates and verifies outputs.	Enhances diagnostic reliability through expert-level validation and oversight.
[65]	Senior doctors supervise junior doctors, identifying cognitive biases and guiding refinements.	Reduces diagnostic errors by incorporating expert oversight in medical training simulations.
[122]	PI agents guide research discussions, while Scientific Critic agents validate expert opinions.	Ensures scientifically sound decision-making by incorporating expert scrutiny.
[143]	AI-based expert agents evaluate student diagnoses, providing feedback and knowledge summaries.	Enhances medical training by integrating expert-led assessment and iterative improvement.

Table 8: Summary of Studies on Expert-Guided Decision-Making in AI Hospitals

Study	Methodology	Key Contribution
[92]	Iteratively refines a robust test set and prompt engineering to minimize discrepancies in model evaluation.	Ensures consistency between AI and human evaluations by continuously improving diagnostic assessments.
[93]	Uses iterative self-criticism and expert-guided feedback to refine AI-generated responses.	Improves AI-patient interaction quality through repeated feedback-driven response adjustments.
[164]	Implements recursive query rewriting until accurate and satisfactory outputs are obtained.	Enhances system robustness by iteratively refining knowledge retrieval queries.
[37]	Stores and reuses high-quality interaction trajectories to iteratively refine AI-generated medical questions.	Enables adaptive learning, improving the efficiency and precision of doctor-agent questioning.
[28]	Engages AI specialists in iterative discussions to refine and optimize medical decisions.	Ensures high-quality medical recommendations through structured feedback loops.
[157]	Continuously updates diagnostic decision trees based on new patient symptoms in multi-turn dialogues.	Enhances diagnostic accuracy by dynamically refining medical knowledge with each patient interaction.
[68]	Facilitates iterative expert discussions, adjusting decisions dynamically based on new feedback.	Improves decision reliability through structured multi-agent feedback mechanisms.
[124]	Uses iterative voting and feedback loops to refine expert-approved clinical reports.	Enhances medical decision quality by ensuring expert consensus via multiple discussion cycles.
[142]	Doctor agents critique and revise MetaAgent-generated reports until consensus is reached.	Strengthens diagnostic conclusions through multi-round expert feedback integration.
[65]	Junior doctors iteratively reconsider diagnoses based on repeated analysis and expert feedback.	Improves diagnostic training outcomes through structured, feedback-driven optimization.
[94]	Agents dynamically adjust classification confidence based on peer reasoning and discussion.	Enables consensus-building through iterative debate and self-correction among medical AI agents.
[11]	AI nurse agents perform action-reflection-response cycles for continuous improvement.	Enhances AI responsiveness by refining decisions based on real-time feedback.
[141]	Captures error logs, analyzes failure cases, and iteratively refines AI-generated code.	Improves code reliability through structured self-reflection and multi-step correction.
[122]	PI agents iteratively refine scientific discussions based on multi-round critiques.	Ensures high-quality research decisions through iterative expert review.
[143]	AI medical students refine diagnoses based on multi-round feedback and case adjustments.	Enhances AI-driven medical training by optimizing learning through iteration.
[117]	AI agents iteratively refine responses through structured argumentation and counterpoints.	Improves AI reasoning accuracy through competitive self-improvement cycles.
[114]	AI agents evaluate their own responses and request additional refinement when necessary.	Strengthens AI decision-making through iterative, self-guided quality checks.
[55]	AI agents generate, critique, and refine hypotheses until coherent arguments are formed.	Ensures argument consistency and logical soundness through iterative self-improvement.
[115]	LLM agents refine code through multiple rounds of execution-feedback loops.	Enhances AI coding accuracy through iterative plan refinement.
[91]	Generates candidate solutions and iteratively refines them based on real-world execution feedback.	Improves planning efficiency by eliminating ineffective ideas and refining viable ones.
[81]	AI doctors reflect on errors, formulate reusable principles, and iteratively refine diagnoses.	Strengthens medical AI's diagnostic ability through systematic experience-based learning.
[40]	Doctor agents continuously critique and refine diagnostic decisions until consensus is reached.	Enhances diagnostic accuracy through structured argumentation and iterative discussion.
[84]	AI experts assess information sufficiency and refine knowledge through iterative questioning.	Ensures precise medical conclusions by dynamically refining diagnostic inquiries.
[135]	AI refines generated dialogues through multi-turn self-assessment and iterative improvement.	Improves AI conversational quality by systematically re-evaluating responses.
[128]	AI doctors receive critique, refine responses, and restart consultations iteratively.	Enhances diagnostic accuracy through structured feedback-driven self-improvement.

Table 9: Summary of Studies on Iterative Problem Optimization (IPO)

Study	Methodology	Key Contribution
[164]	Aggregates structured queries, user queries, and knowledge graph data to generate coherent patient responses.	Enhances clinical decision-making by fusing multiple data sources into a unified, context-aware response.
[37]	Automatically stores, validates, and reuses dialogue knowledge through structured libraries.	Improves AI reasoning by enabling dynamic knowledge accumulation and refinement across agents.
[88]	Organizes patient data into long-term, working, and short-term memory for structured response generation.	Enhances AI-patient interaction by integrating historical and real-time patient data dynamically.
[63]	Condenses multi-turn AI-patient conversations into structured summaries for efficient knowledge retention.	Improves AI comprehension by automatically integrating key discussion points into a unified context.
[28]	Merges expert discussions, memory retrieval, and medical tool feedback into final AI decisions.	Strengthens AI-driven diagnostics by consolidating expert insights and real-time data.
[157]	Automatically fuses sensor data, medical knowledge, and patient history into AI-generated diagnoses.	Enables comprehensive and informed decision-making by unifying multimodal medical data sources.
[68]	Uses a multi-stage approach to aggregate expert analyses into a consolidated report.	Improves clinical accuracy by systematically merging insights from diverse expert teams.
[124]	Combines question analysis (QA) and option analysis (OA) results into structured diagnostic reports.	Enhances AI decision-making by unifying disparate analytical outputs into a coherent summary.
[142]	Aggregates DoctorAgent assessments and integrates structured EHR data with medical language models.	Improves AI diagnostics by blending structured and unstructured medical data sources.
[65]	Extracts and synthesizes discussion points from multiple doctor agents into final diagnosis conclusions.	Ensures comprehensive and structured diagnostic summaries through automated data fusion.
[113]	Merges patient history, symptom analysis, and lab results into a unified clinical evaluation.	Enhances decision accuracy by synchronizing diverse patient data streams dynamically.
[94]	Integrates classification results, confidence scores, and reasoning outputs into a consolidated report.	Improves AI reliability by synthesizing agent-driven insights into structured decision frameworks.
[11]	Queries hospital information systems (HIS) and patient records, merging findings into clinical documentation.	Enhances patient record accuracy by seamlessly integrating internal and external medical knowledge.
[122]	Merges expert opinions from parallel discussions into a single, optimized report.	Improves medical research AI outputs by systematically combining multidisciplinary insights.
[117]	Merges reasoning trajectories across agents to refine AI-generated conclusions.	Enhances logical coherence by dynamically integrating AI-driven inference strategies.
[132]	Uses self-attention mechanisms to aggregate expert agent diagnostic probabilities.	Improves diagnostic robustness by dynamically synthesizing expert recommendations.
[165]	Consolidates findings from multiple specialized agents into a unified response.	Enhances decision accuracy by systematically integrating domain-specific insights.
[114]	Uses dedicated agents for retrieval, online search, and self-enhancement to refine trial standards.	Ensures clinical trial reliability by automatically validating and merging diverse knowledge sources.
[55]	Structures AI-generated arguments into a unified abstract argumentation framework.	Strengthens AI reasoning by integrating self-contradictory and supporting arguments dynamically.
[166]	Aggregates outputs from intent reconstruction, knowledge retrieval, and fact verification agents.	Ensures factual accuracy by synthesizing structured agent-derived knowledge into responses.
[40]	Integrates doctor agent insights with patient and diagnostic data into a structured report.	Improves AI-assisted diagnosis by merging multi-source clinical evidence dynamically.
[156]	Merges lab diagnostics, imaging results, and medical evaluations into structured treatment plans.	Enhances AI medical consultation quality by synthesizing diverse diagnostic inputs.

