Abdul Mohaimen Al Radi[1], Xu Cao[1], Fanyang Yu[1], Yuyuan Liu[1], Fengbei Liu[1], Chong Wang[1], Yuanhong Chen[1], Jintai Chen[1], Hu Wang[1], Yanda Meng[1], Zhenyi Wang[1], Chen Chen[1], Mubarak Shah [1], Tianyu Han[1], Christos Davatzikos[1], MacLean P. Nasrallah[1], and Yu Tian[1]

[1]Affiliation not available

September 08, 2025

Abstract

In less than three years, large language models (LLMs) have advanced from passive responders to agentic systems capable of planning, acting, and collaborating with other agents. In medicine, these capabilities open the door to systems that can assist clinicians, coordinate care, and adapt to complex, real-world workflows. But the high-stakes nature of healthcare demands a careful examination of their potential: understanding not only the range of applications but also the challenges of data privacy, ethical responsibility, and safe deployment. This survey systematically analyzes over 140 studies from 2022 to 2025 on LLM-based medical agents, offering three key contributions. First, we introduce a unified taxonomy spanning application domains, autonomy levels, and integration of tools and knowledge, effectively categorizing use cases across various biomedical areas. Second, we explore how state-of-the-art agents integrate biomedical expertise, couple LLM reasoning with external resources like clinical databases, electronic health records (EHRs), APIs, and implement human-in-the-loop mechanisms to mitigate hallucinations and biases. Third, we synthesize crossdomain insights from fields such as education, robotics, and automated scientific discovery, highlighting transferable design principles for improved reliability and interpretability. Our analysis identifies persistent challenges, including hallucinations, biases, data privacy risks, and regulatory complexities, and reviews promising solutions to address them. Finally, we propose open research directions toward creating trustworthy, regulation-compliant agents that augment rather than replace clinical expertise, providing researchers, practitioners, and policymakers a comprehensive roadmap for advancing agentic AI in healthcare. A comprehensive list of agentic AI models studied in this work is available at here

Agentic Large-Language-Model Systems in Medicine: A Systematic Review and Taxonomy

Abdul Mohaimen Al Radi, Xu Cao, Fanyang Yu, Yuyuan Liu, Fengbei Liu, Chong Wang, Yuanhong Chen, Jintai Chen, Hu Wang, Yanda Meng, Zhenyi Wang, Chen Chen, Mubarak Shah, Tianyu Han, Christos Davatzikos, MacLean P. Nasrallah, Yu Tian

***Abstract*—In less than three years, large language models (LLMs) have advanced from passive responders to agentic systems capable of planning, acting, and collaborating with other agents. In medicine, these capabilities open the door to systems that can assist clinicians, coordinate care, and adapt to complex, real-world workflows. But the high-stakes nature of healthcare demands a careful examination of their potential: understanding not only the range of applications but also the challenges of data privacy, ethical responsibility, and safe deployment. This survey systematically analyzes over 140 studies from 2022 to 2025 on LLM-based medical agents, offering three key contributions. First, we introduce a unified taxonomy spanning application domains, autonomy levels, and integration of tools and knowledge, effectively categorizing use cases across various biomedical areas. Second, we explore how state-of-the-art agents integrate biomedical expertise, couple LLM reasoning with external resources like clinical databases, electronic health records (EHRs), APIs, and implement human-in-the-loop mechanisms to mitigate hallucinations and biases. Third, we synthesize cross-domain insights from fields such as education, robotics, and automated scientific discovery, highlighting transferable design principles for improved reliability and interpretability. Our analysis identifies persistent challenges, including hallucinations, biases, data privacy risks, and regulatory complexities, and reviews promising solutions to address them. Finally, we propose open research directions toward creating trustworthy, regulation-compliant agents that augment rather than replace clinical expertise, providing researchers, practitioners, and policymakers a comprehensive roadmap for advancing agentic AI in healthcare. A comprehensive list of agentic AI models studied in this work is available at https://github.com/AIM-Research-Lab/Awesome-AI-Agents-Medicine.**

***Index Terms*—Agentic AI, Large-language Models (LLMs), Medical AI, AI for Healthcare, Medical Imaging, Foundation Models.**

Abdul Mohaimen Al Radi, Zhenyi Wang, Chen Chen, Mubarak Shah, and Yu Tian are with the Department of Computer Science, University of Central Florida, Orlando, Florida, USA (e-mail: ab575577@ucf.edu; zhenyi.wang@ucf.edu; chen.chen@crcv.ucf.edu; shah@crcv.ucf.edu; yu.tian2@ucf.edu). Xu Cao is with the Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA (e-mail: xucao2@illinois.edu). Yuyuan Liu is with the Department of Engineering Science, University of Oxford, UK (e-mail: yuyuan.liu@eng.ox.ac.uk). Fengbei Liu is with the School of Electrical and Computer Engineering, Cornell University and Cornell Tech, New York, USA (e-mail: fl453@cornell.edu). Chong Wang is with the Department of Radiology, Stanford University, CA, USA (e-mail: chongwa@stanford.edu). Yuanhong Chen is with the Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia (e-mail: yuanhong.chen@adelaide.edu.au). Jintai Chen is with the Hong Kong University of Science and Technology (Guangzhou), Guangdong, China (e-mail: jtchen721@gmail.com). Hu Wang is with the MBZUAI, Abu Dhabi, United Arab Emirates. Yanda Meng is with the Computer Science Department at the University of Exeter, Exeter, UK (e-mail: y.m.meng@exeter.ac.uk). Tianyu Han is with the Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA (e-mail: tianyu.han@pennmedicine.upenn.edu). MacLean P. Nasrallah is with the Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, USA. (e-mail: Maclean.Nasrallah@pennmedicine.upenn.edu). Christos Davatzikos and Fanyang Yu are with the Artificial Intelligence in Biomedical Imaging Laboratory (AIBIL), Center for AI and Data Science for Integrated Diagnostics (AI2D), Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. (e-mail: Christos.Davatzikos@pennmedicine.upenn.edu; Fanyang.Yu@Pennmedicine.upenn.edu).

I. Introduction

FROM clinical documentation assistants to drug discovery platforms, LLM-based agents (a language model enhanced with tools or memory to autonomously reason and act on tasks) are rapidly branching into every corner of medicine. However, healthcare applications pose unique machine learning challenges, including heterogeneous multi-modal data integration, domain-specific knowledge representation, stringent requirements for safety and interpretability, and high stakes in clinical decision-making. This survey systematically examines how the agents address these challenges by integrating specialized biomedical knowledge, dynamic interaction with clinical databases, tools, and guidelines, and contextual reasoning over complex patient histories. Despite rapid progress, significant gaps remain in reliably translating these models into safe, trustworthy clinical tools. The recent surge in publications reflects a key trend: as shown in Figure 1, agent papers have now overtaken standalone LLM/VLM works, signaling a clear shift from foundational models to agentic system design.

Existing surveys on AI in medicine predominantly focus either on isolated LLM functionalities or general-purpose clinical NLP applications [1]–[6]. In contrast, this work provides a comprehensive review of autonomous and semi-autonomous LLM-based agents tailored specifically for integration within clinical workflows. Table I provides a summary of traits that differentiate us from the aforementioned survey works.

We introduce a novel taxonomy to categorize agents along three rigorously defined dimensions (application, tool-usage, and autonomy), as visualized in Figure 2.

This taxonomy was developed via an iterative coding process over the reviewed literature and validated through expert consultation, enabling systematic characterization and comparison of diverse agent architectures and functionalities. To summarize the entire landscape of agents in medicine, figure 3 shows the generalized architecture of medical LLM

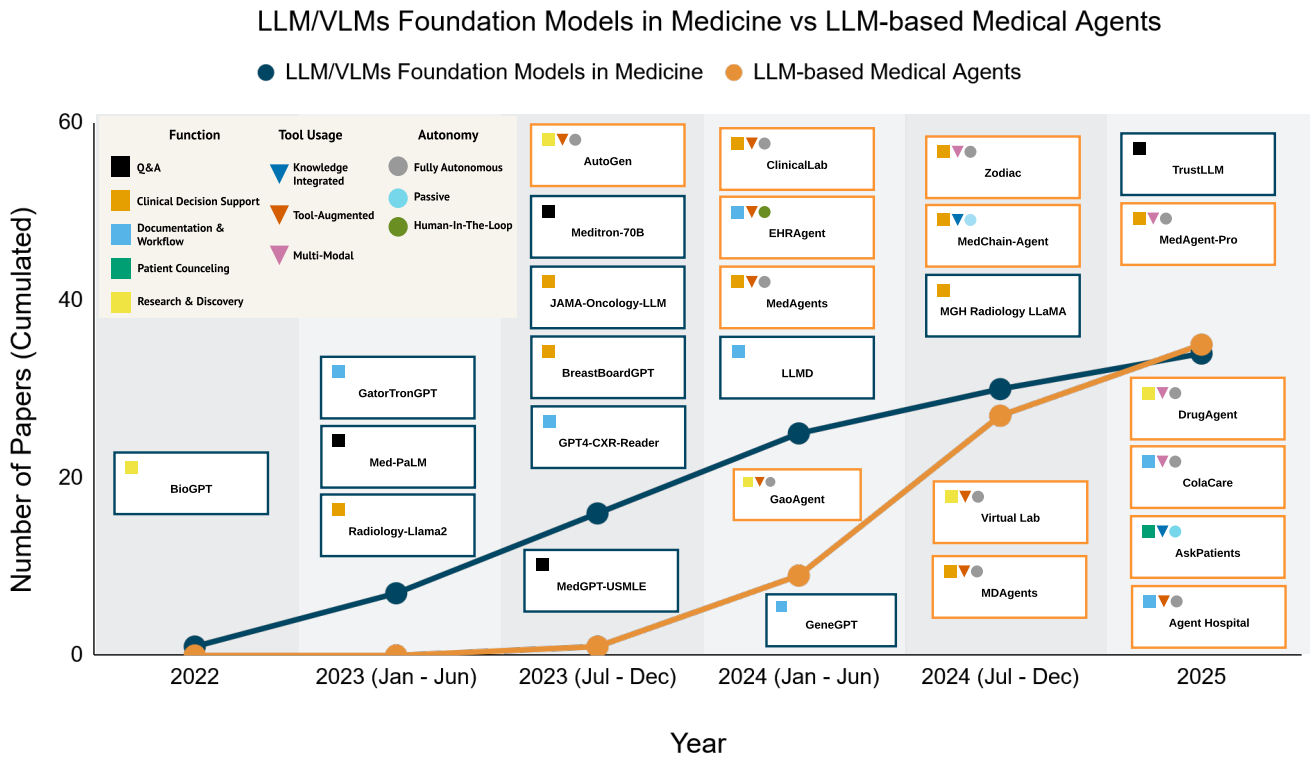

Fig. 1. Cumulative growth of LLM-based medical agents (orange) and standalone LLM/VLM models (blue) from 2022 to 2025. The x-axis is segmented into semiannual intervals to reflect the accelerating publication pace. Notably, by 2025, the cumulative count of agent papers surpasses that of LLM/VLMs, marking a transition from foundational model development to agentic system design. Each notable paper is annotated within its corresponding time interval; blue-bordered boxes indicate non-agent LLM/VLMs, while orange-bordered boxes indicate agentic systems. For each entry, square, triangle, and circle markers respectively denote the system's *functionality*, *tool usage*, and *autonomy*. LLM/VLMs are only labeled with functionality, whereas agent papers exhibit all three dimensions. Recurring combinations such as *Clinical Decision Support* + *Tool Augmented* + *Fully Autonomous* emerge as dominant agent archetypes.

agents, positioning the LLM as the central cognitive module that orchestrates task planning, external tool invocation like knowledge base queries, medical imaging analysis, and memory components for longitudinal patient data retention. This architecture generalizes prior agent designs by emphasizing modularity and the integration of heterogeneous clinical resources, enabling flexible deployment across diverse healthcare workflows.

