Recontextualizing NLP in Healthcare: A Survey on LLM-Based Multi-Agent AI Hospitals

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly being used as autonomous agents in high-stakes domains, yet their behavior in complex, real-world environments remains under-explored. This survey introduces the concept of AI hospitals—LLM-driven multi-agent ecosystems that simulate clinical workflows and support a wide range of medical applications. We review 70+ recent studies and propose a taxonomy covering core components and application areas. By analyzing how these systems integrate language, knowledge, and interaction in dynamic settings, we highlight AI hospitals as a powerful testbed for evaluating LLMs beyond static benchmarks. We also outline open challenges in aligning LLM behavior with clinical reasoning, safety, and patient-centered goals, offering a roadmap for the future at the intersection of NLP and healthcare.

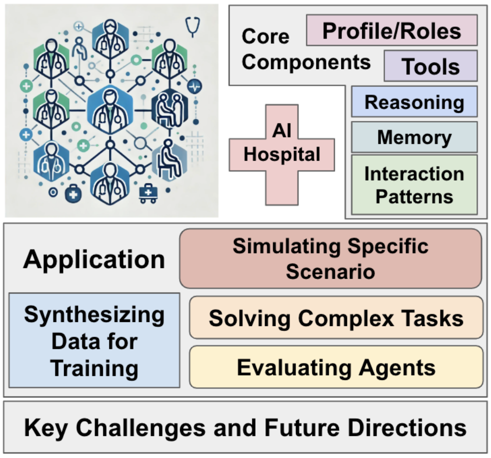

Figure 1: Overview of the LLM-based multi-agent AI Hospital. Figure 2 & 3 present the taxonomy of core components and applications. Table 2 and 3 in the appendix outline the key challenges and future directions.

1 Introduction

Over the past decade, natural language processing (NLP) and artificial intelligence (AI) have achieved significant advances across tasks such as translation, summarization, and question answering. In recent years, Large Language Models (LLMs) have emerged as a transformative force, demonstrating strong generalization, reasoning, and interaction capabilities. Beyond text generation, LLMs are increasingly being deployed as autonomous agents capable of decision-making and collaboration in real-world systems. A promising example of this shift is the AI hospital: a multi-agent simulation framework in which LLMs act as diverse clinical agents—doctors, nurses, patients, researchers—within simulated hospital environments. These systems go beyond static benchmark evaluation by enabling dynamic, interdisciplinary assessments of agent behavior in clinical decision-making, education, mental health support, and collaborative research. Despite growing interest, research on AI hospitals remains fragmented. Existing studies often focus on isolated multi-agent applications, lacking a unifying framework to connect them. To date, no prior survey has systematically examined these efforts through the lens of AI hospitals, nor analyzed their core components, applications, and open challenges.

This survey addresses this gap by organizing 70+ recent studies into a structured taxonomy across three dimensions: **1) Core Components**: Analyzing the fundamental elements of AI hospitals, including agent roles, interaction patterns, tool integration, memory management, and reasoning mechanisms. **2) 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. **3) Key Challenges & Future Directions**. By providing a cohesive framework, this work aims to strengthen collaboration between NLP and healthcare communities and to recontextualize LLMs as agents within real-world systems.

2 Core Components

2.1 Agent Roles

Patient-Centered Agents are designed to simulate patients with different demographic backgrounds, health conditions, and communication abilities. **Patient Agent** supports various applications in AI hospitals, such as clinical training, patient education, and medical history collection. Many works (Bao et al., 2024; Wang et al., 2023a) focus on enhancing the realism of patient agents. Recent studies (Du et al., 2024; Li et al., 2024d; Yu et al., 2024; Liu et al., 2025) also leverage evolutionary learning, fine-tuning techniques, Chain-of-Thought (CoT), and Retrieval-Augmented Generation (RAG) to enhance patient agents' consistency, realism, and role-playing stability while reducing hallucinations. **Psychological Patient Agent** (PPA) simulates mental health conditions for AI-driven treatment training (Wang et al., 2024b; Wei et al., 2024a). Unlike general patient agents, PPAs must replicate mood changes, cognitive distortions, and treatment resistance, with studies focusing on authenticity through expert-guided prompt engineering (Louie et al., 2024), structured cognitive modeling (Wang et al., 2024d), and simulations fostering adaptive communication (Chen et al., 2023b). **Resident Agents** model general populations transitioning into patient agents when ill, autonomously navigating healthcare processes while also supporting public health simulations and epidemiological modeling by incorporating disease progression, healthcare-seeking behavior, and policy interventions (Li et al., 2024b; Williams et al., 2023).

Medical Professional Agents can perform tasks such as patient consultation, medical history collection, clinical reasoning, diagnostic decision-making, emotional support, care coordination, and auxiliary examinations. **General Doctor Agent** , often called primary care physician (PCP), performs initial patient assessments and oversees the diagnostic process. Several studies have explored various aspects of these agents, including their questioning strategies (Liu et al., 2025), autonomous learning for diagnostic optimization (Du et al., 2024), reasoning in clinical conversations (Johri et al., 2023), adaptive multi-agent collaboration (Kim et al., 2024), the role of PCP in diagnosis (Wang et al., 2024a), and their integration into AI hospital environments (Fan et al., 2024). **Specialist Agent** represents domain-specific medical experts such as cardiologists, radiologists, and hematologists, for handling complex cases and contributing expert knowledge to diagnostic and treatment decision-making. Specialist agents require high-precision reasoning, deep medical expertise, and the ability to collaborate effectively in multidisciplinary team (MDT). Many works (Chen et al., 2024e; Kim et al., 2024) highlight the benefits of structured expertise, domain-specific knowledge, and coordinated decision-making in the AI Hospital. **Therapist Agent** provides emotional support, psychological intervention and psychotherapy (Wang et al., 2024b; Qiu and Lan, 2024; Chen et al., 2023b). **Nurse Agent** facilitates triage, basic care and patient coordination (Bao et al., 2024; Li et al., 2024b). **Medical Technician Agents** aid diagnostic procedures, ensuring accurate test results (Schmidgall et al., 2024b). **Medical students & Examiner agent** Simulate clinical training to improve medical history collection and diagnostic skills (Li et al., 2024d; Yao et al., 2024b).