Table 10: Summary of Studies on Automated Knowledge Integration (AKI)

Study	Methodology	Key Contribution
[37]	Implements a multi-agent medical dialogue system where patient and doctor agents engage in multi-turn conversations for diagnosis.	Demonstrates multi-agent collaboration in medical diagnosis, with AI agents simulating doctor-patient interactions and multidisciplinary teamwork.
[133]	Uses LLMs to role-play clients and therapists in mental health consultations, simulating real-world patient-therapist interactions.	Establishes a multi-agent framework for evaluating AI therapist performance in psychotherapy through client-therapist role-based coordination.
[88]	Evaluates medical LLMs through multi-round dialogues between doctor agents and patient simulators.	Develops a structured multi-agent evaluation framework where AI doctors and patient agents interact for clinical decision-making.
[63]	CRAFT-MD framework integrates multiple AI agents (clinical LLM, patient AI, and scoring AI) for medical dialogue reasoning evaluation.	Introduces a multi-agent approach for assessing clinical reasoning by simulating real-world doctor-patient consultations.
[28]	Constructs a multi-agent system where patient and doctor agents collaborate in clinical decision-making.	Models a real-world multidisciplinary consultation (MDT) scenario through AI-based role coordination.
[107]	Uses LLMs to role-play therapist-client interactions in mental health counseling across multiple dialogue rounds.	Demonstrates AI-driven multi-agent collaboration in psychotherapy, ensuring realistic and structured counselor-client interactions.
[142]	Defines doctor agents and meta-agents to simulate multidisciplinary case reviews, integrating EHR-based decision-making.	Enhances multi-agent clinical coordination by incorporating document retrieval, decision synthesis, and iterative refinement.
[65]	Implements a four-agent clinical team framework that mirrors human medical team interactions.	Captures task specialization in clinical teams, modeling real-world medical decision-making dynamics.
[113]	AgentClinic constructs a multi-agent environment where different agents play specialized roles in patient care.	Simulates comprehensive clinical workflows by assigning AI agents to distinct roles such as patient, physician, and evaluator.
[11]	Uses three AI agents (receptionist, patient, and supervisor) to recreate an outpatient clinical reception workflow.	Enhances medical role-based coordination by structuring multi-agent interactions for service flow simulation.
[143]	MEDCO framework employs AI agents to train medical students in multi-role, multi-specialty medical education scenarios.	Establishes a structured multi-agent educational system to simulate collaborative medical training.
[132]	Develops an agent-based multi-expert consultation framework for diagnostic decision-making.	Implements AI-driven role-based collaboration in multi-expert medical consultations.
[81]	Builds a virtual hospital where patient and medical AI agents dynamically interact across the entire clinical workflow.	Simulates the full hospital ecosystem, enabling AI agents to learn from iterative interactions and role-specific responsibilities.
[40]	AI Hospital framework assigns different NPC agents (patients, examiners, doctors) to structured medical dialogue roles.	Realistically models patient-centered care through structured multi-agent interactions.
[156]	ClinicalAgent decomposes clinical workflows into multi-stage tasks assigned to specialized AI agents.	Advances role-based coordination by integrating multi-agent decision-making into structured clinical pathways.
[27]	Develops two chatbot agents representing a psychiatrist and a patient for multi-turn psychiatric diagnosis.	Implements AI-driven mental health diagnostics through structured doctor-patient interactions.
[84]	MEDIQ framework simulates doctor-patient interactions using two AI agents with structured medical decision-making.	Enhances medical AI role-playing by structuring clinical reasoning into modular AI interactions.
[135]	Utilizes two LLM agents to simulate multi-turn physician-patient consultations with dynamic prompt control.	Ensures realistic AI-driven medical consultations through controlled role-based dialogue generation.
[163]	MedQA-CS employs AI agents to play medical students and examiners in clinical scenario-based evaluations.	Develops an AI-driven OSCE (clinical exam) simulation for standardized medical assessments.
[128]	Constructs a multi-agent system where AI models assume patient, doctor, moderator, and evaluator roles for medical diagnosis simulation.	Introduces an advanced AI-driven diagnostic simulation system leveraging structured multi-agent interactions.

Table 11: Summary of Studies on Role-based Coordination

Study	Methodology	Key Contribution
[68]	Implements an MDT model where AI agents exchange messages in multiple rounds, with a moderator agent resolving disagreements.	Establishes a structured debate process that enables AI agents to iteratively discuss and refine clinical decisions until consensus is reached.
[124]	Uses an expert consultation phase where agents debate and vote on proposed reports, iteratively refining conclusions.	Enables multi-agent deliberation through structured voting and iterative argumentation, ensuring robust expert consensus.
[94]	Employs a collaborative discussion stage where agents present classification results, reasonings, and confidence scores, stopping early if consensus is reached.	Introduces Byzantine consensus-inspired mechanisms to improve debate efficiency and ensure reliable decision-making.
[122]	Structures multi-agent discussions where specialized AI agents respond to agenda items, while a critique agent provides feedback before a PI agent synthesizes final conclusions.	Enhances scientific discourse in clinical decision-making through iterative argumentation and structured critique.
[117]	Develops a multi-agent debate (MAD) framework where different AI agents, including adversarial roles, engage in multi-round discussions to refine answers.	Implements structured opposition and argumentative refinement to enhance reasoning accuracy and knowledge validation.
[55]	ArgMed-Agents generate arguments, undergo validation challenges, and iteratively refine conclusions based on critique.	Establishes an AI-driven medical debate system where adversarial argumentation improves diagnostic accuracy.
[40]	Uses a collaborative diagnostic model where multiple physician agents debate diagnostic discrepancies under a central agent's moderation.	Simulates structured physician discourse, ensuring consensus-driven, high-accuracy diagnosis.
[80]	MAC framework enforces multi-round discussions among physician agents, where they critique each other's diagnostic reasoning, with a supervisor agent moderating.	Models real-world clinical discussions through structured debate, ensuring deep, multi-perspective diagnostic refinement.

Table 12: Summary of Studies on Multi-Round Interactive Debate.