We identify and analyze high-impact application domains, including radiology workflow, clinical documentation generation, and large-scale synthesis of biomedical literature for knowledge discovery. Concurrently, we critically assess technical limitations and risk factors, such as hallucination phenomena, dataset bias, and the challenge of verifying model outputs in high-stakes scenarios like cancer treatment planning. These concerns underscore the indispensable role of human oversight and robust evaluation frameworks in deploying LLM-based agents in clinical practice.

Drawing on advances in adjacent domains such as educational technology, robotics, and scientific discovery, we extract transferable methodological insights relevant to healthcare agents. For example, curriculum learning paradigms from educational agents inform adaptive medical training systems; hierarchical planning frameworks from robotics inspire structured clinical procedure modeling; and multi-agent collaboration techniques from scientific discovery platforms suggest novel architectures for complex multi-step clinical reasoning. These cross-domain analogies highlight promising avenues for enhancing agent robustness, flexibility, and interoperability.

By synthesizing interdisciplinary findings and establishing a rigorous taxonomy, this survey offers researchers, clinicians, and policymakers a comprehensive roadmap for designing, evaluating, and deploying LLM-based healthcare agents that augment, rather than replace, clinical expertise. The paper proceeds as follows: Section II defines foundational concepts and technical preliminaries; Section IV presents a systematic analysis of agent architectures and ML strategies; Section III surveys concrete clinical use cases; Section V discusses evaluation methodologies tailored to healthcare contexts; and Section VII outlines open challenges and future research directions essential for realizing trustworthy clinical AI agents.

II. Background of AI Agents

Agentic AI represents the next stage in the evolution of LLMs, extending them with sophisticated reasoning and iterative planning to autonomously address complex, multi-step problems. In healthcare, these capabilities underpin systems that can integrate diverse data sources, interact with clinical tools, and adapt to dynamic care environments. This section reviews the foundation models on which medical agents are built, provides an overview of agent architectures, and examines agents with specialized knowledge or tool integration.

TABLE I
COMPARISON OF OUR SURVEY WITH EXISTING REVIEWS ON LLMS IN HEALTHCARE (2024–2025). WE HIGHLIGHT DIFFERENCES IN AGENT FOCUS, TOOL INTEGRATION, MODALITY COVERAGE, TAXONOMY STRUCTURE, AND BREADTH OF APPLICATION DOMAINS.

Survey	Year	Modality	Agent Focus	Tool Usage	Taxonomy	Domains Covered	Notable Limitations
[1] (Ye et al.)	2025	Multimodal	No	No	Partial (Model-centric)	Imaging, VQA	Focused on MLLMs; lacks agentic framing
[2] (Aljohani et al.)	2025	Text	No	No	No	Trust, Risk, Ethics	Strong on safety, lacks technical taxonomy
[3] (Wang et al.)	2025	Text	Yes	Partial	Partial (High-level)	NLP, Diagnosis, Workflow	Lacks structured application taxonomy or system-level design
[4] (Khan et al.)	2025	Text	No	No	No	Foundation Models in Medicine	General overview; agents briefly discussed
[5] (Liu et al.)	2024	Text	No	No	No	LLM model types	Focused on categorizing LLMs, not on use as agents
This Survey (Ours)	2025	Multimodal	Yes	Yes	Yes (3D: Function, Autonomy, Tool Use)	Radiology, CDS, Docs, Patients, Research	First comprehensive agent-centric review with deep architectural and system analysis

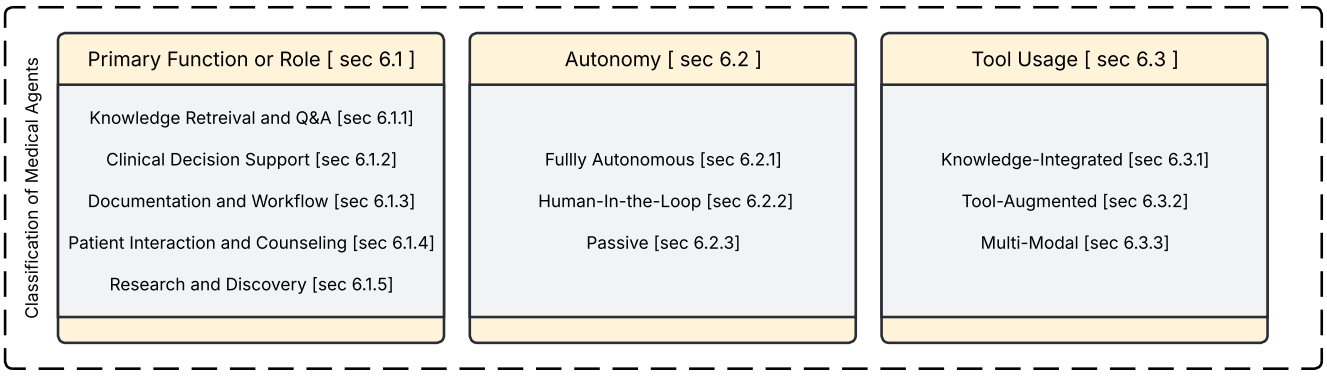

Fig. 2. Taxonomy of LLM-based medical agents organized by primary function, level of autonomy, and type of tool usage. This framework sets the stage for systematically categorizing agents based on what they do, how they operate, and the tools they use.

A. Agentic Foundation Models

Large language models are trained on vast text corpora to predict and generate text; through this process, they acquire a broad range of knowledge and linguistic competence. Key technical breakthroughs such as the Transformer architecture with self-attention, scaling laws for model size, and instruction tuning have enabled LLMs to achieve human-level performance on domain-specific tasks and model distillation to deploy LLMs in real-time scenarios. General models like OpenAI's GPT-3/4 [7] and Google's PaLM (540B) [8] demonstrate emergent abilities like zero-shot reasoning that were not apparent in smaller predecessors. In the medical domain, specialized LLMs have been introduced to encode biomedical knowledge more effectively. For example, BioGPT [9] is a generative Transformer pre-trained on 15M PubMed abstracts. It achieved state-of-the-art on biomedical QA and relation extraction tasks with a 78.2% accuracy on PubMedQA by 2022. Similarly, Google's Med-PaLM [10] was created by instruction-tuning a 540B model (Flan-PaLM) on medical Q&A. The resulting model's answers were judged to be aligned with clinical consensus in 92.6% of cases, approaching clinician performance (92.9%). These domain-specific LLMs provide a foundation for building reliable medical agents.

Recent advances have also extended LLMs into the multimodal domain. LLaVA-Med [11] adapts vision-language instruction tuning for medical contexts, enabling models to process clinical images alongside textual inputs. It demonstrates strong zero-shot performance on radiology and pathology benchmarks by aligning medical images with expert-level language supervision. Similarly, BiomedGPT [12] introduces an open-source generalist biomedical vision-language model trained under a multi-task contrastive objective across 15+ modalities and 25+ datasets. It outperformed GPT-4V and Med-PaLM on tasks including medical VQA, radiology report generation, and summarization, demonstrating strong zero-shot generalizability, even at a much smaller scale than closed-source counterparts.

MedGemma [13], built on the open-source Gemma architecture, combines large-scale biomedical vision-language pretraining with medical instruction tuning. It achieves state-of-the-art results across chest X-ray diagnosis, medical VQA, and multi-step agentic workflows. These medical multi-modal large language models (MLLMs) represent a new generation of foundation models capable of jointly reasoning over diverse

clinical inputs such as images, EHRs, and natural language.
Key Takeaways:

- Foundation models in medicine are large LLMs trained or adapted with biomedical knowledge to achieve expert-level reasoning in narrow clinical tasks.
- Recent medical MLLMs like LLaVA-Med, MedGemma, and BiomedGPT extend capabilities to process both text and clinical images.
- These models provide a strong knowledge base but lack planning, memory, or tool-use components on their own.

B. LLM Agents

An LLM agent extends a pre-trained language model by equipping it with the ability to perceive, reason, and act in different environments. The LLM serves as the cognitive core, while surrounding modules handle planning, memory, and tool integration. These augmentations allow the model to perform dynamic, task-oriented behaviors beyond static text generation [10], [14].

Planning mechanisms enable agents to decompose complex tasks (e.g., diagnostic reasoning or report generation) into intermediate steps. This is often achieved via prompting strategies such as chain-of-thought [15] or structured sequences like ReAct [16]. Memory components can be short-term (in-prompt) or long-term (external memory stores) [17], [18], allowing agents to preserve longitudinal patient context.

Tool use enables agents to interface with APIs, databases, and calculators [19], [20], bridging internal reasoning with external capabilities. Some systems also support multi-modal input, like radiology images or EHR tables [21], [22].

Multi-agent systems are an emerging direction. These frameworks assign specialized roles to collaborating LLM agents, such as planner, retriever, or verifier, to increase robustness and reliability [23]–[25]. For example, DrugAgent [24] divides biomedical discovery into modular sub-tasks handled by different agents. Clinical systems like MAGDA [26] and RareAgents [27] similarly employ role-based LLM collaborations to improve decision support and guideline adherence.

These extensions collectively enable LLM agents to interact with users and environments in a more intelligent, flexible, and context-aware manner, marking a fundamental shift from static model use toward dynamic agentic behavior.
Key Takeaways:

- LLM agents extend base models with planning, memory, and tool-use components, enabling dynamic interaction with clinical workflows.
- Planning and memory modules support multi-step reasoning and contextual continuity over time.
- Tool integration empowers agents to interface with external resources such as databases, APIs, or medical calculators.
- Multi-agent systems assign specialized roles like planner, retriever, verifier to improve performance, interpretability, and reliability.
- These architectures represent a shift from static model use to interactive, agentic intelligence tailored for healthcare tasks.

C. Tool-Augmented and Knowledge-Integrated Agents

Many clinical tasks require up-to-date or patient-specific information that falls outside a language model's pretraining corpus. Tool-augmented and knowledge-integrated agents address this limitation by invoking external resources such as calculators, APIs, clinical databases, or internet search engines [20], [28], [29].

One widely adopted approach is Retrieval-Augmented Generation (RAG), which retrieves relevant passages and incorporates them into the prompt for grounded output [30]. BioRAG [31] and Self-BioRAG [32] enhance biomedical factuality by combining domain-specific retrieval with iterative reasoning and self-reflection.

Beyond retrieval, many agents are equipped with functional tools. These include lab value calculators [19], [33], drug databases [24], and structured EHR interfaces [34], [35]. For example, ChemCrow [36] augments LLMs with chemistry tool APIs; similar architectures have been translated to medical domains for clinical recommendations [37].