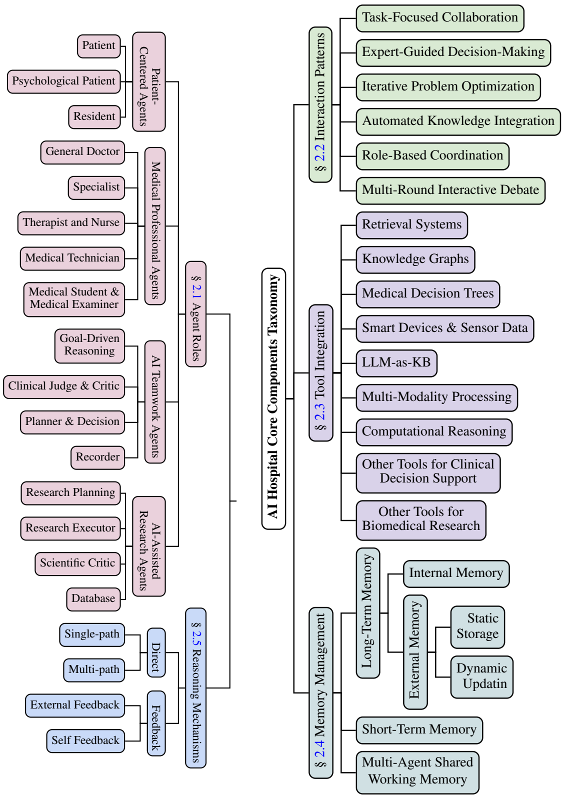

Figure 2: Taxonomy of AI hospital core components.

Medical AI Teamwork Agents collaborate to tackle complex AI hospital tasks beyond a single agent's capacity. They handle information extraction, reasoning, and decision-making in disease analysis, diagnosis, patient triage, medical planning, and final decisions. **Goal-Driven Reasoning Agent** coordinates multi-step reasoning using structured pipelines, dual-agent frameworks, and symbolic reasoning (Yu et al., 2024; Hong et al., 2024; Shi et al., 2024b). **Clinical Judge Agent** ensures AI-driven diagnoses meet accuracy, effectiveness, and guideline adherence (Johri et al., 2023; Yue et al., 2024a). **Critic Agent** refines reasoning, mitigates biases, and enhances reliability through structured feedback (Ke et al., 2024; Hong et al., 2024). **Planning Agent** decomposes tasks, optimizes workflows, and improves triage and structured conversations (Yue et al., 2024a; Shi et al., 2024a). **Decision Agent** mediates conflicting assessments and synthesizes insights for coherent, evidence-based diagnoses (Tang et al., 2023; Wang et al., 2024g). **Recording Agent** logs key medical insights (Ke et al., 2024; Yu et al., 2024).

AI-Assisted Research Agents optimize new knowledge discovery, research support, and scientific review. **Research Planning Agent** plays a crucial role in structuring research tasks and ensuring efficient problem decomposition in complex domains, leveraging hierarchical decision-making and adaptive optimization to refine research strategies and enhance scientific impact (Swanson et al., 2024; Xiao et al., 2024). **Research Executor Agent** facilitates clinical research by assisting in hypothesis testing, statistical analysis, and experiment interpretation, leveraging domain-specific expertise to optimize research workflows and minimize execution failures (Swanson et al., 2024; Xiao et al., 2024). **Scientific Critic Agent** is responsible for assessing the quality and validity of AI-generated solutions, ensuring reliable decision-making in research and clinical settings(Xiao et al., 2024). **Database Agent** is designed to retrieve, manage, and integrate medical information for improved decision-making (Shi et al., 2024b).

2.2 Interaction Patterns

AI Hospital employs different interaction patterns to enhance efficiency, reliability, and decision-making. **Task-Focused Collaboration** decomposes complex medical tasks into structured subtasks for efficiency and consistency. Modular architectures follow predefined workflows to accomplish tasks, such as the ERRG workflow (Extract, Retrieve, Rewrite, Generate) (Li et al., 2024d). Multi-agent systems like AIPatient(Yu et al., 2024), ClinicalAgent (Yue et al., 2024a), and EHRAgent (Shi et al., 2024b) assign roles and execute tasks sequentially to enhance reasoning and decision-making. **Expert-Guided Decision-Making** ensures AI-driven medical decisions are clinically reliable. Multiple studies (Du et al., 2024; Chen et al., 2024e; Kim et al., 2024; Tang et al., 2023) emphasize expert integration in decision-making, research, and medical education, ensuring domain expertise and consensus validation. **Iterative Problem Optimization (IPO)** refines problem-solving through feedback loops. AI agents iteratively adjust queries (Yu et al., 2024), refine diagnostic interactions via conversational and reflection-based corrections (Du et al., 2024; Bao et al., 2024), and critique each other's reasoning (Tang et al., 2023). Programming agents iteratively enhance code accuracy (Shi et al., 2024b). **Automated Knowledge Integration (AKI)** merges diverse medical knowledge and patient data for accurate, context-aware decision-making. Techniques include knowledge-enhanced retrieval (Shi et al., 2024a), memory-based integration (Liao et al., 2024), and Directed Acyclic Graph (DAG)-based structuring (Du et al., 2024). Multi-modal approaches combine structured and unstructured EHR data, sensor inputs, and medical evidence (Yang et al., 2024a), while team-based models apply adaptive fusion (Wang et al., 2024a), confidence validation (Lu et al., 2024), and structured reasoning (Hong et al., 2024). **Role-based Coordination** assigns AI agents specific roles (e.g., physicians, therapists, or patients) to simulate medical interactions and enhance diagnosis, training, and decision-making (Du et al., 2024; Wang et al., 2024b; Qiu and Lan, 2024). Multi-disciplinary AI teams integrate specialists' insights into comprehensive diagnoses (Wang et al., 2024g; Chen et al., 2024e). Systems like AgentClinic (Schmidgall et al., 2024b) and Agent Hospital (Li et al., 2024b) expand role-based AI applications to triage, reception, and follow-ups. **Multi-Round Interactive Debate** fosters structured discussions where AI agents critique, resolve disagreements, and refine conclusions (Fan et al., 2024; Li et al., 2023b; Kim et al., 2024). Approaches employ voting (Tang et al., 2023), debate strategies (Smit et al., 2023), and confidence-based stopping (Lu et al., 2024). AI-driven research teams apply debate mechanisms to synthesize findings (Swanson et al., 2024).