Study	Methodology
	Retrieval Systems
[105]	Utilizes SBERT embeddings and cosine similarity to perform semantic search for retrieving the most relevant predefined questions or scripts.
[37]	Implements retrieval-augmented generation (RAG) to extract relevant information from stored records, ensuring context retention in long conversations.
[28]	Employs patient embedding retrieval from RareBench to dynamically retrieve the top-k most similar cases for diagnosis support.
[157]	Uses a Retriever model (based on OpenAI's text-embedding-ada-002) to fetch relevant knowledge from sensor data and medical textbooks.
[68]	Introduces MedRAG, a systematic RAG framework that retrieves biomedical, clinical, and general medical knowledge to enhance agent responses.
[142]	Implements a retrieval module to match patient records with relevant documents from a medical corpus for contextual decision support.
[113]	Adopts an adaptive RAG approach that enables doctor agents to retrieve clinical information from internet and textbook databases for medical research.
[94]	Incorporates RAG to fetch emergency severity index (ESI) handbook information, integrating it into LLM-generated clinical reports.
[141]	Uses RAG (via Vertex AI Search) to retrieve medical literature, software documentation, and code discussions to enhance clinical reasoning.
[143]	Employs a key-value memory system with ChromaDB and OpenAI embeddings to retrieve stored medical case knowledge for simulated training.
[114]	Leverages retrieval agents to extract indexed medical data, while online search agents supplement clinical trial standards with external sources.
[166]	Utilizes a Knowledge Retriever to fetch the most relevant documents from external medical databases based on intent reconstruction.
[78]	Uses ChatCAD+ to retrieve knowledge from the Merck Manual, enhancing response accuracy in MMedAgent's retrieval-augmented generation.
[1]	Extracts the latest medical data from literature and verified online sources to enhance clinical decision-making.
[121]	Implements PubMedBERT embeddings and Faiss indexing to retrieve relevant medical abstracts from 5.3 million PubMed articles for enhanced LLM responses.
	Other MedRAG-style retrieval tools [150, 152, 151, 167, 139, 87]
	Knowledge Graphs
[85]	Constructs structured "skeleton" and "case graphs" using named entity recognition (NER) and relationship extraction, storing medical entities and relationships in an RDF-based graph database for efficient querying.
[164]	Develops the AIPatient KG using NER and Neo4j to structure EHR-extracted medical entities and enables efficient querying via a reasoning-driven retrieval system.
[28]	Uses DDI-graph to model drug-drug interactions in a structured graph format, supporting medication recommendation tasks.
[165]	Integrates DrugBank for structured drug-target interactions and Hetionet for interconnected biomedical relationships, aiding in drug repurposing and disease analysis.
[1]	Incorporates external structured medical knowledge bases to enhance AI-driven clinical decision-making.
[135]	Utilizes MedSpaCy and QuickUMLS to extract and constrain dialogue generation around clinically relevant concepts, reducing hallucinations and ensuring medically accurate interactions.
	Other MedGraphRAG style retrieval [90, 146, 70, 95, 66, 46]
	Medical Decision Trees
[105]	Implements a confidence-based decision tree to determine whether a user query should be accepted or rephrased, ensuring accurate responses through rule-based branching logic.
[157]	Converts graphical diagnostic guidelines into an if-else structured decision tree using OCR, enabling step-by-step symptom-based diagnosis and treatment recommendations.
	Other studies discussing how to build or use Medical Decision Trees [77, 172]
	LLMs as Knowledge Bases
[165]	Utilizes LLMs as dynamic medical knowledge bases to extract, extrapolate, and generate novel biomedical insights, including hypothetical drug interactions and potential therapeutic targets.
[42]	Introduces MedGENIE, a generate-then-read framework where domain-specific LLMs construct artificial contexts through prompting, outperforming retrieval-based approaches in medical question answering.

Table 13: Summary of Studies about Retrieval Tools.

Study	Methodology
	Smart Devices & Sensor Data Integration
[157]	Integrates real-time physiological data from smart devices like Fitbit Sense, including step count, sleep score, SpO2, and heart rate, to provide objective health monitoring for clinical diagnosis.
[1]	Aggregates and processes medical data from EHRs, smartphones, and wearable devices to support comprehensive health monitoring and medical decision-making.
	Multi-Modality Processing
[85]	Implements TTS (Text-to-Speech) and STT (Speech-to-Text) models to enable bidirectional voice-text conversion, enhancing interactive multi-modal communication.
[157]	Integrates text-based medical knowledge, numerical sensor data, and image-based diagnostic guidelines using OCR, enabling comprehensive cross-modal reasoning.
[143]	Develops Radiology Tool and ReportVQA Tool to process and fuse radiological images and textual reports, enhancing multi-modal diagnostic support.
[78]	Employs Grounding DINO, MedSAM, G-Seg, and BiomedCLIP to extract, segment, and classify medical images, enabling advanced medical image analysis.
[1]	Supports multi-modal communication channels (text, audio, images, gestures) and a data pipeline to store and process multi-modal inputs for enhanced decision-making.
	Computational Reasoning
[141]	Implements self-reflection and iterative debugging to analyze error logs, refine generated code, and improve Pass@1 accuracy in computational tasks.
[55]	Utilizes a symbolic solver within the ArgMed-Agents framework to perform formal reasoning and construct a directed argument graph for decision support.
[115]	Employs LLM-generated executable code and a code execution engine to automate data processing, translating natural language tasks into iterative coding workflows.
	Some studies discussing using medical calculators. [167, 61, 67, 50]
	Other Tools for Clinical Decision Support
[28]	Integrates external diagnostic and treatment tools (Phenomizer, LIRICAL, Phenobrain, DrugBank, and DDI-graph) via APIs to enhance automated clinical decision-making.
[132]	Enhances LLM-based medical practitioners with NIH-provided specialized disease knowledge to improve domain-specific clinical expertise.
[165]	Extracts and utilizes ClinicalTrials.gov data to train predictive models (enrollment, drug risk, disease risk) for optimizing clinical trial design and recruitment.
[78]	Uses MRG (Medical Report Generation) tools, such as ChatCAD trained on MIMIC-CXR, to generate structured medical reports from imaging and clinical data.
[1]	Employs AI analytics and translation tools to extract insights, detect events, and enhance accessibility for multilingual clinical health assistants (CHAs).
[121]	Develops a pathology-specific CV model library, integrating classification, detection, segmentation, and generation models to support AI-assisted pathology diagnostics.
	Other Tools for Biomedical Research
[122]	Develops a virtual laboratory integrating ESM, AlphaFold-Multimer, and Rosetta to accelerate computational nanobody design for biomedical research.
[62]	Enables LLMs to interact with NCBI Web API, facilitating easy and efficient access to biomedical databases such as Entrez and BLAST without requiring local infrastructure.
[148]	Implements Tool Retrieval and Code Sandbox to automate tool selection and securely execute LLM-generated code for single-cell data analysis, improving reproducibility.
[91]	Integrates DrugAgent with specialized tools for dataset processing, molecular fingerprint extraction, and ChemBERTa-based tokenization, bridging AI with pharmaceutical research needs.

Table 14: Summary of Studies about Other Tools.