Multi-agent systems also fall under this paradigm when each agent accesses tools independently or in a coordinated pipeline. Systems like MedAgent-Pro [21] and DrugAgent [24] use multiple agents to analyze multi-modal inputs and verify one another's outputs, increasing system robustness.

These designs improve factual accuracy, reduce hallucinations, and expand the functional capacity of LLM-based medical agents. However, they also increase system complexity, requiring reliable orchestration, error handling, and auditability for safe clinical deployment.
Key Takeaways:

- Tool-augmented agents access real-time or patient-specific data through APIs, databases, and calculators to enhance medical reasoning.
- Retrieval-Augmented Generation (RAG) frameworks improve factual accuracy by grounding model outputs in external evidence.
- Agents can use tools independently or as part of a multi-agent pipeline, increasing robustness and reducing hallucinations.
- Common tools include EHR access, clinical calculators, drug interaction APIs, and literature search engines.
- These systems demand careful orchestration and validation to ensure safety and reliability in clinical settings.

III. Applications of LLM-based Agents in Healthcare

In this chapter, we cover major application areas of agents and highlight the representative studies along with their use-cases, technologies used, evaluation, limitations, and potential open challenges for each area. We also chose an agent at the frontier of each application area and illustrated their workflow. Additionally, in table II we summarize the maturity level of each domain, as well as the depth of integration of technologies like tool-use, memory, multi-agent, and human-in-the-loop.

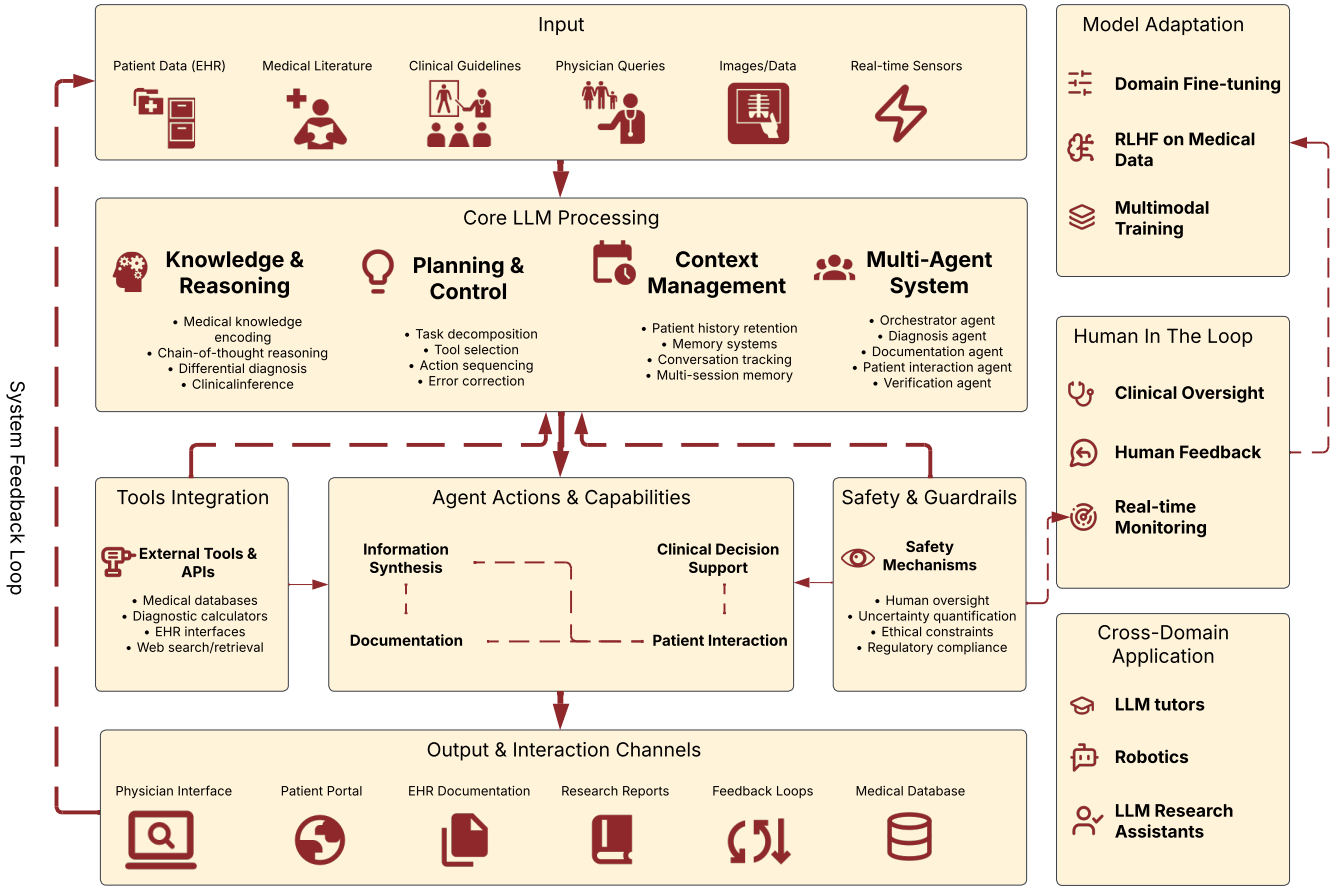

Fig. 3. This framework depicts the workflow of LLM-based agents in medicine. Starting with clinical inputs, the system processes information through adapted models with knowledge representation, memory, and planning capabilities. These enable key functions: information synthesis, clinical decision support, documentation, and patient interaction. External tools enhance capabilities while human oversight and safety guardrails ensure reliable operation. The framework supports various clinical, patient, and research applications with a continuous feedback loop for improvement.

A. Medical Imaging and Radiology

Radiology is a high-volume, high-interpretation workload domain with structured workflows, repetitive reporting tasks, and increasing multi-modal data complexity, making it especially conducive to agentic automation. The field already relies on Picture Archiving and Communication System (PACS), Radiology Information System (RIS), and templated documentation, offering a natural opportunity for integration of agents to act as intelligent intermediaries between human users and complex data systems. Figure 4 illustrates a multi-agent design for X-ray report summarization. Early studies confirm that LLMs can already interpret radiology text inputs and support decision-making at or near expert levels.

Radiology has seen some of the earliest explorations of LLMs in clinical workflows [38]–[41]. A prominent use case is automated patient triage and imaging protocol selection. Gertz et al. (2023) [42] demonstrated that GPT-4 could read free-text radiology request forms (clinical orders for imaging) and determine the appropriate imaging study and protocol 84% of the time, matching the decisions of radiologists in the majority of cases. This suggests LLMs can serve as front-end agents to streamline radiology workflows by prioritizing cases and recommending scan protocols based on clinical descriptions.

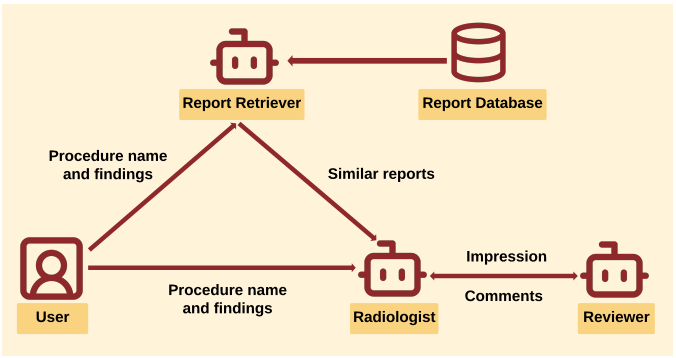

Fig. 4. Workflow of the *Rad-Council* system for chest X-ray summarization. The LLM agents collaborate over a shared memory and receive both visual features and textual findings to generate structured radiology impressions.

Another active application scenario for agents is radiology report generation and summarization. Radiologists are required to interpret medical images and manually draft reports, which typically contain structured sections such as findings and impressions. Recent studies have evaluated the ability of LLMs

TABLE II
COMPARISON OF AGENTIC PROPERTIES EXHIBITED BY AGENTS ACROSS MAJOR HEALTHCARE APPLICATION DOMAINS. THE TABLE EVALUATES FIVE KEY DIMENSIONS: TOOL USE, MEMORY HANDLING, MULTI-AGENT COLLABORATION, HUMAN-IN-THE-LOOP SUPERVISION, AND LEVEL OF AUTONOMY. THESE DIMENSIONS REFLECT THE DEGREE TO WHICH AGENTS OPERATE INDEPENDENTLY, USE EXTERNAL RESOURCES, AND INTERACT WITH HUMANS OR OTHER AGENTS. A GREATER NUMBER OF CHECKMARKS (✓) INDICATES STRONGER INTEGRATION OF THAT PROPERTY. THIS COMPARISON ILLUSTRATES HOW APPLICATION CONTEXTS SHAPE THE DESIGN AND CAPABILITIES OF LLM-BASED HEALTHCARE AGENTS.

Domain	Tool Use	Memory	Multi-Agent	Human-in-the-Loop	Autonomy	Maturity Level
Medical Imaging & Radiology	✓	.	✓	✓	**Medium**	Early Clinical
Clinical Decision Support	.	✓	✓	✓✓	**Low**	Experimental
Clinical Documentation	✓✓	✓	.	✓	Medium	Pilots Running
Patient Interaction	✓	.	.	✓✓	Low	Emerging
Drug Discovery & Research	✓✓✓	✓	✓✓	.	High	Preclinical Use

to act as a reporting assistant, like generating a concise impression summary given the detailed findings. Sun et al. (2023) [43] found GPT-4 could produce plausible impression summaries from chest X-ray findings. In a blinded evaluation, however, radiologist-written impressions still outperformed GPT-4's on coherence, completeness, and factual correctness [44]. Other studies noted that while LLMs sometimes overstated confidence or introduced minor unsupported statements [45], [46], their outputs were often rated highly coherent and acceptable by referring physicians [47]–[49]

Agents also have the potential to serve as diagnostic assistants in medical imaging. [50] explored GPT-4 for the differential diagnosis generation task, which is to give a description of imaging findings, the model needs to suggest likely diagnoses. Impressively, in 94% of cases, the LLM's differential diagnoses were deemed acceptable. The results indicate that LLMs, even without direct vision input, can leverage text descriptions of images combined with clinical context to propose diagnoses. Such an agent might take radiology findings ("MRI shows multiple demyelinating plaques...") and list possible conditions, like multiple sclerosis, neuromyelitis optica, along with recommended next steps [51]. Wang et al. (2025) [21] proposed a multi-agent framework for multi-modal medical diagnosis via an evidence-based reasoning approach. The diagnostic workflow consists of both task and case levels, where a diagnostic plan is first generated by knowledge-based reasoning from RAG and planner agent, then multiple tool agents continue to process multi-modal patient data and output a final diagnosis based on the qualitative and quantitative evidence. These agents can aid radiologists as second readers or training tools and could further facilitate the clinical diagnosis procedure and improve treatment planning.

Clinical interaction and education in radiology have also benefited from LLMs. ChatGPT, despite not being trained specifically on radiology, was able to answer radiology-related questions, like explaining an MRI finding or a procedure at a level useful for radiologists, trainees, and even patients [52]. [53] reported that GPT-4, with its more advanced reasoning. One study noted that ChatGPT nearly passed a multiple-choice radiology board exam (when image-based questions were excluded) [54], underscoring the model's strong knowledge base. This opens up possibilities of using LLM agents as virtual tutors for radiology residents, simulating exam Q&A or explaining difficult cases in plain language [51].