2.3 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 doctor agents in contextual reasoning (Du et al., 2024; Kim et al., 2024). **Knowledge graphs** structure medical knowledge into interconnected networks, enabling AI systems to navigate relationships between symptoms, treatments, and medical histories for informed decision support (Li et al., 2024d; Yu et al., 2024; Chen et al., 2024e). **Medical decision trees** provide structured diagnostic pathways, ensuring AI-driven recommendations align with established clinical guidelines and expert knowledge (Yang et al., 2024a; Li et al., 2023a). **LLM-as-KB** transforms LLMs into dynamic knowledge repositories, allowing AI to synthesize medical insights beyond static databases (Yue et al., 2024a; Frisoni et al., 2024). **Smart devices and sensor data** integration facilitate real-time health monitoring, merging wearable data with EHR insights to enhance predictive analytics and personalized care (Yang et al., 2024a; Abbasian et al., 2023). **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 (Li et al., 2024d; Yang et al., 2024a; Li et al., 2024a). **Computational reasoning tools** equip AI with logical inference and code execution capabilities, supporting automated clinical research and data-driven modeling (Wang et al., 2024f; Hong et al., 2024). Finally, some **other clinical decision support tools** optimize diagnostic accuracy by leveraging external APIs, existing predictive models/systems, and structured reporting systems (Wang et al., 2024a; Li et al., 2024a). And some **other biomedical research tools** accelerate drug discovery and genomic analysis, enabling AI-powered advancements in computational biology and molecular medicine (Swanson et al., 2024; Jin et al., 2023; Liu et al., 2024).

2.4 Memory Management

AI Hospital leverages structured memory management for adaptive learning and decision-making.

Long-Term Memory (LTM) retains knowledge across sessions, integrating internal model updates and external databases for enhanced reasoning. **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, Li et al. (2024d) 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. Wang et al. (2024e) integrates internal memory by fine-tuning ChatGPT with real patient clinical records, resulting in more accurate adverse event and drug predictions. **External Memory** supplements AI hospital systems with structured knowledge from databases, knowledge graphs, and retrieval systems while enabling real-time adaptation. **Static Storage** maintains long-term, structured knowledge, such as NIH resources for disease-specific agents (Wang et al., 2024a), CCD for patient history (Wang et al., 2024d), and structured ESI manuals (Lu et al., 2024). Medical knowledge databases, textbooks, and diagnostic guidelines serve as stable references (Yang et al., 2024a; Shi et al., 2024a; Yue et al., 2024b), while drug knowledge graphs and clinical trial registries support evidence-based decision-making (Chen et al., 2024e; Yue et al., 2024a; Liu et al., 2024). **Dynamic Updating** integrates real-time knowledge via retrieval systems and APIs, refining AI behavior with expert feedback (Louie et al., 2024), synchronizing clinical guidelines (Yang et al., 2024a), and leveraging PubMed or GitHub updates (Wang et al., 2024f). Additionally, long-term memory enhances task execution by retrieving past cases (Shi et al., 2024b; Schmidgall et al., 2024b; Bao et al., 2024), preserving user preferences like recurring health concerns for personalized responses.

Short-Term Memory (STM) and **Multi-Agent Shared Working Memory (WM)** serve complementary roles in AI hospitals and medical dialogue systems, ensuring context retention, reasoning consistency, and collaborative decision-making. STM is a temporary, agent-specific memory that maintains coherence during task execution but is cleared afterward (Liu et al., 2025). Medical dialogue systems use dialogue history, entity extraction, or summaries to mitigate forgetfulness and enhance reasoning. In contrast, WM is a globally shared memory facilitating knowledge synchronization, feedback integration, and structured reasoning across agents. It supports dynamic inference buffers, execution trace retention, and cross-agent coordination. For instance, Lu et al. (2024) updates summary reports for diagnostic consistency, while Hong et al. (2024) structures symbolic inference steps. WM also optimizes iterative decision-making (Kim et al., 2024; Xiao et al., 2024), reducing redundancy by storing shared task outcomes (Xiao et al., 2024). Feedback integration enhances refinement, as seen in expert voting (Tang et al., 2023), meta-doctor consolidation (Wang et al., 2024g), and structured critique cycles (Swanson et al., 2024).

2.5 Reasoning Mechanisms

Direct: derives conclusions through structured logic without external feedback. **Single-path** follows a linear progression, where each step builds on the previous one, as seen in ERRG (Li et al., 2024d), cognitive conceptualization maps (Wang et al., 2024d), ClientCAST (Wang et al., 2024b), and medical diagnostic frameworks like MDAgents (Kim et al., 2024) and expert systems (Yan et al., 2024). CoT-based approaches include AgentClinic (Schmidgall et al., 2024b), AI nurse simulators (Bao et al., 2024), CoT-driven coding (Wang et al., 2024f), least-to-most reasoning in clinical agents (Yue et al., 2024a), and Chain-of-Diagnosis models (Chen et al., 2024d). **Multi-path** enables parallel inference for flexible decision-making, integrating multi-agent systems like EvoPatient (Du et al., 2024), RareAgents (Chen et al., 2024e), MDAgents (Kim et al., 2024), and MedAgents (Tang et al., 2023). Other methods leverage multi-agent collaboration (Wang et al., 2024g), expert self-consistency (Li et al., 2024c), and symbolic reasoning (Wang et al., 2024a; Hong et al., 2024). Additionally, LLM planners (Liu et al., 2024) generate parallel solutions before validation, while simulated medical research meetings (Swanson et al., 2024) synthesize discussions into optimal decisions.

Feedback-Based: adjusts reasoning by integrating feedback to refine. **External Feedback** enhances AI agents by incorporating real-time data, expert input, and structured resources, enabling agents to refine their understanding through interactions and external tools (Chen et al., 2024d; Johri et al., 2023). Medical consultation systems iteratively update diagnoses through patient interactions, while decision-making agents query external resources like Phenomizer and DrugBank for real-time clinical knowledge (Li et al., 2024c). **Self Feedback** enables AI agents to refine reasoning internally by evaluating logic, correcting inconsistencies, and iteratively improving outputs (Louie et al., 2024; Yu et al., 2024). Reflection-based techniques such as Reflection CoT and self-play mechanisms further enhance AI models by structuring error analysis and collaborative discussions (Schmidgall et al., 2024b). Applications extend to code generation, drug discovery, medical research, and medical exam question generation (Wang et al., 2024f).