Study	Methodology	Key Contribution
	Long-Term Memory (LTM) –> Internal Memory	
[85]	Utilizes LLMs' internal memory to synthesize missing attributes in clinical case graphs based on embedded common-sense knowledge.	Demonstrates how LLMs' pre-trained knowledge can be leveraged to generate plausible clinical attributes, supporting zero- or few-shot tasks.
[140]	Updates ChatGPT's parameters using fine-tuning API to incorporate real-time patient records, improving predictions of adverse events and medication outcomes.	Enhances LLM internal memory by fine-tuning with real-world data, enabling dynamic adaptation to new medical information.
[114]	Employs self-enhancing agents that utilize LLMs' internal knowledge to refine and complete original clinical standards.	Showcases how LLMs' pre-trained knowledge can autonomously refine medical guidelines, enhancing consistency and applicability.
[128]	Builds AMIE on top of the pre-trained PaLM-2 LLM and enhances its medical dialogue and reasoning capabilities via instruction tuning.	Demonstrates the role of instruction tuning in leveraging internal memory to improve LLMs' performance in medical reasoning and dialogue generation.
	Long-Term Memory (LTM) –> External Memory –> Static Storage	
[105]	Implements a knowledge base module as a database for managing and storing structured medical knowledge across sessions.	Ensures stable cross-session knowledge retention, enabling consistent long-term decision-making in AI hospital systems.
[136]	Uses a Cognitive Conceptualization Diagram (CCD) to structure and store patients' historical and cognitive information for maintaining consistency in patient simulation.	Provides a structured long-term storage mechanism to maintain coherent cognitive modeling in multi-turn AI-patient interactions.
[133]	Extracts and stores psychological profiles and reference dialogues from original consultations as static information sources for guiding LLM-based simulations.	Enhances LLM-based patient simulations by leveraging pre-stored psychological data to maintain behavior consistency.
[88]	Stores long-term patient information, including medical history and personal health data, to ensure consistency in patient interactions.	Maintains stable and reliable patient records across interactions, supporting AI-driven medical consultations.
[157]	Constructs a medical expert knowledge base using pre-collected medical dialogues, textbooks, and diagnostic guidelines, storing them in vectorized format.	Facilitates structured knowledge retrieval to improve AI medical decision-making and diagnostic accuracy.
[142]	Employs a retrieval module that matches patient records with a pre-built medical document corpus to provide relevant knowledge.	Supports AI-assisted clinical review generation by retrieving pertinent information from static medical databases.
[94]	Utilizes a structured ESI manual as a reference knowledge base to guide domain-specific decision-making.	Provides stable domain knowledge to AI agents, ensuring consistency in emergency triage and decision-making.
[132]	Integrates NIH-provided expert disease knowledge into specialized AI agents for domain-specific expertise.	Enhances AI-based medical expertise by embedding stable, trusted disease knowledge into specialized agents.
[165]	Incorporates external databases like DrugBank, HetioNet, and ClinicalTrials.gov as stable knowledge sources for drug efficacy and safety evaluation.	Supports AI-driven pharmacological assessments by leveraging pre-structured external knowledge repositories.
[114]	Uses an indexed medical database for retrieving structured medical knowledge to assist AI decision-making.	Provides reliable long-term knowledge storage to enhance AI-generated medical recommendations.
[166]	Utilizes a knowledge retriever module that queries external repositories like Wikipedia for structured knowledge supplementation.	Enhances AI models' factual accuracy by incorporating structured, pre-stored external knowledge sources.
	Long-Term Memory (LTM) –> External Memory –> Dynamic Updating	
[85]	Constructs a case graph using open information extraction and dynamically updates it based on user queries to ensure dialogue consistency.	Enables real-time adaptation of patient case representations by integrating newly inferred attributes during interactions.
[93]	Implements a feedback-based pipeline where expert feedback (appreciation, criticism, rewriting) dynamically updates behavioral guidelines for AI patients.	Ensures continuous refinement of AI patient behavior through expert-driven adaptive rule adjustments.
[164]	Stores an AI Patient knowledge graph (KG) in a Neo4j database, allowing dynamic retrieval and integration of new patient insights.	Supports AI-driven patient simulations by dynamically incorporating evolving clinical knowledge.
[37]	Utilizes an "Attention Library" and "Trajectory Library" to collect and refine high-quality dialogue interactions for future retrieval.	Enhances diagnostic performance by dynamically updating and leveraging past high-quality conversations.
[28]	Maintains personalized long-term memory for AI doctors, integrating patient histories and prior consultation data for real-time decision support.	Improves diagnostic accuracy by continuously updating patient-specific medical histories.
[140]	Updates a patient's digital twin by incorporating the latest EHR data and nearest-neighbor patient records for real-time event prediction.	Enables personalized and continuously evolving risk assessments for clinical decision-making.
[157]	Synchronizes sensor data and medical knowledge updates, ensuring real-time integration of patient monitoring and clinical guidelines.	Provides continuous adaptation of medical insights by incorporating dynamic real-world patient and clinical data.
[113]	Uses a "notebook" memory feature that retains and updates key medical notes across patient interactions while employing RAG to retrieve updated knowledge.	Enhances AI doctors' memory by dynamically integrating newly acquired clinical insights.
[11]	Updates patient records dynamically within a hospital information system (HIS) to provide real-time access to the latest medical data.	Ensures AI-assisted clinical workflows remain aligned with the most recent patient information.
[141]	Employs retrieval-augmented generation (RAG) to fetch real-time medical knowledge from PubMed, GitHub, and StackOverflow.	Improves AI decision-making by dynamically incorporating the latest domain knowledge.
[117]	Uses a key-value memory system to store and retrieve expert feedback and case knowledge, dynamically updating learning outcomes.	Enables AI medical trainees to learn and adapt through dynamically evolving case-based reasoning.
[115]	Maintains a long-term memory (LLL) for storing past successful code examples, dynamically retrieving the most relevant samples for few-shot learning.	Enhances AI coding assistants by continuously updating and reusing high-quality knowledge snippets.
[81]	Updates a medical records library with verified QA pairs and an experience repository that refines diagnostic principles based on past errors.	Ensures AI decision-making improves over time by continuously refining its diagnostic and reasoning abilities.
[84]	Allows expert systems to dynamically expand their knowledge based on real-time patient interactions, integrating new details as they emerge.	Improves clinical reasoning by continuously incorporating patient-specific contextual information.
[135]	Uses MedSpaCy and QuickUMLS to extract and dynamically update clinical keywords guiding AI-driven dialogue generation.	Enhances conversational AI's ability to generate clinically relevant responses through real-time knowledge extraction.
[169]	Transfers frequently accessed knowledge from short-term memory (STM) to long-term memory (LTM) while preserving and refining user preferences over time.	Enables AI assistants to provide highly personalized and continuously evolving patient interactions.
	Short-Term Memory (STM)	
[92]	Retains full dialogue history within patient simulation models to ensure contextual coherence.	Enables seamless multi-turn interactions in AI-driven doctor-patient simulations.
[85]	Extracts core entities and relationships from dialogue context to match with case graphs dynamically.	Supports adaptive dialogue by leveraging short-term contextual memory for entity-based reasoning.
[164]	Maintains multi-turn dialogue history in a reasoning-based retrieval-augmented generation (RAG) framework.	Enhances AI-driven consultations by ensuring continuity in reasoning across multiple exchanges.
[37]	Implements an "instant memory" for recent dialogue continuity and a "summary memory" to consolidate past interactions.	Balances memory retention for real-time coherence and efficient context management in diagnostic conversations.
[133]	Uses dialogue context to maintain consistency in AI-driven therapist-patient conversations.	Ensures realistic and coherent therapeutic interactions through dynamic context retention.
[88]	Stores and updates short-term conversation history in AI-driven medical consultations.	Enhances response coherence by dynamically integrating prior dialogue turns.
[63]	Records multi-turn dialogue history during AI-patient interactions to track patient responses over time.	Supports consistent medical dialogue generation by preserving past interaction context.
[157]	Uses preceding and runtime prompts to store and retrieve patient symptoms and sensor data.	Enables AI-driven medical reasoning by dynamically integrating prior consultation data.
[68]	Maintains dialogue records across multiple MDT discussions to ensure coherent exchanges.	Enhances collaborative medical decision-making by preserving discussion history.
[124]	Tracks expert modifications across multi-turn collaborative consultations.	Ensures iterative improvement in AI-generated reports through sequential refinement.
[65]	Logs conversational exchanges among AI agents for dynamic adaptation.	Enables coordinated decision-making through multi-agent memory sharing.
[11]	Iteratively extracts symptoms and medical history from multi-turn conversations.	Ensures AI models retain relevant patient information throughout diagnostic dialogues.
[122]	Shares agendas, summaries, and context across AI-driven team discussions.	Facilitates structured multi-agent collaboration within AI-driven medical teams.
[117]	Maintains short-term dialogue history for multi-agent debate and discussion.	Enables AI agents to participate in continuous, logically coherent debates.
[55]	Retains dialogue sequences for argumentation-based AI medical interactions.	Enhances AI-driven medical debates by preserving structured argument flow.
[148]	Stores local task-specific memory for intermediate steps, resetting after task completion.	Supports efficient multi-step reasoning while preventing context overload.
[27]	Appends contextual reminders in dialogue prompts to mitigate instruction forgetting.	Ensures AI models maintain instructional adherence throughout extended conversations.
[84]	Preserves patient history and expert QA sequences within a single consultation session.	Enables AI-driven clinical reasoning by maintaining intra-session memory.
[135]	MedSpaCy and QuickUMLS-powered interactive planning	Dynamically updates extracted clinical terms to guide context-aware dialogue generation and maintain response accuracy
[128]	Retains previous dialogue turns for maintaining conversation continuity in medical simulations.	Improves AI doctor-patient interactions by leveraging stored conversational context.
[169]	STM stores the most recent user interactions within a session, ensuring real-time adaptation and contextual coherence.	Enables AI to capture and respond to immediate user inputs, dynamically adjusting responses based on real-time feedback.
	Multi-Agent Shared Working Memory (WM)	
[68]	Stores prior dialogue records and integrates feedback from multi-agent discussions to iteratively refine decisions.	Enhances collaborative AI decision-making by preserving intermediate reasoning steps and optimizing responses through feedback integration.
[124]	Implements a voting-based refinement process where domain experts iteratively modify and validate AI-generated reports.	Ensures high-quality medical reporting through structured multi-agent feedback and consensus-building.
[142]	The meta-doctor agent consolidates feedback from multiple AI doctor agents, analyzing inputs to guide further discussions.	Improves AI-driven collaborative decision-making by systematically integrating expert agent insights.
[65]	Uses a recorder agent to extract learning points, compile differential diagnosis lists, and refine discussions.	Supports AI-driven medical learning and iterative diagnostic improvement.
[94]	Aggregates classification results, confidence scores, and reasoning outputs to generate iterative reports in multi-agent discussions.	Enables continuous decision-making support by dynamically storing and refining critical inference data.
[122]	Integrates multi-agent critiques and iteratively refines responses in scientific discussions and team meetings.	Strengthens AI-driven scientific reasoning by enabling structured multi-agent feedback loops.
[55]	Maintains an argumentation framework by recording iterative reasoning processes and identifying semantic relationships between arguments.	Facilitates structured, transparent, and continuous AI reasoning in complex decision-making.
[166]	Temporarily stores execution trajectory data across decision steps, enabling context-aware decision-making.	Enhances multi-agent collaboration by preserving relevant decision context dynamically.
[148]	Stores final outputs of sub-tasks in a global memory shared by expert AI agents, reducing redundant computations.	Optimizes AI-driven complex decision-making by ensuring efficient knowledge reuse across iterative tasks.
[91]	Logs tool-building failures in the LLM Planner module and dynamically adjusts candidate ideas based on iterative analysis.	Improves AI-assisted planning and execution through failure-driven adaptation and optimization.