Integration of LLM agents with radiology IT systems is under exploration. Researchers envision LLM-based agents serving as a natural language interface to PACS and hospital records [55]. This type of agent could also auto-generate follow-up recommendations. These applications involve tool use: the LLM must query databases and possibly invoke search or filtering functions. Early prototypes indicate this is feasible, but significant engineering is needed to ensure reliability and privacy.

In conclusion, radiology provides a fertile ground where LLM-based agents can assist with workflow optimization, report generation, diagnostic reasoning, Q&A, and data retrieval. The agentic AI tools will significantly augment radiologists' capabilities when properly integrated into the clinical routine.

B. Clinical Decision Support and Treatment Planning

Beyond imaging, LLM agents are being studied as clinical decision support tools for diagnosis and treatment planning in complex cases. A striking example of that is MedAgent-pro, which uses multi-modal patient data for disease diagnosis. Figure 5 illustrates the workflow of MedAgent-Pro [21].

One high-profile study by [56] evaluated whether current conversational LLMs could assist oncology tumor boards in precision oncology (personalized cancer treatment). In this study, four LLMs (including ChatGPT and a biomedical model) were prompted with the molecular profile of 10 fictional advanced cancer patients to generate personalized treatment options. The results were sobering: the LLMs' recommended treatments often deviated substantially from expert oncologist recommendations, and many suggestions were recognized by physicians as AI-generated and not trustworthy.

On the positive side, the LLMs did correctly identify several important treatment strategies and even proposed a few novel options that experts had not considered. This indicates that while LLM agents are not yet ready to autonomously plan cancer therapy, they could act as a brainstorming aid, generating alternative ideas for the medical team to review. The authors conclude that current LLMs should not be used in routine clinical decision-making for oncology, but their rapid improvement suggests potential as support tools in the near future. Crucially, such agents would need strict human oversight to vet each suggestion.

Another line of research has tested whether an LLM assistant can improve physician diagnostic reasoning in general

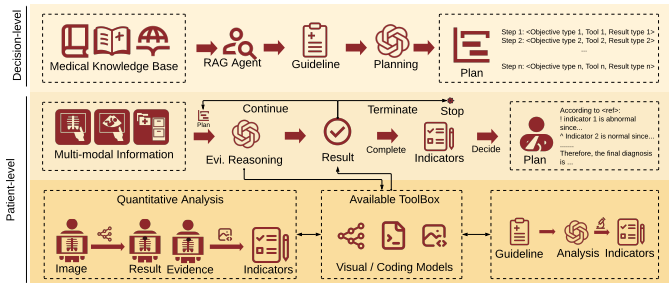

Fig. 5. Workflow of the *MedAgent-Pro* system for clinical decision support using multi-modal patient data. The pipeline begins with patient information such as imaging, lab results, and clinical notes. A planner agent generates a diagnostic reasoning plan, followed by tool-augmented LLM agents that retrieve evidence, analyze inputs, and verify intermediate hypotheses. The final diagnosis and recommendations are synthesized based on accumulated evidence and structured reasoning.

practice. In a randomized trial (2023) with 50 physicians, participants either used their standard resources or had access to ChatGPT (GPT-4 based) as an AI diagnostic aid while solving clinical cases [57]. The outcome was perhaps unexpected: having the LLM "assistant" did not significantly improve the physicians' diagnostic accuracy or confidence. This trial highlights an important limitation. If not carefully integrated, an LLM agent can introduce distraction or false reassurance, yielding no net benefit. It underscores that simply providing an AI tool doesn't guarantee improved outcomes; how clinicians interact with and trust the agent is critical. Future decision-support agents might require refined UX design or training for users to harness the AI effectively.

Despite these challenges, there are niche decision-support domains where LLMs have excelled. One example is a clinical guideline and literature analysis. [58] showed that a medically-tuned LLM (Med-PaLM 2) could answer complex medical questions by synthesizing information from clinical guidelines at an expert level. Another study found GPT-4's answers about specialized clinical topics like rhinology guidelines were surprisingly complete and accurate, in some cases on par with specialists [59].

The difference from the oncology case may lie in problem structure: guideline-based queries or board-style questions have a clearer correct answer, whereas open-ended cancer therapy planning is far more complex with many acceptable approaches. In practice, one could imagine an LLM agent integrated into an electronic health record that, given a specific patient profile, retrieves relevant practice guidelines or similar cases and presents a few evidence-based options. Early prototypes of this idea have shown promise in fields like gastroenterology and cardiology [60], where LLMs could assist with diagnostic scopes and suggest workups in challenging cases.

Overall, clinical decision support agents remain experimental. Their current value lies in augmenting retrieval guidelines, summarizing similar cases, and suggesting evidence-based options, while final decisions must remain human-led. Human-in-the-loop designs are essential to ensure safety and reliability.

C. Clinical Documentation and Narrative Generation

Clinical care involves enormous amounts of documentation: writing progress notes, discharge summaries, referral letters, etc. Agents are being developed to lighten this load through automatic summarization and document generation. Recent work indicates that adapted LLMs can even outperform human experts in certain medical summarization tasks.

For example, Shi et al. [61] introduced *EHRAgent*, a clinical agent designed to perform complex reasoning tasks over structured EHR data using large language models. EHRAgent can answer complex clinical questions such as "What medications have been discontinued in the past 48 hours?" with high accuracy and transparency. Figure6 shows the workflow of the system, demonstrating how it connects natural language prompts to precise EHR data operations.

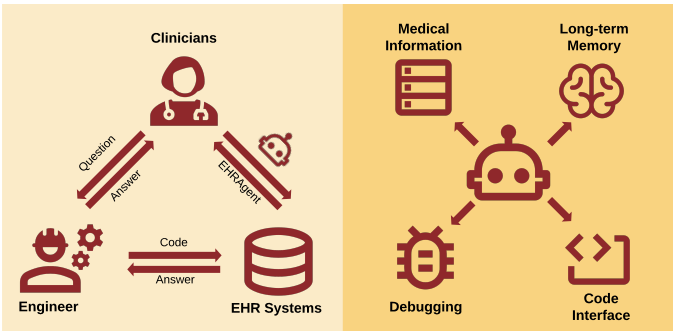

Fig. 6. Workflow of the *EHRAgent* system for clinical decision support and documentation over structured electronic health records (EHRs). Given a natural language query from a clinician, the system decomposes the request into executable operations on structured patient data (e.g., labs, medications, vitals), using LLM-guided code generation and interpretation. The agent then performs reasoning over the extracted results to produce a grounded response or summary. Adapted from [61].

This result was the first evidence of LLMs exceeding human performance in clinical note summarization, which suggests huge potential for reducing documentation burden. Documentation can be subdivided into specific application areas, which include:

- Discharge Summaries: These are concise documents given to patients at hospital discharge, summarizing their hospitalization and next steps [62].
- Patient Queries to Clinical Notes: Patients often send messages or have questions that require reading through their charts. LLM agents can summarize patient histories or multi-visit records to answer specific questions [63].
- Transcription to Documentation: Startups and studies have used LLMs to convert raw transcripts of doctor-patient conversations into structured clinical notes. This involves not just summarizing but also organizing content by sections.

In these applications, the prompting strategy plays a big role. For summarization, techniques like few-shot prompting (providing examples of good summaries) or instruction tuning greatly improve output quality [64]. LLM agents might also use iterative refinement: generate a summary, then potentially refine it upon user request. Some systems incorporate a validation step where another model or a rule-based checker

verifies that no key facts are missing. This approach addresses the risk of omissions or hallucinations.

One remarkable aspect reported by van Veen et al. [65] is that standard NLP metrics for summarization did not always correlate with physician preferences. This underscores the importance of human evaluation in medical NLP: an LLM agent's success should be measured by clinician end-users' satisfaction and safety outcomes, not just BLEU or ROUGE scores. The study's findings that LLM summaries were often preferable to human ones in terms of completeness and correctness suggest that, at least for factual summarization tasks, these agents are nearing practical usability.

Nonetheless, caution is warranted. Hallucination, when the AI makes up information, is dangerous in medical notes. For example, an agent might fabricate a lab result or mention a symptom that was not actually documented. To mitigate this, some research uses retrieval-augmented generation (RAG): the agent first retrieves relevant snippets from the patient's records and is forced to base the summary on those snippets [66]. This can improve factual accuracy by grounding the generation in actual record text. Additionally, human clinicians will need to review AI-generated documents for the foreseeable future. Even if an agent correctly writes 90% of a discharge summary, the clinician must verify its content. Over time, if trust is built and error rates are exceedingly low, more autonomy could be given.

In summary, clinical documentation stands out as a domain where LLM-based agents have already shown tangible benefits: reducing tedious writing tasks and possibly improving the consistency of records. With careful design, these agents could give clinicians back valuable time and reduce burnout associated with paperwork. The key will be ensuring factual reliability and integrating these tools smoothly into clinical workflows, like embedded in the EHR interface.

D. Patient Interaction and Conversational Agents

Interacting with patients via natural language is another promising area for LLM-based agents. Applications range from symptom checkers and health chatbots to patient education and counseling. Compared to clinician-facing tools, patient-facing agents demand extra caution around trust, empathy, and safety. Still, recent peer-reviewed studies have begun examining LLM performance in answering patient questions, often focusing on specific domains.

In one study, [67] asked whether ChatGPT and similar LLMs (with internet access) could answer the questions of prostate cancer patients accurately and help democratize medical knowledge. They connected ChatGPT to an online medical database and evaluated its responses to common patient concerns about prostate cancer. The LLM was able to retrieve information and provide answers covering topics like treatment options, side effects, and prognosis. While the study found the answers were often comprehensive, it also noted variability in quality among different LLMs and occasional outdated or irrelevant information. Following up on that, Zhu et al. [68] proposed *APP*, a patient-facing conversational agent designed to support human-centric medical dialogue through grounded reasoning. The agent engages in multi-turn conversations with patients, incrementally refining its understanding of symptoms and diagnostic hypotheses. By integrating retrieval-augmented generation (RAG), the system retrieves relevant clinical knowledge mid-dialogue to reduce hallucinations and support evidence-based interaction. APP also incorporates a self-reflection loop, enabling the agent to revise or verify its responses based on newly retrieved information. Figure 7 illustrates the agent's reasoning-driven workflow.

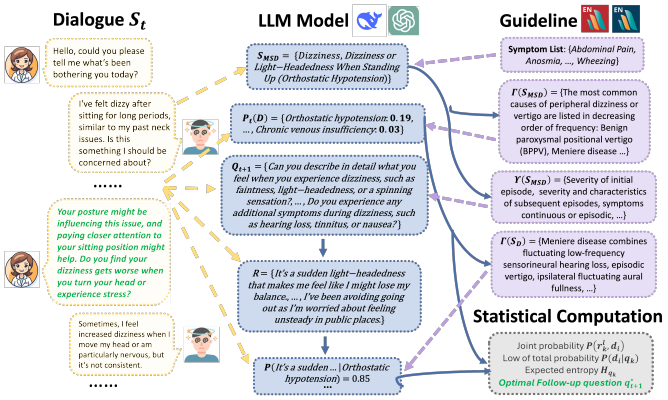

Fig. 7. Workflow of the *APP* agent system for interactive patient counseling and diagnosis refinement. The agent receives a patient's natural language input, extracts symptoms, and queries external medical knowledge sources using retrieval-augmented generation (RAG). It then reasons over the retrieved content to form a diagnosis, engages in follow-up dialogue to clarify uncertainties, and iteratively verifies its conclusions using a self-reflection loop. Adapted from [68].