3 Applications

3.1 Simulating Specific Scenarios

Clinical Workflow Simulation employs multi-agent to model patient care, from consultation to diagnosis. Some works simulate the full consultation workflow, where patient, doctor, and evaluator agents interact. Liu et al. (2025) segmented consultations into four stages and identifies the weakest stage as the limiting factor, akin to Liebig's law. Johri et al. (2023) proposed CRAFT-MD, using doctor agents interacting with structured patient agents and an automatic grading system. Li et al. (2024c) developed MEDIQ, integrating abstention strategies, rationale generation, and self-consistency to refine diagnosis. Fan et al. (2024) introduced AI Hospital, where doctor agents engage in multi-round discussions, mediated by a Central Agent to resolve disagreements. Schmidgall et al. (2024b) presented AgentClinic, a multimodal benchmark incorporating cognitive biases and incomplete information to evaluate LLM-based doctor agents. Another direction expands simulations beyond consultation to the entire patient journey. Bao et al. (2024) developed PIORS, an outpatient reception system using a Service Flow-aware Medical Scenario Simulation framework to enhance department recommendations. Li et al. (2024b) proposed Agent Hospital, a fully autonomous system covering disease onset to recovery. Its MedAgent-Zero framework enables doctor agents to refine their diagnostic accuracy via case-based learning and RAG, mirroring real-world physicians' iterative knowledge refinement and boosting medical evaluation performance. Given the communication-centric nature of mental health care, a large body of work also focuses on **Psychological Counseling and Mental Health Interaction**, which can be viewed as a specialized form here. Examples include Roleplay-doh (Louie et al., 2024), which turns expert feedback into behavior rules; PATIENT-Ψ (Wang et al., 2024d), which incorporates CBT principles; and Chen et al. (2023b), which aligns interactions with DSM-5 criteria.

Multi-Disciplinary Medical Team Simulation replicates real-world medical teams' collaborative processes, optimizing communication, information sharing, and decision-making for complex clinical scenarios. For rare disease, Chen et al. (2024e) introduced RareAgents, where a patient agent presents symptoms, an attending physician agent assembles an MDT, and specialists iteratively

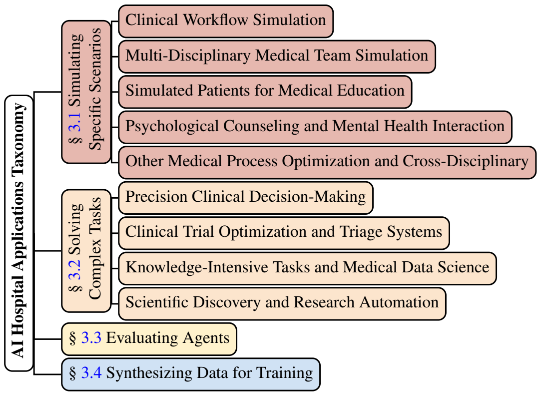

Figure 3: Taxonomy of AI hospital applications.

refine diagnoses using dynamic memory and medical toolkits. Similarly, Kim et al. (2024) proposed MDAgents, employing a hierarchical collaboration strategy where a single doctor handles simple cases, while MDTs, moderated by an external knowledge-integrating agent, address complex ones. Tang et al. (2023) introduced MEDAGENTS, structuring MDT collaboration into four phases—expert recruitment, independent analysis, collaborative consultation, and final decision-making—to enhance reasoning without training. In EHR modeling, Wang et al. (2024g) proposed ColaCare, where DoctorAgent processes structured EHR data with medical guidelines, while MetaAgent refines clinical decisions through iterative assessments, improving predictive modeling by integrating numerical predictions with textual reasoning.

Simulated Patients for Medical Education improve student training in communication, clinical reasoning, and diagnosis within a controlled setting. Advances in LLM-driven simulations enhance fidelity and interactivity. Du et al. (2024) introduced EvoPatient, a multi-agent framework where doctor-patient agents iteratively co-evolve using RAG and personality traits. Wei et al. (2024a) proposed MEDCO, integrating structured training, interdisciplinary collaboration, and multimodal inputs with memory and peer discussion modules. For assessment, Mehandru et al. (2024) proposed AI-SCE for process-focused training, while Yao et al. (2024b) introduced MedQA-CS with simulated student interactions and structured evaluations.

Other Medical Process Optimization and Cross-Disciplinary Simulation AI-driven methodologies have been explored for optimizing medical processes and enabling cross-disciplinary simulations. Swanson et al. (2024) introduced a multi-agent "Virtual Lab," where LLM-powered agents (e.g., principal investigator, biologist, scientific critic) collaborate using biomedical tools like ESM and AlphaFold-Multimer to design nanobody treatments for SARS-CoV-2 variants, showcasing AI's potential in accelerating interdisciplinary research. Similarly, Williams et al. (2023) proposed a generative AI-enhanced epidemic modeling platform, where LLM-driven agents autonomously assess health status and public health data to simulate pandemic dynamics, improving traditional agent-based modeling. These works demonstrate AI's role in advancing scientific discovery and public health modeling through intelligent agent-based decision-making.

3.2 Solving Complex Tasks

Many AI Hospital works leverage multi-agent frameworks to enhance diagnosis, triage, research, and discovery in dynamic clinical settings.

Clinical Decision-Making: AI hospitals improve diagnostic accuracy and transparency, especially for rare or complex diseases. Systems like RareAgents (Chen et al., 2024e), MMedAgent (Li et al., 2024a), and DrHouse (Yang et al., 2024a) integrate tools, memory, and retrieval for consistent, multimodal reasoning. Others focus on interpretability: DiagnosisGPT (Chen et al., 2024d), ArgMed-Agents (Hong et al., 2024), and MedAgents (Tang et al., 2023) use structured reasoning or argumentation to reduce bias and enhance trust.