Table 15: Summary of Studies on Memory in AI Hospitals

Study	Methodology	Key Contribution
	Direct Reasoning –> Single-path Reasoning	
[85]	Implements an Extract-Retrieve-Rewrite-Generate (ERRG) process, forming a fixed sequential logic chain.	Establishes a structured, step-by-step knowledge retrieval and response generation system for AI hospital dialogue systems.
[136]	Constructs a cognitive conceptualization map (CCD) through a fixed, sequential extraction of cognitive elements.	Provides structured cognitive reasoning for AI-driven patient dialogue modeling.
[133]	Simulates client behavior in therapy using predefined psychological profiles and structured responses.	Enables AI-based therapy simulations with deterministic, single-path reasoning.
[68]	Employs few-shot prompting to guide a primary care AI agent in simple case handling.	Uses a deterministic, stepwise approach for AI-driven simple medical cases.
[113]	Utilizes Chain-of-Thought (CoT) reasoning in AI models for structured problem-solving.	Enhances AI reasoning through explicitly guided, stepwise inference.
[11]	The AI nurse agent follows a predefined decision-making logic based on prior interactions.	Provides structured, deterministic guidance in AI-based nurse-patient interactions.
[141]	Uses Chain-of-Thought (CoT) prompting to guide LLMs in generating structured responses in clinical decision-making.	Enhances AI-driven medical reasoning through structured logic prompts.
[165]	Implements Least-to-Most reasoning, solving problems progressively in increasing complexity.	Ensures stepwise, sequential problem-solving for medical decision support.
[156]	Follows a stepwise diagnostic process (preliminary diagnosis $\rightarrow$ differential diagnosis $\rightarrow$ final diagnosis $\rightarrow$ treatment).	Implements a structured reasoning flow for AI-driven medical diagnosis.
[24]	Uses CoT-based diagnostic reasoning to match symptoms with potential conditions in a structured flow.	Enhances AI-based medical diagnosis through structured, progressive inference.
[161]	Forms a fixed sequential logic chain for generating and iteratively improving USMLE-style questions.	Establishes a structured, step-by-step question generation and refinement system, enhancing medical question quality and reasoning accuracy through iterative LLM feedback.
	Direct Reasoning –> Multi-path Reasoning	
[37]	Uses a multi-disciplinary approach where different doctor agents collaborate via a directed acyclic graph (DAG) for diagnosis.	Enables flexible and diverse diagnostic reasoning by integrating multiple expert perspectives.
[28]	Employs a multi-round discussion framework where multiple specialist agents iteratively refine their diagnostic consensus using external medical tools.	Enhances diagnostic accuracy by leveraging multiple expert viewpoints and integrating tool-assisted reasoning.
[68]	Uses a multi-disciplinary team (MDT) framework where multiple agents engage in iterative discussions to form a consensus diagnosis.	Improves diagnostic reliability by synthesizing diverse medical expertise through multi-round interactions.
[124]	Invites multiple domain experts to independently analyze a medical case from different perspectives before integrating their insights.	Strengthens decision-making by aggregating multiple analytical approaches to refine reasoning.
[142]	Individual doctor agents conduct independent initial assessments, and a meta-agent consolidates their findings into a unified report.	Enables comprehensive case evaluation by incorporating multiple reasoning paths into a single decision-making process.
[94]	Utilizes two independent reasoning paths—one with sequential expert-based refinement and another with direct fine-grained classification—before merging conclusions.	Enhances medical triage accuracy by combining complementary diagnostic perspectives.
[122]	Runs multiple parallel clinical discussions and synthesizes their conclusions into a single optimal response.	Ensures thorough decision-making by exploring diverse diagnostic paths before consolidation.
[132]	Uses multiple expert agents to generate independent probabilistic diagnoses, which are then adaptively weighted and merged.	Increases diagnostic robustness by fusing probabilistic predictions from multiple agents.
[55]	Combines multiple reasoning paths from different agents (generator, validator, inferencer) and selects the most coherent decision using symbolic reasoning.	Improves AI decision-making by leveraging argumentation theory to synthesize multiple logical perspectives.
[91]	Generates multiple candidate solutions in parallel and iteratively refines them using experimental validation.	Expands the AI's problem-solving flexibility by considering multiple reasoning paths before selecting the optimal one.
[40]	Multiple doctor agents interact with the same patient record independently, generating diverse diagnostic insights before being consolidated into a final decision.	Enhances diagnostic reliability through iterative multi-agent collaboration.
[84]	Implements a self-consistency mechanism where multiple reasoning trajectories are generated and merged via majority voting or averaging.	Strengthens decision-making by ensuring robustness through multiple internal validation paths.
[80]	Three independent doctor agents analyze a patient case separately, and their insights are later reconciled into a consensus diagnosis.	Enhances diagnostic accuracy by utilizing multiple independent medical reasoning paths.
	Feedback-Based Reasoning –> External Feedback	
[63]	Utilizes a patient AI agent to provide iterative feedback through multi-turn conversations, refining the diagnostic reasoning process dynamically.	Demonstrates how LLMs integrate external patient feedback during the diagnostic process, improving accuracy and adaptability in medical decision-making.
[28]	Employs tool integration (e.g., Phenomizer, LIRICAL, DrugBank) to allow AI doctors to query external diagnostic and treatment knowledge bases.	Enables LLM-based medical agents to dynamically adjust diagnoses and treatment plans based on real-time external feedback from specialized clinical tools.
[140]	Incorporates external feedback from EHR databases and similar patient records to refine medical predictions.	Enhances diagnostic reasoning by leveraging historical patient data as external references, improving contextual accuracy in medical decision-making.
[141]	Integrates human expert instructions and external knowledge sources (web, research papers) to refine model-generated outputs.	Demonstrates how structured external feedback from experts and online sources improves the quality and reliability of AI-driven reasoning in clinical tasks.
[84]	Implements an iterative feedback loop where AI refines its reasoning based on patient-provided responses through multi-turn interactions.	Establishes an interactive AI-driven consultation process where patient feedback incrementally enhances diagnostic accuracy.
[135]	Leverages external structured information (MedSpaCy, QuickUMLS) and expert feedback rules to optimize medical dialogue generation.	Reduces hallucination and ensures clinical logic consistency by dynamically incorporating external structured data and expert insights into medical conversations.
	Feedback-Based Reasoning –> Self Feedback	
[93]	Implements an internal pipeline where generated responses are evaluated against expert-defined rule-based yes/no questions, triggering iterative rewriting if needed.	Establishes a structured self-feedback mechanism where AI systematically refines outputs to comply with predefined rules, ensuring logical consistency.
[164]	Incorporates an internal validation agent that assesses retrieved medical knowledge, reformulating queries iteratively until consistency is achieved.	Introduces a multi-round self-feedback loop where AI autonomously verifies and refines reasoning for reliable medical decision-making.
[140]	Compares AI-generated treatment outcomes with real patient experiences, optimizing predictions through iterative self-correction.	Implements self-feedback loops that enable continuous model adaptation, reducing discrepancies between predictions and actual clinical data.
[68]	Integrates an LLM moderator that oversees multi-agent medical discussions, synthesizing feedback to refine collective decision-making.	Ensures self-correction in AI-driven multidisciplinary diagnosis by iteratively revising conclusions based on internal discussions.
[124]	Facilitates expert voting and feedback loops to revise diagnostic reports iteratively until consensus is reached.	Uses internal expert simulation to enforce self-correction, improving decision robustness in complex cases.
[142]	Engages multiple AI agents in iterative discussion cycles to validate and refine diagnostic conclusions.	Implements structured internal feedback to minimize inconsistencies and optimize clinical decision-making.
[65]	Simulates hierarchical medical roles (junior doctors, senior doctors) to iteratively review, critique, and refine diagnostic reasoning.	Establishes a peer-review-like self-feedback mechanism that enhances the reliability of automated medical diagnoses.
[113]	Introduces an explicit reasoning reflection step where AI revisits its thought process and refines outputs before finalizing conclusions.	Enhances logical coherence in AI reasoning by enforcing systematic self-review before response finalization.
[94]	Allows AI agents to cross-evaluate each other's diagnostic confidence levels, prompting self-adjustments in reasoning.	Strengthens self-feedback mechanisms by enabling agents to refine predictions based on internal peer comparisons.
[11]	Employs an internal supervision agent that critiques nurse agents' decisions, guiding self-revision before finalizing patient care plans.	Improves clinical decision-making through structured self-feedback loops, reducing errors in nursing workflows.
[141]	Uses multi-round error analysis and self-correction to iteratively refine AI-generated code.	Demonstrates self-feedback in complex problem-solving by iteratively improving computational outputs.
[122]	Employs AI-driven scientific debate cycles where agents critique and refine each other's diagnostic reasoning.	Improves medical reasoning accuracy through structured internal argumentation and self-refinement.
[117]	Implements multi-role AI agents (e.g., "Angel" and "Devil") that iteratively challenge and refine each other's reasoning.	Enhances internal self-feedback by fostering argumentative reasoning within AI medical debates.
[114]	Integrates self-verification mechanisms where AI agents assess the correctness and consistency of their knowledge retrieval.	Reduces factual inconsistencies in medical reasoning through structured internal review processes.
[55]	Uses argumentation-based self-feedback, where AI-generated diagnoses undergo recursive critique before confirmation.	Enhances reasoning robustness by enforcing iterative self-verification before reaching a final medical decision.
[166]	Leverages previous trajectory records to iteratively optimize AI-driven medical decision-making.	Improves AI hospital agent adaptability by enabling self-feedback-based continuous refinement.
[115]	Integrates a rubber duck debugging mechanism where LLMs analyze their own execution errors and refine solutions.	Demonstrates AI self-correction in medical automation by iteratively improving execution logic.
[91]	Uses internal feedback loops to refine AI decision-making based on detected failures in prior execution attempts.	Enhances robustness by allowing AI agents to iteratively adjust strategies through self-monitoring.
[81]	Employs an internal repository where AI agents analyze and refine diagnostic decisions based on past learning experiences.	Establishes a continuous self-feedback cycle where AI accumulates and applies experiential knowledge in medical decision-making.
[40]	Implements an internal feedback system where AI agents iteratively resolve conflicting diagnoses through structured discussions.	Strengthens AI hospital reasoning by enabling self-consensus formation through iterative feedback.
[156]	Uses a structured internal validation mechanism where AI assesses and refines its diagnostic reasoning before finalizing medical conclusions.	Enhances medical accuracy by enforcing multi-stage self-verification processes in AI-driven clinical workflows.
[135]	Introduces an internal self-correction module that refines AI-generated medical dialogues iteratively.	Ensures medical dialogue quality by integrating structured self-feedback in AI-driven patient communication.
[128]	Uses a dedicated AI critic to evaluate and refine physician-agent responses based on predefined quality standards.	Demonstrates self-feedback-driven iterative optimization in AI-mediated medical interactions.
[80]	Uses structured self-reflection to enable AI agents to iteratively refine their diagnostic reasoning.	Improves logical coherence by allowing AI systems to assess and adjust their own reasoning before decision finalization.