[69] tested ChatGPT on questions about liver cirrhosis and hepatocellular carcinoma (liver cancer), comparing its answers to those of physicians. Physicians rated the AI's answers on accuracy and completeness. They found ChatGPT could correctly answer a significant portion of questions, often providing layperson-friendly explanations. Nonetheless, the study identified limitations in the model's ability to respond to nuanced or case-specific questions. While current large language models (LLMs) performed well in answering general patient FAQs, such as those related to symptoms or standard treatment options, their effectiveness diminished when addressing personalized or context-dependent queries.

An intriguing use case for patient-facing LLM agents is as a support tool in multidisciplinary care meetings. [70] explored using ChatGPT during a breast cancer tumor board (a meeting where doctors discuss cases). The agent was prompted with details of a breast cancer case and asked to provide insights or treatment suggestions, effectively acting as a non-voting "advisor" in the meeting. While clinicians did not rely on its suggestions to make decisions, they noted it occasionally brought up pertinent points. More importantly, the study assessed how the presence of an AI agent influenced the discussion. Doctors reported that when the AI concurred with their planned treatment, it gave some additional confidence, and when it differed, it prompted them to double-check their rationale, though they would not follow the AI blindly. This scenario exemplifies a human-AI collaborative workflow, where the AI agent is neither fully autonomous nor

just a passive tool, but an active participant that can shape human decision-making dynamics (for better or worse). The authors caution that appropriate safeguards and transparency are needed. The AI's suggestions must be taken as hypotheses, not facts, and any errors it makes should be identified to avoid misinformation in the meeting.

More traditional conversational agents for patients (sometimes called "digital health assistants") are also being enhanced by LLM technology. For mental health counseling, for instance, earlier rule-based chatbots are now being replaced by GPT-based agents that can converse more naturally about stress, anxiety, or lifestyle, while following therapeutic frameworks.

In summary, LLM-based agents show promise in patient interaction settings like health education (explaining conditions, treatments), triage and symptom checking (guiding patients on whether to seek care), and support in care coordination (like summarizing patient questions for doctors, or even interacting in clinical visits as a "AI scribe" that also answers patient queries). Key challenges include alignment (making sure the agent's tone and content are appropriate for patients), factual accuracy with the latest medical knowledge, and clear disclaimers that these agents are not a replacement for professional medical advice. Based on current research, a reasonable near-term scenario is using LLM agents to augment communication. For example, a patient asks an agent a question via a clinic's portal, the agent drafts an answer with references, and a clinician reviews it before it's sent out. This semi-automated workflow could greatly increase efficiency in handling the large volume of patient messages many practices receive, while keeping a human clinician in the loop for quality control.

E. Drug Discovery and Biomedical Research

Outside of direct clinical care, LLM-based agents are being applied to biomedical research problems, including drug discovery, genomics, and literature mining. These applications often involve coupling LLMs with scientific databases and cheminformatics tools, effectively creating research assistant agents. While some work remains in pre-print form, peer-reviewed studies are beginning to appear, especially demonstrating how LLM agents can navigate the vast biomedical literature.

A primary use of LLMs in drug discovery is knowledge extraction and hypothesis generation from texts. The volume of publications in biology and chemistry is enormous; LLMs fine-tuned on this literature can act as mining tools. BioGPT, mentioned earlier, was shown to excel at extracting relationships like drug–disease associations from text [9]. Another model, GatorTronGPT [71], a 2023 generative LLM trained on a mix of clinical notes and biomedical texts (277 billion words), demonstrated the ability to both answer clinical research questions and generate new hypotheses by integrating information across papers. Such models can serve as agents that a scientist queries in natural language to get synthesized answers drawing from many studies.

Beyond text, LLMs are being integrated with chemical structure data. Although molecules are not text, one can represent chemical structures as strings (SMILES notation) and train language models on them. For instance, Wang et al. introduced cMolGPT [72], a conditional GPT that generates molecules with desired properties by conditioning on protein targets. By incorporating protein–ligand interaction knowledge, cMolGPT could propose new compounds likely to bind a given target (like EGFR, an important cancer target). This LLM can be seen as a drug discovery agent that "ideates" new molecules when asked, say, "Design a molecule that inhibits protein X but is similar in structure to known drug Y." It uses its learned chemical language to output candidate structures, which are then evaluated by predictive models or experimental tests. Early results show these AI-generated molecules can indeed exhibit high predicted activity, like cMolGPT achieved >0.75 correlation in activity prediction.

An exciting development is multi-agent systems for automated experimentation. DrugAgent [73], a framework in which multiple LLM-based agents collaborate to design and execute machine learning workflows for drug discovery. The workflow is illustrated in Figure 8 DrugAgent assigns different roles to agents (a planner, a coder, an analyst, etc.) to automate tasks like data preprocessing, model training on bioassay data, and result interpretation.

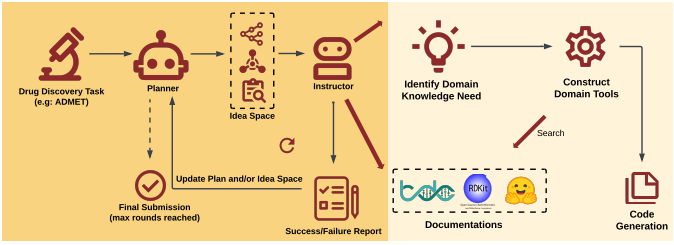

Fig. 8. Overview of the DrugAgent multi-agent workflow. Given a drug discovery task described in natural language, the Planner proposes solution ideas, and the Instructor incorporates domain-specific knowledge to generate working code pipelines. Each candidate is executed, evaluated, and the best-performing solution is identified and returned.

In case studies, this multi-agent system outperformed single-agent approaches (like a lone LLM with ReAct prompting) by about 5% in ROC-AUC on drug-target interaction prediction. The improvement likely comes from the specialization and interplay between agents, mimicking a team of experts, which reduces errors in complex pipelines. For example, one agent might be tasked with calling a chemistry database API (tool use) to fetch compound data, another with writing Python code to train a QSAR model, and another with analyzing the outputs, all coordinated by a top-level planner agent.

This kind of autonomous "AI scientist" setup is still experimental, but it foreshadows how LLM agents could eventually handle routine research tasks. Notably, a system called AI-Scientist [74], [75] has also been proposed to design and analyze experiments end-to-end. While not specific to drug discovery, it could be adapted to, say, plan a series of virtual screening steps or to optimize a synthetic route for a drug molecule.

Tool integration is crucial in these agents. Chemistry-focused agents often integrate specialized toolkits: for exam-

ple, ChemCrow [76] is an LLM agent augmented with 17 chemistry tools for tasks like drawing structures, querying spectra, looking up synthesis routes.

In literature analysis, multi-modal LLMs are emerging that can handle text, tables, and even figures from papers. An agent that can read a PDF of a clinical trial, extract key results, and present them in summary form would be invaluable to researchers and clinicians alike. Some prototypes (unpublished) are exploring connecting LLMs with document parsing tools for this purpose.

In summary, LLM-based agents in drug discovery and biomedicine act as amplifiers of scientific discovery, sifting through knowledge and sometimes generating new hypotheses or molecule designs. They tend to be tool-heavy (given the need for data retrieval and chemical computations) and often operate as part of an automated pipeline rather than an interactive conversation. While not a replacement for human scientists, they can offload grunt work and suggest non-obvious connections. Encouraging results in this realm suggest that the coming years may see LLM agents routinely assisting in tasks like systematic literature reviews, hypothesis generation for grant proposals, or even running virtual experiments in silico before any wet-lab work is done.

IV. Layered Evolution of Techniques and Architectures for LLM-based Medical Agents

The evolution of medical agents displays a clear progression in sophistication. Early efforts began with simple prompt engineering, which treated the model as a static knowledge engine, which is insufficient to tackle the complexity of downstream medical tasks. Newer approaches layered in tool use for grounded reasoning, memory systems for persistent context, and multi-agent architectures for complex task orchestration. At each stage, new capabilities were unlocked at the cost of added system complexity. We now walk through this layered development, from the foundational to the frontier. The overall pros and cons of each emerging enhancement is summarized in table III. We also provide figures explaining each techniques within representative medical agents.

A. Prompting Strategies and Few-Shot Learning

Prompt engineering, the process of crafting the right input to guide the LLM, is often the first step in turning an LLM into a task-specific agent. Many early successes in medical LLM applications came from clever prompting alone, without additional training. For example, simply instructing the model with an explicit role can significantly improve the relevance of its output [78]–[80].

When evaluating GPT-3.5 on medical exam questions, Kung et al. [81] provided exemplars of question-answer pairs in the prompt; this few-shot prompting enabled the model to score around 60% on USMLE-style questions. With GPT-4, which has more emergent reasoning ability, even zero-shot [82] performance on exams jumped near or above the passing threshold. These results reinforced what Brown et al. [83] showed broadly: large LMs are few-shot learners that can infer the format and goal of new tasks from a handful of demonstrations.

A powerful prompting approach for reasoning tasks is chain-of-thought (CoT) prompting [15]. Instead of asking the LLM to directly output an answer, the prompt encourages it to generate a step-by-step explanation first (the ”thoughts”) and then the final answer. In medical contexts, CoT can be very useful, like in a differential diagnosis agent, having the model list its reasoning before concluding can both improve accuracy and provide transparency.

Indeed, researchers have found that CoT prompting elicits more accurate diagnoses from LLMs, as it mitigates some of the ”jump to conclusion” errors. Yao et al. [77] took this further with Tree-of-Thoughts, where the model explores multiple reasoning paths (like different possible diagnoses) in a tree search before finalizing an answer. Figure 9 illustrates the prompting strategies as well as tree-of-thought. Such techniques could be particularly relevant for complex medical decision-making where a linear thinking process (CoT) might miss alternatives that a tree search could catch, like considering rare diagnoses.

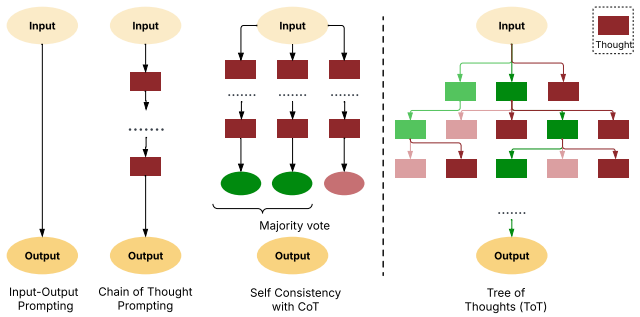

Fig. 9. Overview of prompting strategies including chain-of-thought and tree-of-thought.

Prompt order and phrasing can also matter. Lu et al. [84] showed that LLM few-shot prompting is sensitive to the order of examples. To get reliable outputs, one might need to try different example orders or use techniques like ordering ensembles.