Triage and Clinical Trials: Agent-based systems like TriageAgent (Lu et al., 2024), PIORS (Bao et al., 2024), and ClinicalAgent (Yue et al., 2024a) improve emergency triage, outpatient routing, and trial matching using guideline-based retrieval and reasoning strategies.

Knowledge-Intensive Workflows: AI agents support data science tasks such as EHR analysis (Shi et al., 2024b), code generation (Wang et al., 2024f), fact-checking (Yue et al., 2024b), and question generation (Yao et al., 2024a), streamlining clinical research.

Scientific Discovery: Multi-agent labs like VirtualLab (Swanson et al., 2024), CellAgent (Xiao et al., 2024), and DrugAgent (Liu et al., 2024) automate biomedical discovery, integrating reasoning agents with domain tools to accelerate hypothesis generation, molecular analysis, and drug development.

3.3 Evaluating Agents

AI hospital evaluations are shifting from static benchmarks to interactive, multi-agent simulations that capture real-time reasoning, collaboration, and patient engagement (Johri et al., 2023; Schmidgall et al., 2024b; Li et al., 2024c). Recent work emphasizes state-aware evaluation, using patient simulators like SAPS (Liao et al., 2024) and role-play settings (Louie et al., 2024; Wang et al., 2024b) to test an agent's adaptability and coherence across turns. Multi-agent frameworks such as AI Hospital (Fan et al., 2024) and ClinicalLab (Yan et al., 2024) assess inter-agent collaboration, dispute resolution, and cross-department knowledge exchange. Multimodal evaluation is also gaining traction: MMedAgent (Li et al., 2024a) combines imaging and text-based reasoning, while others assess tool-assisted clinical calculations (Khandekar et al., 2024). Finally, OSCE-style benchmarks like MedQA-CS (Yao et al., 2024b), OSCEBot (Pereira et al., 2023), and AI-SCE (Mehandru et al., 2024) offer comprehensive, scenario-based evaluations of real-world clinical skills.

3.4 Synthesizing Data for Training

Synthetic data generation in AI hospitals supports realistic, privacy-preserving training for medical LLMs. Multi-agent co-evolution frameworks (Du et al., 2024; Li et al., 2024b) simulate diagnostic dialogues, refine agent reasoning, and improve generalization to benchmarks. NoteChat (Wang et al., 2023a) transforms clinical notes into role-played, polished conversations via planning, simulation, and feedback. AMIE (Tu et al., 2024) uses self-play and auto-feedback to enhance history-taking and reasoning. These methods reduce annotation costs while maintaining clinical validity, enabling scalable training for downstream applications.

4 Key Challenges & Future Directions

[1] **Agent Roles** Recontextualizing NLP models as clinical agents demands role-consistent behaviors reflecting real-world complexity. Doctor agents must demonstrate diverse diagnostic reasoning, clinical decision-making, and personalized communication styles, while patient agents require nuanced disclosure of medical histories influenced by social determinants of health. Techniques like memory modules, inverse reinforcement learning, and dynamic knowledge graphs will be essential to capture these complexities. A critical interdisciplinary challenge is integrating social, psychological, and behavioral theories into NLP frameworks, enhancing realism, fairness, and patient diversity in clinical simulations.

[1] Due to space limitations, we include detailed discussions and Table 2 and 3 for this section in Appendix A.

Interaction Patterns The design of interaction patterns in multi-agent healthcare remains challenging, particularly regarding meaningful human participation. Current studies typically limit human roles to evaluation, rather than interactive partners, restricting NLP's practical impact. A key direction is exploring hybrid human-AI interaction models, clearly distinguishing human contributions from autonomous agent behaviors. Techniques from fields such as game theory, human-computer interaction, and cognitive science could enrich NLP's methods for modeling realistic and beneficial human-agent collaboration.

Tool Integration While current works integrate diverse tools, systematic evaluation frameworks remain underdeveloped, limiting assessment of their true interdisciplinary impact. Future research should move beyond static NLP benchmarks, leveraging AI hospital ecosystems as dynamic environments to rigorously evaluate how well integrated tools improve real clinical workflows and outcomes. Interdisciplinary validation frameworks must assess tools' contributions to patient safety, decision quality, and healthcare accessibility.

Memory Management remains critical for integrating NLP within longitudinal patient care. While current models rely on static EHRs and retrieval-augmented generation, accurately capturing temporal disease progression and dynamic patient profiles requires advanced interdisciplinary solutions. Temporal knowledge graphs and dynamic memory retrieval methods should be explored to align NLP outputs with patient comprehension levels, enabling more personalized, adaptive, and clinically relevant interactions.

Reasoning Mechanisms Most NLP reasoning approaches remain limited to single-path inference, insufficient for complex, uncertain clinical scenarios. Future research should integrate adaptive reasoning frameworks combining single-path, multi-hop, and probabilistic approaches, leveraging insights from clinical reasoning literature. Bayesian inference, Markov Decision Processes (MDPs), and decision-theoretic methods from cognitive science and medicine could enhance NLP agents' ability to handle clinical uncertainty, improve safety, and support rigorous interdisciplinary evaluations of clinical reasoning.

Simulating Specific Scenario & Solving Com-

plex Tasks AI hospital simulations must address challenges in modeling realistic clinical scenarios extending beyond acute patient visits, including chronic disease management and public health emergencies. Interdisciplinary NLP research must incorporate socio-behavioral dynamics and broader environmental contexts to accurately represent real-world complexity. Ensuring robustness in multi-agent architectures also requires addressing technical bottlenecks such as hallucinations, biases, and computational scalability. Enhanced error-handling, uncertainty quantification, and human expert oversight are critical for meaningful interdisciplinary deployment in healthcare.

Evaluating Agents Evaluating NLP-driven clinical agents requires shifting from accuracy-focused metrics toward interdisciplinary frameworks aligning with real-world medical practice. Future evaluations should incorporate measures reflecting clinical utility, usability, patient-centered outcomes, and cost-effectiveness, leveraging feedback loops involving clinicians and patients. Balancing inference efficiency and resource costs, as well as integrating medical-specific domain knowledge with general-purpose LLM capabilities, represents an interdisciplinary evaluation challenge critical to impactful real-world deployment.