Table 16: Summary of Studies on Reasoning Patterns in AI Hospital.

Study	Methodology	Key Contribution
Clinical Workflow Simulation		
Online Medical Consultation Simulation [92]	Multi-agent AI hospital framework simulating closed-loop interactions from patient consultation to diagnosis using a patient simulator and structured questioning.	Enhances diagnostic accuracy and workflow optimization by analyzing the impact of different inquiry strategies on final diagnoses.
CRAFT-MD [63]	Multi-agent clinical reasoning assessment framework with AI-driven doctor-patient dialogues and automatic diagnosis evaluation.	Improves diagnostic reasoning by quantifying LLM performance in patient interactions and refining information integration through conversation summarization.
AgentClinic [113]	Multi-modal agent benchmark replicating full clinical workflows with sequential decision-making and bias simulation.	Evaluates LLMs across nine specialties and seven languages, introducing cognitive bias challenges to assess robustness in clinical decision-making.
PIORS [11]	Multi-agent intelligent outpatient reception system integrating LLM-based agents with hospital information systems (HIS).	Enhances triage accuracy and operational efficiency by simulating patient interactions and dynamically adjusting inquiry strategies.
Agent Hospital [81]	Virtual hospital with LLM-driven agents simulating triage, consultation, diagnostics, and follow-ups, incorporating self-learning from case histories.	Improves diagnostic accuracy through memory-based retrieval and self-adaptive learning, reducing dependence on manually labeled data.
AI Hospital [40]	Multi-agent system replicating complete clinical workflows, integrating a multi-view medical evaluation (MVME) benchmark for LLM performance assessment.	Enhances diagnosis and decision-making by incorporating dispute resolution and physician-agent collaboration.
MEDIQ [84]	Interactive framework modeling dynamic doctor-patient dialogues with confidence-based inquiry strategies.	Improves diagnostic accuracy by introducing strategic questioning and reasoning consistency mechanisms to mitigate incomplete information issues.
NoteChat [135]	Multi-agent framework simulating clinical dialogues and documentation processes with keyword extraction, role-play, and text refinement modules.	Enhances clinical note generation and workflow efficiency by aligning AI-generated conversations with professional medical documentation standards.
Multi-Disciplinary Medical Team (MDT) Simulation		
RareAgents [28]	Multi-agent system modeling MDT workflows for rare disease diagnosis, dynamically forming specialist teams with long-term memory and medical tool integration.	Enhances complex disease diagnosis and personalized treatment by simulating real-world interdisciplinary expert collaboration and decision-making.
MDAgents [68]	Adaptive multi-agent framework adjusting team composition based on query complexity, integrating external medical knowledge sources (MedRAG).	Optimizes collaborative clinical decision-making by dynamically scaling expertise levels and refining multi-step medical reasoning.
MedAgents [124]	Multi-agent framework structuring expert recruitment, independent analysis, multi-round discussion, and collaborative decision-making.	Improves diagnostic accuracy and medical reasoning by mimicking expert team deliberation without requiring additional model training.
ColaCare [142]	Multi-agent MDT simulation combining structured EHR analysis with authoritative medical guideline retrieval, using DoctorAgent and MetaAgent coordination.	Enhances clinical decision-making by integrating structured and unstructured medical data for more interpretable and precise mortality risk predictions.
Simulated Patients for Medical Education and Training		
GPT-3.5 Patient Simulator [54]	Single-agent chatbot simulating patient dialogues based on predefined disease scripts for history-taking training.	Offers a cost-effective, reusable platform for medical students to practice clinical questioning and patient communication.
EvoPatient [37]	Multi-agent co-evolutionary simulation framework using patient and doctor agents with RAG-enhanced dialogue generation.	Provides an adaptive virtual training environment for medical professionals by improving questioning strategies and patient responses over time.
AIE & SAPS [88]	Automated interaction evaluation framework with state-aware patient simulators for dynamic clinical dialogues.	Captures real-time decision-making behaviors of medical AI models, improving the assessment of diagnostic reasoning and patient interaction.
MEDCO [143]	Multi-agent training platform simulating multi-role interactions (patients, students, radiologists, experts) with multimodal data integration.	Enhances clinical communication and diagnostic skills by enabling cross-disciplinary and interactive learning experiences.
AI-SCE [96]	Agent-based modeling (ABM) framework simulating structured clinical examinations (SCEs) for comprehensive workflow analysis.	Provides a detailed evaluation of AI-driven clinical decision-making by capturing both final outputs and intermediate reasoning steps.
MedQA-CS [163]	AI-structured clinical examination (AI-SCE) framework with LLMs simulating both medical students and clinical examiners.	Provides a more realistic and challenging alternative to MCQ-based assessments by structuring medical student evaluations through real-time clinical scenario simulations.
OSCEBot [105]	Single-agent patient simulation system integrating NLP and semantic matching for structured OSCE training.	Improves medical education by aligning student interactions with clinical scripts using sentence embedding techniques and real-time feedback.
Psychological Counseling and Mental Health Interaction Simulation		
Roleplay-doh [93]	Single-agent system optimizing AI patient responses based on expert-driven feedback-to-guideline transformation.	Improves AI-based therapy training by dynamically aligning simulated patient behaviors with expert-defined expectations in real-time conversations.
PATIENT-Ψ [136]	Single-agent CBT-based patient simulator integrating cognitive models and diverse conversation styles for training new therapists.	Enhances psychological training by offering structured roleplay environments with expert-defined patient behavior modeling and real-time feedback.
ClientCAST [133]	Multi-agent AI framework simulating therapist-client interactions with personalized psychological profiles and post-session evaluations.	Enhances AI-driven mental health simulations by enabling detailed personality-based patient modeling and comparative analysis across multiple LLMs.
Therapy Simulation [107]	Multi-agent system modeling structured therapy sessions with LLM-based patient and therapist roles using integrative therapy frameworks.	Provides a scalable and cost-effective platform for evaluating AI-assisted mental health support, generating high-quality synthetic training data.
Psychiatric AI Simulation [27]	Multi-agent AI hospital framework with DSM-5-guided doctor and emotionally responsive patient chatbots.	Provides realistic mental health diagnostic simulations by integrating structured psychiatric assessment with empathetic patient dialogue modeling.
Other Medical Process Optimization and Cross-Disciplinary Simulation		
Virtual Lab [122]	Multi-agent AI-driven collaborative research platform with domain-specific expert agents for interdisciplinary biomedical discovery.	Accelerates cross-disciplinary scientific problem-solving, exemplified by AI-assisted design of SARS-CoV-2 nanobody therapeutics.
AI-Driven Epidemic Simulation [145]	Multi-agent system combining generative AI (LLMs) with agent-based epidemiological modeling (ABM).	Improves pandemic response simulations by enabling dynamic, autonomous decision-making in synthetic populations based on real-time health and behavioral data.