Another recent idea is self-refinement [85] prompts: after an initial answer, the agent can be prompted with ”Check your answer for errors or missing info.” This mimics a human double-checking their work. In the oncology LLM study, the researchers allowed ChatGPT to regenerate answers after being told its first attempt might have missed something; this sometimes led to more comprehensive treatment options, as the model would then mention additional therapies [86].

Finally, prompt-based conditioning has been used to control style and safety. For patient-facing agents, prompts often include explicit instructions like ”If you do not know the answer or if it would be unsafe to answer, respond with a disclaimer.

In summary, prompting is a powerful lever to adapt general LLMs to specific medical tasks. It is often the most accessible method (not requiring model retraining). However, prompts can be brittle and may not guarantee consistency. Thus, many

TABLE III
COMPARISON OF CORE ARCHITECTURAL TECHNIQUES USED IN LLM-BASED MEDICAL AGENTS ACROSS FIVE SYSTEM DIMENSIONS. WE SUMMARIZE KEY STRENGTHS, LIMITATIONS, TYPICAL USE CASES, AND REPRESENTATIVE SYSTEMS FOR EACH APPROACH. THIS TABLE COMPLEMENTS THE ARCHITECTURAL DIAGRAMS IN FIG. 3 AND OFFERS PRACTICAL GUIDANCE FOR RESEARCHERS AND PRACTITIONERS.

Technique	**Key Strengths**	**Common Limitations**	**Best Use Cases**	**Representative Systems**
Prompting / CoT	Simple, training-free; elicits reasoning; interpretable	Brittle to prompt changes; limited factual grounding; no memory or tools	QA, summarization, basic decision support	GPT-4 (CoT), Medprompt, Tree-of-Thoughts [77]
Tool-Augmentation, like RAG	Improves factuality and recency; supports API, search, or calculator use	Integration overhead; tool invocation errors; maintenance burden	Evidence-based reasoning, score calculation, guideline retrieval	BioRAG [31], Almanac [29], ChemCrow [36]
Memory-Augmentation	Supports longitudinal tasks; retains interaction history	Memory update/decay management; privacy risks	Chronic care agents, follow-up note summarization	Self-BioRAG [32], EHRNoteQA [34]
Multi-Agent Systems	Modular, composable; role separation improves specialization	Complex orchestration; communication overhead; error propagation	Research pipelines, multimodal workflows, planning-heavy tasks	DrugAgent [24], MedAgent-Pro [21]
Human-in-the-Loop	Enhances trust and safety; ideal for high-stakes settings	Slower throughput; requires clinician engagement	Clinical decision support, note validation, oncology planning	Benary et al. [56], APP [68], Radiology LLM studies [51]

efforts combine prompting with more robust frameworks like fine-tuning or tool use for reliability.

B. Tool-Augmented and Knowledge-Integrated Agents

Many medical tasks require information not contained in the LLM's trained parameters. For example, up-to-date research findings, a patient's personal health records, or performing a calculation on data. Tool-augmented LLM agents address this by connecting the model to external resources.

One common form is Retrieval-Augmented Generation (RAG) [87]. In RAG, when the LLM gets a query, the system first retrieves relevant documents, like PubMed abstracts, or sections of a textbook or clinical guidelines and prepends them to the model's input. The LLM's answer is thereby grounded in those references. This approach has been successfully used in medical QA benchmarks [88].

For instance, Med-PaLM [10] employed a variant of RAG when answering medical exam questions, retrieving medical knowledge snippets so that the model's generation could quote and build on verified information. RAG greatly reduces hallucination and improves factual accuracy, at the cost of needing a database and a retriever component.

Beyond retrieval, LLM agents can be given calculators, databases, and APIs. A clear example is the ChemCrow [89] agent in chemistry, which had tools like a calculator for molecular weight and a database for compound properties. In a healthcare context, analogous tools could be: a drug database API (to check dosages, interactions), a lab interpretation tool (to compute MELD score from lab values, for instance), or simply a general calculator for medical formulas, like body mass index, chemotherapy dose by BSA. An agent with a calculator tool would not make the arithmetic mistakes that a pure LLM might. Schick et al. [20] notably created Toolformer, showing that LLMs can even learn when to call tools by themselves.

Search and Internet access have been tested as well. For rapidly evolving fields like COVID-19, any model trained on data before 2020 is obsolete. But an agent with internet search capability can fetch the latest guidelines from the web. Several studies, like [67], found that connecting ChatGPT to the internet improved its ability to answer patient questions correctly. However, using internet tools introduces new risks: the agent might encounter unreliable sources. Thus, careful curation or site-specific search is often implemented.

A special type of tool is the Electronic Health Record (EHR) interface. Experimental agents have been built that, given natural language instructions, perform operations in the EHR like "pull the latest MRI report" or "summarize the medications list". This essentially treats the EHR as a database that the agent can query via an API. If realized, this could allow clinicians to simply ask the agent for information from the chart (lab values, past visit notes), instead of manually clicking through multiple screens. Microsoft's internal Copilot for EHR projects and others are exploring this [29], but peer-reviewed outcomes are not yet published.

A lightweight but effective example of a medical agent that is both tool-augmented and knowledge-integrated is BioRAG [31]. This system enhances biomedical question answering by equipping an LLM with a domain-specific retrieval module that interfaces with a curated corpus of medical literature. Figure 10 illustrates the workflow of BioRAG.

In summary, tool augmentation greatly expands what medical LLM agents can do, moving from purely knowledge-based tasks to action-based tasks (querying, calculating, updating records, etc.). Most literature supports the combination of LLM + tools as the path to make agents useful in real clinical environments, since no static model can contain all up-to-date medical knowledge or perform all functions. The trade-off is added complexity: these agents have more moving parts and potential points of failure (tool API errors, retrieval mistakes, etc.). Nonetheless, early successes in both general domains and biomedical ones point to tool use as a critical feature of advanced LLM-based healthcare agents.

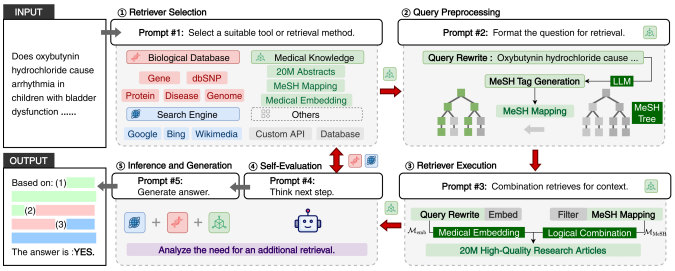

Fig. 10. The architecture of the proposed BioRAG framework consists of five iterative components that collectively enhance biological question reasoning. The system begins by selecting the most appropriate information source through a retriever selection module. The input query is then pre-processed to improve retrieval quality, including rewriting and assigning topic tags from a predefined knowledge hierarchy. Retrieved context is gathered from a biomedical knowledge base using a hybrid retrieval strategy. A self-evaluation step assesses the adequacy of the retrieved information and determines whether further retrieval is necessary. Finally, the LLM performs inference and generates a response grounded in the retrieved evidence. Adapted from [31]

C. Memory and Long-Term Context

Memory architectures addresses the context window limitation of LLMs. In clinical scenarios, agents need to remember information with large context, which these memory architectures facilitate.

One approach is to use an external vector database to store and retrieve past interactions. For example, an agent in a chronic disease management app may have dozens of chat sessions with a patient over months. Key facts from each session can be embedded and stored. When a new session begins, the agent retrieves the most relevant past facts to "remind" itself [18], [88]. This is similar to retrieval augmentation, but specifically tailored to conversation history.

Researchers have also experimented with summarizing interactions on the fly. After each encounter, the agent could generate a succinct summary and store it. Next time, it only needs to read the summary. This hierarchical memory can keep the context manageable.

In non-medical domains, techniques like long-term memory modules (key-value memory networks, etc.) have been integrated with LLM agents [17]. In healthcare, this is still nascent, but one can envision each patient as having a dedicated memory store that the agent accesses. Privacy is a concern here: how and where those memories (which may contain PHI) are stored needs to be handled in compliance with regulations (likely on secure servers, not the public cloud).

A strong example of a memory-augmented medical agent is the personalized LLM assistant proposed by Zhang et al. [90], which coordinates both short-term and long-term memory to support patient-specific dialogue. Figure 11 shows the overview of this memory integration into agentic systems. The short-term memory module captures recent conversational context, while the long-term memory module stores and retrieves key health events, previous questions, and agent responses across interactions. This allows the agent to personalize recommendations and avoid redundant prompts, closely mimicking human clinicians who build rapport and reasoning over time. The system demonstrates how layered memory coordination can improve both user experience and clinical task performance, especially in domains like chronic care, symptom tracking, and follow-up management.

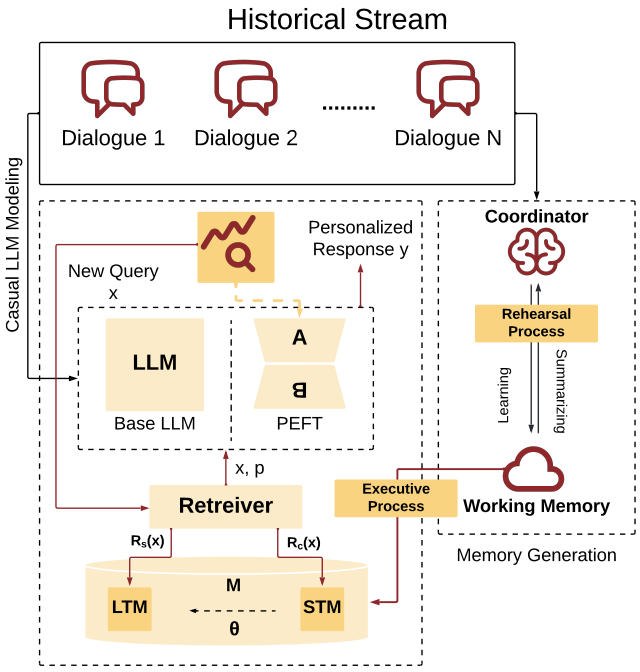

Fig. 11. Overview of the MaLP framework. The user's past conversations are processed by a coordinator and a fine-tuned LLM to build memory using a dual-process mechanism (DPeM). Once memory is constructed, new user queries are matched with relevant past information through a retriever. The LLM then generates personalized responses based on both retrieved memory and previous dialogue history.

In summary, while not heavily covered in early 2018-2023 healthcare LLM papers, memory augmentation is likely to become important as agents move from one-off tasks to continuous roles (like a long-term health coach or a primary care assistant that "knows" the patient's story). Techniques borrowed from general AI are vector stores, summarization, and hierarchical memory. These techniques will need to be validated in the medical context.

D. Multi-agent Architectures

Some systems use more than one language model agent to divide the work. These multi-agent systems are designed to handle complex clinical tasks by having different agents take on different roles. Each agent can specialize in a part of the workflow, such as planning, retrieving information, verifying decisions, or generating text.

One common design is a planner agent that breaks down the task into smaller steps, and then delegates those steps to worker agents. For example, in a diagnostic agent, one component might collect patient history, another might suggest possible conditions, and a third might verify whether the suggestion matches clinical guidelines. This kind of role separation mirrors how real clinical teams work, with different specialists contributing to the final decision. It also makes the system easier to scale or modify, since each agent can be replaced or retrained independently.