Synthesizing Data for Training Synthesizing realistic, unbiased, and privacy-compliant data remains challenging for training NLP-driven medical agents. While reinforcement learning and self-play approaches offer promise, applying them in clinical contexts faces limitations from data scarcity and ethical concerns. Future interdisciplinary directions include dynamic synthetic data generation through multi-agent collaboration, multimodal integration, and fairness-driven evaluation metrics. Interdisciplinary collaboration involving domain experts, ethicists, and clinicians is essential for generating synthetic data capable of reliably informing real-world clinical practice.

Governance, Ethics, and the Roles of AI Researchers and Medical Practitioners The deployment of AI hospital systems introduces significant governance and ethical challenges related to transparency, responsibility allocation, security, and equitable healthcare access. As these systems increasingly influence medical decision-making, establishing clear accountability frameworks for errors or adverse outcomes becomes critical. In particular, implicit biases in resource allocation—reflected in training data or agent behavior—may exacerbate social inequalities if left unaddressed. A robust governance framework must ensure compliance with ethical standards, protect patient privacy, and support interdisciplinary oversight.

Transparent version control and model evolution tracking are necessary to monitor changes in behavior, mitigate unintended consequences, and ensure reproducibility across deployments. Addressing these challenges will require collaboration among NLP researchers, clinicians, ethicists, and policymakers, as well as international cooperation to establish consistent norms and regulations.

From the perspective of AI&NLP researchers, a key challenge lies in fully leveraging the AI hospital as a testbed to iteratively address the technical, behavioral, and evaluation challenges discussed throughout Section 4. This includes simulating failures, validating safety interventions, and aligning agent behaviors with domain expectations—not only for technical excellence but for responsible impact For medical practitioners, the challenge is to integrate AI systems into clinical practice in a way that improves equity and efficiency without increasing burden or disrupting workflows. AI hospitals must be designed not to replace, but to assist human judgment—enhancing clinician decision-making through trustworthy collaboration. This requires deep involvement of healthcare professionals throughout the system design and testing process. By embedding clinical expertise into development, AI hospitals can be grounded in real-world needs, bridging the gap between language-based AI systems and practical, ethical medical applications.

5 Conclusion

As large language models increasingly take on agentic roles, AI hospitals provide a compelling framework for reimagining NLP in complex, high-stakes domains like healthcare. By simulating multi-agent clinical workflows and enabling dynamic, role-based evaluation, these systems move beyond static benchmarks—offering new ways to assess reasoning, collaboration, and safety in real-world contexts. More broadly, AI hospitals illustrate how recontextualizing NLP systems within interdisciplinary environments can surface both limitations and opportunities. They challenge us to bridge linguistic modeling with clinical reasoning, decision-making under uncertainty, and societal considerations such as fairness, trust, and impact.

6 Limitations

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 original papers and code repositories for full implementation details. Our literature review mainly covers *ACL, NeurIPS, ICLR, ICML, AAAI, select medical journals, and preprints (arXiv, medRxiv, bioRxiv), so some relevant work may be overlooked. Given the field's rapid evolution, we remain committed to updating our perspectives and incorporating new advancements.

References

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.

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.

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

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.

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.

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.

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.

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.

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.

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*.

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*.

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.

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.

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

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.

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.

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.

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

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

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

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

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024d. Cod, towards an interpretable medical agent using chain of diagnosis. *ArXiv*, abs/2407.13301.

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. 2023a. Huatuogpt-ii, one-stage training for medical adaption of llms. *ArXiv*, abs/2311.09774.

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

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

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

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, Shao-Kang 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.

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

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.

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*.

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.

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*.

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

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*.

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.

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.

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

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

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

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.

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*.

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

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.

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.

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.

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

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

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

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

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.

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.

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*.

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.

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.

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.

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.

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*.

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*.

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*.

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.

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

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

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

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

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

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

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

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

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*.

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.

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.

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*.

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*.

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.

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).

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.

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*.

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.

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

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).

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.

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*.

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.

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.

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*.

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.

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.

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*.

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

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

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

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. 2024b. 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*.

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.

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*.

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.

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.

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

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*.

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.

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.

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*.

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*.

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*.

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.

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.

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.

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.

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

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

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

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

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024d. 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.

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

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

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

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

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

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

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

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*.

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*.

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.

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

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*.

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

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.

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*.

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.

Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024a. 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.

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

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

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

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

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*.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024a. 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.

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

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.

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.

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*.

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*.

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*.