Table 17: Summary of Studies on Simulating Specific Scenarios

Study	Methodology	Key Contribution
Precision Clinical Decision-Making & Bias Mitigation		
RareAgents [28]	Multi-agent AI hospital framework using MDT-style collaboration with dynamic long-term memory and medical tool integration.	Enhances rare disease diagnosis and treatment recommendations through specialized agent collaboration, outperforming existing medical LLMs in accuracy and drug recommendation F1 score.
DrHouse [157]	Multi-agent system integrating sensor data, medical guidelines, and interactive reasoning with adaptive uncertainty evaluation.	Achieves 18.8% improvement in diagnostic accuracy and 91.7% patient satisfaction by combining multi-source data fusion, multi-agent collaboration, and real-time medical updates.
MDAgents [68]	Adaptive multi-agent decision framework adjusting from single-agent to MDT or ICT based on case complexity.	Improves clinical decision-making accuracy by 11.8% through adaptive expert recruitment and self-consistency-enhanced reasoning.
MedAgents [124]	Multi-agent expert discussion with iterative refinement for medical reasoning and decision-making.	Outperforms Zero-shot CoT and Self-Consistency methods in 9 medical datasets, reducing hallucinations and improving diagnostic reliability.
ColaCare [142]	DoctorAgent and MetaAgent framework combining structured EHR analysis with LLM reasoning.	Enhances clinical decision transparency and bias correction through a multi-perspective collaborative consultation mechanism.
AI Diagnostic [65]	GPT-4-based multi-agent system with four roles: initial diagnostician, opposer, mentor, and recorder.	Achieves a 71.3% improvement in initial diagnostic accuracy by simulating clinical team decision-making and mitigating cognitive biases.
AMSC [132]	Multi-expert collaboration using LLMs with adaptive probability fusion for diagnosis.	Reduces reliance on large-scale LLM fine-tuning and enhances fairness by leveraging multi-agent voting and weighted decision fusion.
ArgMed-Agents [55]	Argumentation-driven multi-agent system with generator, validator, and reasoner for clinical reasoning.	Improves diagnostic accuracy and explainability by enabling iterative self-argumentation to minimize hallucination and knowledge conflicts.
MMedAgent [78]	Multi-modal LLM-based multi-agent system with dynamic tool selection for medical tasks.	Outperforms GPT-4o in various clinical tasks by integrating specialized medical tools through instruction tuning.
TBI Rehabilitation [87]	Multi-agent framework using guideline-based GPT agents to improve clinical QA accuracy.	Enhances diagnostic accuracy, interpretability, and empathy compared to GPT-4 while reducing hallucinated answers.
AMIE [128]	Self-play multi-agent AI with simulated doctor-patient interactions and feedback optimization.	Achieves superior diagnostic accuracy and patient interaction quality compared to human PCPs in a randomized controlled trial.
DiagnosisGPT [24]	Chain-of-Diagnosis (CoD) framework decomposing LLM decision-making into modular diagnostic steps.	Increases transparency and reduces bias by structuring the LLM-driven diagnostic process into discrete, explainable stages.
MAC [80]	Multi-agent conversational framework simulating real-world clinical team collaboration for rare disease diagnosis.	Improves diagnostic accuracy and test recommendation quality compared to single-agent LLMs like GPT-4.
MEDAIDE [144]	Multi-agent LLM framework leveraging retrieval-augmented generation, intent recognition, and agent collaboration for handling complex medical tasks.	Enhances strategic reasoning and decision-making in healthcare by integrating domain-specific knowledge retrieval, fine-grained intent identification, and collaborative agent-based analysis.
Clinical Trial Optimization and Triage Systems		
TRIAGEAGENT [94]	Heterogeneous multi-agent framework leveraging RAG for multi-stage reasoning and real-time ESI guideline integration.	Reduces triage error rates and enhances interpretability by dynamically optimizing decision-making through multi-agent collaboration and confidence-based stopping.
PIORS [11]	Multi-agent system with LLM-based nurse and hospital assistant using SFMSS framework for patient triage and personalized reception.	Improves outpatient triage accuracy and reduces wait times by simulating diverse patient interactions and integrating hospital information systems.
ClinicalAgent [165]	Multi-agent system incorporating GPT-4 with ReAct and Least-to-Most reasoning for clinical trial result prediction and optimization.	Enhances clinical trial recruitment and safety evaluation by integrating structured knowledge from DrugBank and HetioNet with agent-based decision-making.
MAKA [114]	Multi-agent knowledge augmentation framework dynamically supplementing missing information for clinical trial matching.	Increases patient-trial matching F1 scores by 6-10% through specialized agents identifying knowledge gaps and integrating external medical data sources.
Knowledge-Intensive Tasks and Medical Data Science		
Clinical Data Science Assistant [141]	Multi-agent AI hospital system integrating LLMs with human workflows for code generation, data analysis, and visualization using CoT and Self-Reflection.	Increases clinical data science efficiency by reducing coding errors (60% improvement) and accelerating task completion through AI-assisted code generation and iterative refinement.
SMART [166]	Multi-agent framework with specialized agents (intent reconstructor, knowledge retriever, fact locator, and response generator) optimizing medical knowledge retrieval and clinical reasoning.	Enhances diagnostic reliability and patient education by reducing hallucinations and improving structured medical knowledge retrieval.
EHRAgent [115]	Multi-agent LLM system integrating interactive code generation with execution feedback for complex EHR queries.	Improves multi-table reasoning and clinical data extraction by autonomously generating, debugging, and executing queries, reducing reliance on data engineers.
MCQG-SRefine [161]	A multi-step self-refinement pipeline that integrates retrieval-augmented generation, LLM-based critique and correction, and iterative feedback loops to generate high-quality USMLE-style multiple-choice questions.	Enhances the quality and difficulty of LLM-generated medical exam questions by leveraging structured topic and test point selection, retrieval-based example referencing, and iterative critique-correction cycles to ensure domain relevance and rigorous evaluation.
Scientific Discovery and Research Automation		
Virtual Lab [122]	Multi-agent AI hospital framework with a PI agent leading domain-specific AI researchers in interdisciplinary scientific discovery.	Successfully designs SARS-CoV-2 nanobodies by integrating AI-driven protein modeling tools (ESM, AlphaFold, Rosetta), demonstrating AI's role in accelerating cross-disciplinary research.
AI Scientist Framework [44]	Multi-agent system with hypothesis-driven AI agents collaborating through LLMs, machine learning tools, and experimental platforms.	Automates hypothesis generation, experimental validation, and scientific debate, reducing research costs and increasing discovery speed in biomedicine.
CellAgent [148]	Multi-agent system for scRNA-seq data analysis with role-based agents (planner, executor, evaluator) executing end-to-end bioinformatics workflows.	Automates and optimizes single-cell RNA sequencing analysis, enhancing reproducibility, accuracy, and accessibility for non-expert researchers.
DrugAgent [91]	Multi-agent AI framework for drug discovery integrating ML-based feature engineering, model training, and idea-space optimization.	Improves ADMET predictions (F1 = 0.92) by reducing human intervention in ML-based pharmaceutical research, accelerating drug development workflows.