A good example is RareAgents [27], a framework tailored to rare disease diagnosis and treatment. It simulates a multi-disciplinary medical team: a central "attending physician" agent selects relevant specialist agents from a predefined pool. Those specialists discuss iteratively to reach a consensus on differential diagnosis and medication plans. Each agent has dynamic long-term memory built from past consultations, and can use medical tools like phenotype matchers or drug interaction databases. RareAgents outperforms state-of-the-art clinical models and GPT-4o in both differential diagnosis and rare disease medication recommendation. It also introduces a new benchmark dataset, MIMIC-IV-Ext-Rare, supporting further research. Figure 12 shows the multi-agent framework.

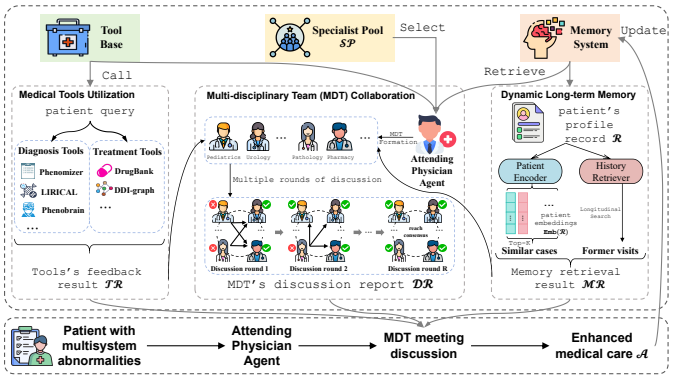

Fig. 12. Overview of the RareAgents framework. When a patient presents with multi-organ or complex symptoms, the Attending Physician Agent assembles a multi-disciplinary team by selecting specialist agents from a predefined pool. These agents reach a diagnostic consensus through iterative discussion. Each specialist agent is equipped with dynamic long-term memory for retrieving relevant cases and tools to support diagnosis and treatment planning. Adapted from [27]

Another example is MAGDA [26], a diagnostic assistant built around clinical guidelines. It includes a retriever agent that finds the right guideline, a reasoning agent that applies it to the patient's case, and a verifier agent that checks for consistency. This setup allows the system to reason with medical rules while keeping the logic traceable.

Some agents debate with each other to improve results. In this setup, one agent might suggest a diagnosis, while another agent challenges it. Through back-and-forth critique, the system can catch errors or consider rare possibilities. This idea, sometimes called self-reflection or adversarial review, helps reduce hallucinations and improve reliability in difficult cases like rare disease diagnosis. Studies have shown that even simple critique loops can reduce errors in LLM agents.

Multi-agent systems can also combine different data types. For instance, a vision model can analyze a medical image and describe what it sees, and a language model can then read the description and explain the results to a patient or doctor. Liu et al. [43] showed this setup using a vision model to read chest X-rays and GPT-4 to write the impression. This design supports multi-modal reasoning in tasks like radiology reporting, surgical planning, and patient monitoring.

However, multi-agent systems are harder to build. Agents need to pass messages clearly, manage shared memory, and recover from failures. If one agent makes a mistake, others can repeat or amplify it. To address this, some systems use a final verifier agent to check all outputs before they are returned. Frameworks like AutoGen [91] and LangGraph [92] help manage communication and coordination among agents.

Although many multi-agent medical systems are still early in development, this design is gaining traction. ClinicalAgent [93], RareAgents [27], and Zodiac [60] all explore multi-agent collaboration in trial matching, guideline-driven care, and cardiology. As the field grows, it is likely that more systems will use teams of agents to reflect how real clinical care is delivered with collaboration, specialization, and safety checks built into the process.

E. Human-in-the-Loop and Safety Mechanisms

Across all uses of LLM medical agents, one theme is paramount: human oversight. Nearly every study and commentary emphasizes that these agents should assist, not replace, human healthcare professionals [94]. The concept of human-in-the-loop can take several forms:

- **Approval before action:** An agent drafts a recommendation or document, and a human approves or edits it before it is finalized. This is currently the norm in note generation (the doctor edits the AI-produced note) and likely will be the norm if agents propose treatments (the doctor must approve the plan). Benary et al.'s oncology study explicitly had oncologists review all LLM suggestions, and none were applied to patients without thorough expert validation [56].
- **Selective usage:** Clinicians might learn when to consult the agent and when not to. For example, a physician may use an LLM agent for a second opinion on a straightforward case to save time, but rely on their own judgment for a complex case. Alternatively, they might use the agent early in reasoning (for broad differential ideas) but switch it off when narrowing down to a final decision to avoid bias. A trial found that physicians did not significantly benefit from LLM help in diagnosis [95], possibly because they weren't trained in how to effectively incorporate the AI. Future training of medical professionals may need to include working with AI assistants as a skill.
- **Real-time monitoring:** In interactive settings (like a patient chatbot), a human moderator might oversee multiple agent-user conversations at once, ready to step in if the agent output seems problematic. This is analogous to how some mental health chatbot services operate: they have humans in the loop who get alerted if certain keywords (suicidal ideation, etc.) appear, then the human can join the chat.
- **Reinforcement Learning from Human Feedback (RLHF):** This technique, used to align ChatGPT, is essentially a human-in-loop training phase. Models like GPT-4 were refined by showing them good and bad outputs (ranked by humans) and optimizing accordingly [96]. For medical agents, a similar approach can be applied: have medical experts review agent outputs and use their feedback to fine-tune the agent.

- **Guardrails and policies:** Developers often hand-craft certain rules that act as a backstop. For example, an agent could be programmed to never give explicit medical advice without a disclaimer like, "I am not a medical doctor, but..." (if patient-facing), or to refuse certain requests (like providing a diagnosis purely based on very limited information). OpenAI's models have system instructions not to do certain things; similarly, a medical agent might have a policy: "If the user asks for a drug recommendation, always include a note that a doctor's evaluation is needed before starting any medication." These rules embody a human-in-loop philosophy at design time.

Despite these measures, safety and ethical considerations remain a major concern, as noted in multiple studies [45]. Hallucinations and unwarranted confidence are particular issues with LLMs in medicine [97]. An LLM might state an incorrect dosage with full confidence, which could be dangerous. The literature suggests a few mitigation strategies: (1) constraining outputs to a fixed set of options (reduces free-text errors), (2) requiring the agent to cite sources for factual claims (so the user can verify), and (3) using ensemble or "consultation" approaches (have two independent models, and if they disagree, flag for human review).

Another interesting benchmark is the tendency of LLMs to exhibit biases. Like treating demographic groups differently. While not yet extensively studied in agents, we know from general LLM research that biases exist [98]. A medical agent might inadvertently provide lower-quality answers for underrepresented groups if not carefully evaluated. This calls for bias testing as part of the validation of any clinical AI.

Regulatory bodies are starting to pay attention. If an LLM agent is used in a way that influences clinical decisions (like triaging patients or recommending treatments), it could be deemed a medical device by regulators such as the FDA. This would require rigorous clinical trials to prove safety and efficacy, much like a new drug or a diagnostic device. As of 2025, few if any LLM-based agents have undergone that level of regulatory approval. Most are in pilot or research stages, or deployed in limited, non-critical roles.

In conclusion, human-in-the-loop approaches are not just advisable but likely mandatory for LLM agents in healthcare at the current stage. The literature uniformly advocates for keeping the clinician or scientist as the final authority, using the AI as a supporting tool. Over time, if we accumulate evidence of reliability in narrow tasks, some agents might earn more autonomy under supervision (just as an autopilot is allowed to fly a plane but a human pilot must be there to intervene). Achieving an optimal synergy between AI agents and healthcare professionals by combining the tireless processing and knowledge recall of machines with the intuition and contextual understanding of humans is a key goal of the next phase of research.

V. Lessons from Other Domains: Adaptation of LLM Agents to Healthcare

Research in education, robotics, and scientific discovery has pioneered many LLM-agent designs that healthcare can leverage. We highlight a few cross-domain insights and how they might translate to medical applications:

A. Education (LLM Tutors and Student Agents)

In education, LLM-based agents have been used as virtual tutors, capable of teaching or quizzing students in a conversational manner [99], [100]. This idea could be adapted to medical training. For example, a pair of LLM agents simulating a doctor and patient, allowing a medical student to practice history-taking in a low-risk environment. Another educational use is LLMs as automated graders or evaluators [101].

The education domain has also explored reinforcement learning to improve LLM generalization in interactive simulations [102], which might inform how we train medical agents to handle the open-ended nature of clinical encounters. Overall, the lesson is that LLMs can assume pedagogical roles, something we see hints of in how they are used for physician exam preparation [55], and careful prompt designs (like Socratic questioning style) can be imported from education research to make medical agents better teachers or trainers.

B. Robotics (LLM planners for actions)

Robotics has embraced LLMs as high-level planners that interface with low-level controllers. A notable example is PaLM-SayCan by Google, where an LLM (PaLM) was used to interpret commands and suggest feasible actions for a robot, while a separate affordance function checked which actions are physically possible [103]. The phrase "the robot is the LLM's hands and eyes" encapsulates this synergy.

In healthcare, we can imagine this concept in contexts like surgery or nursing. A future surgical robot could have an LLM-based agent that understands a surgeon's high-level instruction ("stitch the wound in layer X closure") and translates it into the robot's motion commands, all while checking safety constraints. Closer to current reality, consider assistive robots in elder care. An LLM agent could process a spoken request from a patient ("I dropped my medicine, can you help me?") and plan a sequence like finding the pill on the floor, picking it up, and finally placing it on the table.

The key takeaway from robotics is the importance of grounding. LLM agents must be tied to the real world via sensors and effectors (or, in healthcare, via data and actions in the medical record or devices). Roboticists have also been working on safety in LLM plans, making sure a physical robot doesn't do something harmful. This aligns with healthcare's need for fail-safes. An LLM agent controlling a syringe pump or an insulin dispenser must have layers of safety checks, much like robots have emergency stop mechanisms.

C. Science and Research (LLM Scientific Assistants)

In scientific research, LLM agents are being used to write code (e.g., Copilot, GPT-Engineer) and even to autonomously design experiments. A recent direction is using LLMs to generate hypotheses and design the next experiment in a loop with laboratory robots. This is essentially closing the loop of a scientific discovery process.

This was demonstrated in some form in materials science and chemistry experiments by 2023 (though often using GPT-3 plus domain tools) [24]. Translating this to healthcare, one could imagine an agent that proposes a small clinical trial or A/B test in a hospital. For instance, noticing that two antibiotics are used for a condition with uncertain relative efficacy, the agent might suggest a trial protocol. It could then monitor outcomes (with human IRB approval, of course) and analyze results.

While this is futuristic, elements of it, like automatically analyzing clinical data to generate new insights, are quite plausible in the near term. Already, LLMs like GPT-4 can assist in writing research papers or summarizing data for systematic reviews. The "AI Scientist" concept from other sciences foreshadows AI epidemiologists or pharmacologists that could trawl through health records and scientific literature to generate new medical knowledge (for example, identifying a previously unnoticed side effect of a drug by linking disparate case reports). The multi-agent approach taken in some of these systems (planner, experimenter, analyst) can be mirrored in medical research contexts as discussed above in drug discovery.