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	(Liu et al., 2025)
2	Leveraging Large Language Model as Simulated Patients for Clinical Education	arXiv	No	(Li et al., 2024d)
3	A GPT-Powered Chatbot as a Simulated Patient to Practice History Taking	JMIR Med Edu	No	(Holderried et al., 2024)
4	Designing and building OSCEBot ® for virtual OSCE – Performance evaluation	Med Edu Online	No	(Pereira et al., 2023)
5	Roleplay-doh: Enabling domain-experts to create LLM-simulated patients	EMNLP24	link	(Louie et al., 2024)
6	AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow	arXiv	link	(Yu et al., 2024)
7	LLMs Can Simulate Standardized Patients via Agent Coevolution	arXiv	link	(Du et al., 2024)
8	PATIENT-Ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals	EMNLP24	link	(Wang et al., 2024d)
9	Towards a Client-Centered Assessment of LLM Therapists by Client Simulation	arXiv	link	(Wang et al., 2024b)
10	Automatic Interactive Evaluation for LLMs with State Aware Patient Simulator	arXiv	link	(Liao et al., 2024)
11	Guidelines For Rigorous Evaluation of Clinical LLMs For Conversational Reasoning	medRxiv	link	(Johri et al., 2023)
12	RAREAGENTS: Autonomous Multi-disciplinary Team for Rare Disease Diagnosis	arXiv	No	(Chen et al., 2024e)
13	TWIN-GPT: Digital Twins for Clinical Trials via Large Language Model	TOMM24	No	(Wang et al., 2024e)
14	Interactive Agents: Simulating Counselor-Client Psychological Counseling	arXiv	link	(Qiu and Lan, 2024)
15	DrHouse: An LLM-empowered Diagnostic Reasoning System	IMWUT	No	(Yang et al., 2024a)
16	MDAgents: Adaptive Collaboration of LLMs for Medical Decision-Making	NeurIPS 2024	link	(Kim et al., 2024)
17	MEDAGENTS: Large Language Models as Collaborators for Medical Reasoning	ACL Findings 2024	link	(Tang et al., 2023)
18	ColaCare: Enhancing EHR Modeling through LLM Multi-Agent Collaboration	arXiv	link	(Wang et al., 2024g)
19	Mitigating Cognitive Biases in Clinical Decision-Making via Multi-Agent LLMs	JMIR	No	(Ke et al., 2024)
20	AgentClinic: A Multimodal Agent Benchmark for Simulated Clinical Environments	arXiv	link	(Schmidgall et al., 2024b)
21	TRIAGEAGENT: Multi-Agents for LLM-Based Clinical Triage	EMNLP Findings 2024	link	(Lu et al., 2024)
22	PIORS: Personalized Intelligent Outpatient Reception Using Multi-Agents	arXiv	link	(Bao et al., 2024)
23	Can Large Language Models Replace Data Scientists in Clinical Research?	arXiv	No	(Wang et al., 2024f)
24	The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies	BioRxiv	link	(Swanson et al., 2024)
25	MEDCO: Medical Education Copilots Using Multi-Agent Framework	arXiv	No	(Wei et al., 2024a)
26	Should We Be Going MAD? Multi-Agent Debate Strategies for LLMs	ICML2024	link	(Smit et al., 2023)
27	Beyond Direct Diagnosis: Multi-Specialist Agent Consultation for Diagnosis	arXiv	No	(Wang et al., 2024a)
28	ClinicalAgent: Clinical Trial Multi-Agent System with LLM Reasoning	BCB '24	link	(Yue et al., 2024a)
29	Enhancing Clinical Trial Patient Matching via Multi-Agent Knowledge Augmentation	arXiv	No	(Shi et al., 2024a)
30	ArgMed-Agents: Explainable Clinical Decision Reasoning via Argumentation	BIBM2024	No	(Hong et al., 2024)
31	Synergistic Multi-Agent Framework with Trajectory Learning	AAAI25	link	(Yue et al., 2024b)
32	Empowering Biomedical Discovery with AI Agents	Cell	No	(Gao et al., 2024)
33	MedDM: LLM-executable Clinical Guidance Tree for Decision-Making	arXiv	No	(Li et al., 2023a)
34	Text2MDT: Extracting Medical Decision Trees from Texts	arXiv	No	(Zhu et al., 2024)
35	BioKGBench: A Knowledge Graph Benchmark	arXiv	link	(Lin et al., 2024)
36	Medical Graph RAG: Safe LLMs via Graph Retrieval-Augmented Generation	arXiv	link	(Wu et al., 2024)
37	HeCiX: Integrating Knowledge Graphs and LLMs for Biomedical Research	arXiv	No	(Kulkarni et al., 2024)
38	KRAGEN: Knowledge Graph-Enhanced RAG for Biomedical Problem Solving	Bioinformatics	link	(Matsumoto et al., 2024)
39	Accelerating Medical Knowledge Discovery via Automated Knowledge Graphs	KGSWC 2024	No	(Khalid et al., 2024)
40	Augmented Non-Hallucinating LLMs as Medical Information Curators	npj Digital Medicine	No	(Gilbert et al., 2024)
41	Benchmarking Retrieval-Augmented Generation for Medicine	ACL Findings 2024	link	(Xiong et al., 2024a)
42	Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions	arXiv	link	(Xiong et al., 2024b)
43	Almanac — Retrieval-Augmented Language Models for Clinical Medicine	NEJM AI	link	(Zakka et al., 2024)
44	Augmenting Black-box LLMs with Medical Textbooks for Biomedical QA	EMNLP Findings 2024	link	(Wang et al., 2023b)
45	To Generate or Retrieve? Effectiveness of Artificial Contexts in Medical QA	ACL 2024	link	(Frisoni et al., 2024)
46	AgentMD: Empowering Language Agents for Risk Prediction	arXiv	link	(Jin et al., 2024)
47	MedCalc-Bench: Evaluating LLMs for Medical Calculations	NeurIPS 2024	link	(Khandekar et al., 2024)
48	Augmenting ChatGPT with Clinician-Informed Tools for Medical Calculations	medRxiv	No	(Goodell et al., 2023)
49	GeneGPT: Augmenting LLMs with Domain Tools for Biomedical Information	Bioinformatics	link	(Jin et al., 2023)
50	EHRAgent: Code-Empowered LLMs for Few-shot Complex Tabular Reasoning	EMNLP 2024	link	(Shi et al., 2024b)
51	MMedAgent: Learning to Use Medical Tools with Multi-modal Agent	EMNLP Findings 2024	link	(Li et al., 2024a)
52	Conversational Health Agents: A Personalized LLM-Powered Agent Framework	arXiv	link	(Abbasian et al., 2023)
53	PathAsst: A Generative AI Assistant for Pathology Analysis	AAAI Technical Track	link	(Sun et al., 2023)
54	GPT-agents Based on Medical Guidelines for Traumatic Brain Injury Rehabilitation	Scientific Reports	No	(Li et al., 2024e)
55	CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis	arXiv	No	(Xiao et al., 2024)
56	DrugAgent: Automating AI-Aided Drug Discovery via LLM Multi-Agent Collaboration	arXiv	No	(Liu et al., 2024)
57	Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents	arXiv	No	(Li et al., 2024b)
58	AI Hospital: Benchmarking LLMs in a Multi-agent Medical Interaction Simulator	COLING 2025	link	(Fan et al., 2024)
59	ClinicalLab: Aligning Agents for Multi-Departmental Clinical Diagnostics	arXiv	link	(Yan et al., 2024)
60	LLM-empowered Chatbots for Psychiatrist and Patient Simulation	arXiv	No	(Chen et al., 2023b)
61	MediQ: Question-Asking LLMs and a Benchmark for Interactive Clinical Reasoning	NeurIPS 2024	link	(Li et al., 2024c)
62	Epidemic Modeling with Generative Agents	arXiv	link	(Williams et al., 2023)
63	NoteChat: A Dataset of Synthetic Patient-Physician Conversations	ACL 2024 Findings	link	(Wang et al., 2023a)
64	Evaluating Large Language Models as Agents in the Clinic	npj Digital Medicine	No	(Mehandru et al., 2024)
65	MedQA-CS: Benchmarking LLMs Clinical Skills Using an AI-SCE Framework	arXiv	link	(Yao et al., 2024b)
66	Towards Conversational Diagnostic AI	arXiv	No	(Tu et al., 2024)
67	LLM-based Medical Assistant Personalization with Short- and Long-Term Memory	NAACL 2024	link	(Zhang et al., 2023)
68	CoD: Towards an Interpretable Medical Agent Using Chain of Diagnosis	arXiv	link	(Chen et al., 2024d)
69	Multi-Agent Conversation Framework Enhances Rare Disease Diagnosis in LLMs	Preprint	link	(Li et al., 2023b)
70	MEDAIDE: Towards an Omni Medical Aide via Specialized LLM-based Multi-Agent Collaboration	Preprint	No	(Wei et al., 2024b)
71	RAG-Gym: Optimizing Reasoning and Search Agents with Process Supervision	Preprint	link	(Xiong et al., 2025)
72	MCQG-SRefine: Multiple Choice Question Generation and Evaluation with Iterative Self-Critique, Correction, and Comparison Feedback	NAACL 2025	link	(Yao et al., 2024a)