Table 18: Summary of Studies on Solving Complex Tasks with Multi-Agent AI Hospitals

Study	Methodology	Key Contribution
[92]	Extracts real patient interaction strategies to train a patient simulator and enables multi-turn diagnosis simulations with LLM-based doctors.	Establishes an automated evaluation framework with hallucination rates, irrelevant response rates, and human-likeness scores to comprehensively assess medical AI agents.
[88]	Implements an Automated Interactive Evaluation (AIE) framework and a State-Aware Patient Simulator (SAPS) to track dialogue state and assess LLMs' adaptability.	Provides a dynamic evaluation system that quantifies LLM performance in multi-turn medical consultations, bridging the gap between static knowledge assessment and real-world clinical scenarios.
[63]	Uses LLM-based doctor, patient, and grader agents to evaluate clinical reasoning, medical history collection, and diagnostic accuracy in interactive dialogue settings.	Integrates multi-agent collaborative assessment, leveraging real-time patient interaction and structured scoring for an AI-driven medical evaluation system.
[113]	Builds a multi-agent system for interactive clinical evaluation, incorporating real-time role-playing across multiple medical specialties and languages. Incorporates external tools (e.g., calculators, retrieval mechanisms, memory notebooks) to assess AI decision-making in complex clinical scenarios.	Introduces a bias-aware assessment that evaluates the impact of cognitive and implicit biases on LLM clinical decision-making and patient trust. Demonstrates how tool integration improves LLM diagnostic accuracy and reasoning ability in a hospital simulation environment.
[40]	Simulates real-world medical interactions with multi-agent role-playing, including doctors, patients, and coordinators, while integrating multi-view evaluation benchmarks.	Establishes a collaborative decision-making system where LLM-driven doctors resolve diagnostic disagreements via debate and consensus-building.
[156]	Designs an end-to-end multi-specialty evaluation benchmark with case-driven assessment metrics.	Creates a structured, multi-agent collaborative system that accurately mimics hospital workflows to evaluate LLMs' clinical decision-making capabilities.
[84]	Develops an AI hospital system with patient and expert agents that engage in multi-turn automated interactions using state-aware tracking.	Enhances LLM evaluation by introducing modular expert systems that systematically assess clinical reasoning and adaptive questioning in uncertain information settings.
[128]	Uses multi-agent self-play and feedback loops to refine LLM-based doctor-patient interaction for clinical reasoning and communication.	Demonstrates LLMs' superiority over human physicians in patient interaction and diagnostic accuracy through randomized, double-blind studies.
[93]	Implements expert-driven qualitative feedback to optimize LLM role-playing in mental health simulations.	Establishes an interactive, expert-guided LLM evaluation pipeline that ensures AI-generated patient responses adhere to realistic psychological behaviors.
[136]	Combines LLMs with CBT-based patient cognitive models to create a multi-agent AI hospital framework for psychotherapy training.	Develops an interactive, memory-based patient simulator that personalizes psychological training scenarios and evaluates LLM therapists' adaptability.
[133]	Simulates patient-therapist interactions using LLM-generated clients and therapist agents to evaluate AI in counseling.	Introduces a double-layer agent system that assesses LLM therapists' empathy, communication, and effectiveness through structured psychological metrics.
[107]	Uses a dual-agent approach where LLMs act as both therapists and patients in multi-turn consultations.	Provides a scalable, privacy-conscious evaluation platform for AI-based psychological counseling with automated response assessment.
[85]	Uses virtual simulated patients (VSPs) with graph-driven context management and retrieval-augmented generation.	Creates a dynamic, interactive evaluation system for medical trainees, measuring their ability to engage in structured and naturalistic clinical dialogues.
[96]	Simulates high-fidelity clinical exams (e.g., OSCE) to assess LLMs' diagnostic reasoning and decision-making.	Develops a structured, interactive evaluation approach that tests LLMs beyond multiple-choice assessments by integrating dynamic case-based scenarios.
[163]	Uses an AI-structured clinical exam (AI-SCE) with LLM agents as medical students and clinical examiners.	Provides a benchmark that aligns LLM assessments with real-world clinical exams, evaluating their ability to conduct medical interviews and differential diagnoses.
[67]	Evaluates LLMs on text-based medical problem-solving and quantitative decision-making using calculation tasks.	Assesses how well LLMs replicate real-world physician workflows that involve medical calculators and quantitative reasoning.
[78]	Integrates multimodal LLMs with specialized medical tools for radiology and diagnostic reasoning.	Establishes an AI-driven, multimodal medical evaluation system that mimics expert workflows in clinical imaging and diagnostics.

Table 19: Summary of Evaluating Agents in AI Hospital.

Study	Methodology	Key Contribution
[37]	Implements a co-evolutionary multi-agent system where virtual doctors and patients engage in staged diagnostic conversations to generate standardized and high-quality training data.	Introduces an automated patient-doctor dialogue synthesis framework that ensures contextual coherence and domain diversity, enhancing AI medical training without relying on sensitive real-world data.
[81]	Uses a fully virtualized multi-agent hospital to synthesize medical training data, evolving patient agents and doctor agents through reinforcement-driven self-improvement.	Establishes an autonomous, scalable data synthesis pipeline that continuously refines AI-driven medical decision-making, bridging virtual training with real-world clinical performance.
[135]	Simulates real-world clinical consultations using structured dialogue planning, role-played interactions, and iterative refinement with medical expert feedback.	Develops a structured, multi-phase dialogue generation system that produces high-fidelity synthetic doctor-patient conversations, optimizing AI models for clinical note generation and medical QA.
[128]	Implements a self-play simulation environment where AI doctor, patient, and critic agents iteratively refine diagnostic reasoning and patient communication.	Enables adaptive learning through synthetic case-based interactions, overcoming data scarcity and improving LLM performance in clinical history-taking and medical reasoning.

Table 20: Summary of Synthesizing Data for Training in AI Hospital