Key Challenges	Future Research Directions
§ 4 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.
§ 4 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.
§ 4 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.
§ 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.
§ 4 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.

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

Key Challenges	Future Research Directions
§ 4 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.
§ 4 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.
§ 4 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.
§ 4 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 3: Key Challenges and Future Directions for different applications in AI Hospital.

A Key Challenges and Future Directions

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 (Kim et al., 2024). 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 (Wang et al., 2023a). Subsequent work may consider better integrating STM/LTM/WM modules to maintain contextual coherence (Zhang et al., 2023). At the same time, recent advancements in role-playing (Chen et al., 2024a) and personalization (Chen et al., 2024a,c; Zhang et al., 2024) methods in the general NLP domain, such as interview-driven persona modeling (Park et al., 2024) and expert feedback-based refinements (Louie et al., 2024), can be leveraged to improve agent behavior.

Another key aspect is managing information asymmetry, a fundamental characteristic of real-world medical conversations (Ariss, 2009; Greco, 2020). Doctor agents seek comprehensive patient information, whereas patient agents may selectively withhold certain details due to privacy concerns or psychological barriers (Gill and Maynard, 2006). 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 (Lehtinen, 2013). 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) (Chadi and Mousannif, 2022) 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 (Snoswell et al., 2024). 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 (Yu et al., 2024; Bakkum et al., 2024). To address this issue, demographic attributes should be supplemented with other factors, such as social determinants of health (SDOH) (Ong et al., 2024). 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.

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 (Balogh et al., 2015; Dhawale et al., 2017; Hu et al., 2024b). Current approaches rely on end-to-end LLM predictions without explicit mathematical modeling. By leveraging some methodologies like game theory (Blake and Carroll, 2016; Sun et al., 2025; Djulbegovic et al., 2015; Glycopantis and Stavropoulou, 2018), we can better model asymmetric information challenges in medical interactions. 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 (Gerstgrasser and Parkes, 2023) to optimize information exchange, with doctor agents guiding patient agents toward informative disclosures. Evolutionary game theory (Bloembergen et al., 2015) 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 (Verma et al., 2019; Chatzimichail and Hatjimihail, 2023).

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 (Wang et al., 2024c). 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 (Xiong et al., 2024a,b, 2025; Zakka et al., 2024; Wang et al., 2023b; Li et al., 2024e), GraphRAG (Lin et al., 2024; Wu et al., 2024; Kulkarni et al., 2024; Matsumoto et al., 2024; Khalid et al., 2024; Gilbert et al., 2024), and LLM-as-KB (Frisoni et al., 2024) 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.

Memory Management Existing research largely relies on static EHRs as memory to represent patients, often utilizing RAG or GraphRAG-based methods (Yu et al., 2024) 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 (Xie et al., 2022). 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 (Rasmussen et al., 2025) 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 (Chen et al., 2024b). 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 (Singh et al., 2024; Zhang et al., 2022). 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 (Cai et al., 2023), 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.

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 (Xu et al., 2024; Huang et al., 2025b; Faray de Paiva et al., 2025; Tran et al., 2024; Hu et al., 2024a). 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 (Balogh et al., 2015; Alli et al., 2024). 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 (Bennett and Hauser, 2013; Polotskaya et al., 2024). 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.

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 [2]. However, current systems lack flexibility to model such disruptions, limiting realism (Gürcan, 2024). 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 (Yang et al., 2024b; Al Owayyed et al., 2024). Integrating

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

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 (Bertl et al., 2023), such as LLMs hallucination generation (Huang et al., 2025a; Zuo and Jiang, 2024; Li et al., 2023c), 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 (Quinn et al., 2021). 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 (Balogh et al., 2015). 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, 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.

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 (Tam et al., 2024). 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 (Baines et al., 2018). These social evaluation mechanisms have not yet been fully integrated into AI hospital system assessments (Moy et al., 2024). Additionally, drawing inspiration from the Turing test (Nov et al., 2023), 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 (Snell et al., 2024). However, in AI hospital research, little attention has been given to how computational resource consumption impacts the practical value of a system (Fan et al., 2024; Smit et al., 2023). 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 (Achiam et al., 2023) and LLaMA (Dubey et al., 2024), with limited exploration of medical-specific LLMs like DoctorGLM (Xiong et al., 2023), HuatuoGPT (Chen et al., 2023a), BianQue (Chen et al., 2023c), BioLLaMA (Tran et al., 2023), BioMistral (Labrak et al., 2024), and Baichuan-M1 (Wang et al., 2025). Some studies, such as MedQA-CS (Yao et al., 2024b), 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" (Frisoni et al., 2024) 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.

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 (DeepSeek-AI et al., 2025), have demonstrated that models can continuously improve through reinforcement learning (RL) (Jayaraman et al., 2024) in specific environments without supervised data, AI hospitals, as complex medical environments, 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 (Li et al., 2024b; Ouyang et al., 2022; Rafailov et al., 2023; Yao et al., 2023; Mishra et al., 2024). 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 (Giuffré and Shung, 2023). 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 (Arora et al., 2023). 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 (Acosta et al., 2022). 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 (Ueda et al., 2023; Schmidgall et al., 2024a).