In all these domains, one recurring theme is that LLM agents work best when integrated with domain-specific tools and constraints. Education agents use curricula or knowledge bases, robotics agents use sensorimotor data, and science agents use experimental data. The healthcare equivalent is integrating medical knowledge bases (like drug databases, clinical guidelines) and patient data streams with LLM reasoning.

Another theme is evaluation. Education has student test scores, robotics has task completion rates, and science has experimental validation. Likewise, healthcare LLM agents need rigorous evaluation on clinical outcomes or user satisfaction to truly prove their worth.

Thus, healthcare can adopt many innovations from other fields, like role-playing prompts and tutoring strategies from education [104], planning and grounding techniques from robotics, and autonomous experimentation cycles from scientific research. The cross-pollination of ideas accelerates progress. At the same time, medicine's unique ethical and safety requirements mean we must adapt these methods with care. A failure in a game or a coding task is minor, but a failure in patient care can be life-threatening. So, while an "AutoGPT"-like agent might autonomously debug software with little oversight, an autonomous agent in a hospital must be introduced gradually, with exhaustive testing. The experience in other fields is encouraging because it shows complex tasks can be handled by LLM agents, but it also serves as a caution that we should implement guardrails and domain checks as those fields do when stakes are high. For instance, a misstep by a factory robot can be dangerous, so the robotics field has developed formal verification for some AI plans, which is something healthcare AI might borrow.

VI. Taxonomy of LLM-based Agents for Medical Applications

To organize the rapidly growing field of medical LLM agents, we introduce a taxonomy based on primary function, level of autonomy, and type of tool integration. These dimensions were chosen because they reflect the most critical and recurring differences among systems observed in the literature. Grouping agents by function helps clarify what real-world tasks they aim to support, from clinical documentation to patient counseling. Autonomy levels distinguish between fully automated systems and those designed for human oversight, which is especially important for safety and regulatory concerns. Tool usage captures the architectural complexity of the agent, whether it retrieves external knowledge, uses APIs, or processes multi-modal inputs. This structured view helps readers compare agents across diverse domains and identify tradeoffs between capability, safety, and deployment readiness. It also offers a practical framework for researchers and practitioners to evaluate or design new agents tailored to specific clinical use cases. Tables IV, V, and VI provide a summary of representative LLM-based agents in healthcare under this taxonomy, and we elaborate on the categories below.

A. By Primary Function or Role

- **Knowledge Retrieval and Q&A Agents:** These agents focus on providing information. For example, answering clinical questions or patient inquiries. They serve as a knowledge interface, often augmented with medical databases. Examples: a radiology Q&A chatbot for trainees, or a patient-facing FAQ agent for chronic disease management [55]. Their outputs are typically advisory (answers, explanations) rather than action-oriented.
- **Clinical Decision Support Agents:** Agents that assist with diagnoses, differential diagnosis generation, and treatment recommendations play a reasoning role, analyzing patient data (often via text input) to suggest next steps [105]. These usually involve complex reasoning and often require human validation (hence low autonomy in deployment). Med-PaLM and GPT-4 when used to propose diagnoses or plans, fall here.
- **Documentation and Workflow Agents:** Agents whose main job is generating or managing text in the clinical workflow, like writing notes, summarizing visits, drafting reports. This category (also called administrative assistants) includes discharge summary generators [62], transcription-based note creators [64], and workflow triage bots that route tasks (like GPT-4 triaging radiology requests). They primarily produce documentation or classifications [29].
- **Patient Interaction and Counseling Agents:** These interact directly with patients or non-expert users. Their function is communication, answering questions, gathering history (some symptom checkers ask questions), and providing counseling (diet, mental health, etc.) [59]. Empathy and clarity are key here. They often overlap with knowledge Q&A agents but are distinct in having a conversational, perhaps longitudinal interaction with patients.
- **Research and Discovery Agents:** Agents used by scientists or data analysts for knowledge discovery, such as literature review assistants, hypothesis generators, or data

analysis bots [73]–[75]. Their function is to accelerate research insights (e.g., summarizing 100 papers or suggesting biological targets for a disease). They might not be patient-facing at all, but they operate in silico based on data and literature.

TABLE IV
DIMENSION 1: PRIMARY FUNCTION OR ROLE

Function	Examples
Knowledge Retrieval and Q&A Agents	[16], [18], [30]
Clinical Decision Support Agents	[38], [50], [106]
Documentation and Workflow Agents	[23], [29], [52]
Patient Interaction and Counseling Agents	[47], [68], [107]
Research and Discovery Agents	[24], [73], [74]

Table IV summarizes the key functional roles LLM agents fulfill in healthcare, highlighting their diverse applications from information retrieval to patient interaction. This classification sets the stage for understanding how autonomy and tool use further differentiate these systems.

B. By Level of Autonomy

- **Autonomous Agents:** Those capable of performing sequences of actions with minimal human input, beyond just generating an answer [108]. These often integrate planning and can initiate tasks (like running database queries [67], scheduling follow-ups) on their own. True autonomy in medicine is rare so far due to safety concerns. An example might be an agent that autonomously monitors ICU data and directly adjusts ventilator settings (hypothetically, none deployed like this yet). In our literature survey, autonomy is mostly limited to simulation settings.
- **Hybrid (Human-in-loop) Agents:** The majority of current systems. They operate independently to a point, like drafting a note or suggesting a diagnosis, but a human approves or intervenes before final decisions. Their autonomy is constrained by oversight. For instance, an agent that generates a prescription order that a physician must sign off is semi-autonomous [68], [109].
- **Passive AI Assistants:** These are more like traditional AI tools where the agent doesn't initiate any action; it only responds to specific prompts, and every step is user-driven. For example, using ChatGPT to answer a single question on demand is essentially a passive use (the LLM isn't proactively managing a process). Many current "agents" in healthcare are of this form. They act when asked, but don't have an ongoing goal or initiative.

Table V categorizes agents by their autonomy, reflecting the spectrum from fully autonomous systems to passive assistants. This dimension emphasizes the varying degrees of human oversight and intervention, critical for ensuring safety and trust in clinical settings.

TABLE V
DIMENSION 2: BY LEVEL OF AUTONOMY

Autonomy	Examples
Fully Autonomous	[32], [67], [108]
Human-In-The-Loop	[51], [110], [111]
Passive	[9], [45], [72]

C. By Tool/Resource Integration

- **Knowledge-Integrated Agents:** Agents connected to knowledge bases or retrieval systems (RAG approach [18], [88]). They pull in patient data or medical literature as needed. Most clinical Q&A and decision support agents fall here, as they use patient records or references.
- **Tool-Augmented Agents:** Agents that can perform non-text actions via tools, like searching the web, performing calculations, updating records, and interfacing with medical devices. For instance, an agent that can actually order a lab test through the hospital's system (via API) when instructed. Few are fully implemented in practice yet, but research systems like ChemCrow (with chemical tools) exemplify this [36]. In a medical prototype, an agent might use a "Medical Calculator" tool if the prompt involves computing an MDRD GFR from creatinine, etc.
- **Multi-modal Agents:** A special sub-case where the agent can process or output non-text modalities, like medical images. GPT-4 Vision and similar multi-modal models hint at this future. An agent that can look at an X-ray image and dictate a report is multi-modal. While GPT-4 Vision (2023) had some capability, a study found it still struggled with detailed radiology reports [112]. But vision integration will likely improve, bridging into the radiology and pathology domains directly.

TABLE VI
DIMENSION 3: TOOLS/RESOURCE INTEGRATION

Tools	Example
Knowledge-Integrated	[56], [67], [88]
Tool-Augmented	[30]–[32]
Multi-modal	[21], [22], [113]

Table VI outlines the integration of external tools and resources, showing how agents range from standalone language models to sophisticated multi-modal systems interfacing with diverse data types and medical devices. This aspect is key to enhancing agent capabilities beyond pure language understanding.

It's worth noting that these dimensions overlap. For example, a "clinical documentation agent" (function) might be semi-autonomous and use knowledge integration (tool). In Table VII, we identify key examples by their primary function, but also note agent type and tool use.

TABLE VII
REPRESENTATIVE LLM-BASED AGENTS IN HEALTHCARE (2022 – 2025)

Function	Autonomy	Tool Use	Example Application	Reference
Knowledge Retrieval & Q&A	Fully Auto.	Multi-modal	Pathology Q&A	[114]
	Fully Auto.	Tool-augmented	Answering medical questions	[30]–[32], [109]
	Passive	Tool-augmented	Information retrieval	[115]–[117]
Clinical Decision Support	Fully Auto.	Tool-augmented	Disease phenotyping	[37], [118]–[122]
	Fully Auto.	Multi-modal	Cardiology diagnosis	[21], [22], [60]
	Human-loop	Knowledge-int.	Radiology report generation	[48], [123]
	Human-loop	Tool-augmented	Diagnostic reasoning	[57], [112], [124]
	Human-loop	Multi-modal	Bedside support	[113]
Documentation & Workflow	Fully Auto.	Multi-modal	Pathology interpretation	[125]
	Fully Auto.	Knowledge-int.	Thrombectomy reporting	[105]
	Fully Auto.	Tool-augmented	EHR navigation	[29], [126]
	Human-loop	Knowledge-int.	EHR navigation	[35]
	Human-loop	Tool-augmented	Radiology reporting	[23], [52]
Patient Interaction & Counseling	Fully Auto.	Multi-modal	Patient Simulation	[127]
	Fully Auto.	Tool-augmented	Differential diagnosis	[128]
	Human-loop	Knowledge-int.	Prostate-cancer counseling	[67]
Research & Drug Discovery	Fully Auto.	Tool-augmented	Drug-target discovery	[24], [25], [129], [130]
	Fully Auto.	Multi-modal	Molecule design	[73], [74]
	Passive	Tool-augmented	De-novo molecule generation	[72], [131], [132]

This taxonomy helps clarify discussions of "LLM agents" by breaking the broad concept into meaningful sub-classes. It shows that not all medical LLM agents are alike. A patient chatbot is very different from an autonomous research synthesis agent. By categorizing them, we can better enumerate requirements and challenges for each type. For instance, a patient-facing agent needs strong ethical safeguards and empathy modeling, whereas a research agent needs heavy integration with literature databases and might prioritize depth of reasoning.

Table VII illustrates the landscape of LLM-based medical agents. We see that radiology has multiple entries, reflecting the active research in imaging applications. Clinical documentation and education also show strong results (with LLMs sometimes matching or beating human performance in narrow tasks [47], [49]). Decision support entries reveal the gap between current LLM capabilities and the rigor required for unsupervised clinical use.

Patient-facing agents are emerging, with promise in information delivery but a clear need for oversight and personalization. In drug discovery and research, the table includes examples to show how advanced agent architectures and tool integrations developed in other fields can be ported into biomedical contexts.

When classifying an LLM agent, one should consider: What is it fundamentally trying to do?, How autonomously is it operating?, and What external resources does it use? Our taxonomy and table together aim to provide a conceptual map to place any given medical LLM agent into context. For instance, an agent that converses with patients to gather symptoms and then writes a draft note actually spans categories: it's a patient interaction agent initially, then a documentation agent, likely semi-autonomous and connected to the EHR. Such an agent would face the combined challenges of those categories (needing both empathy and factual accuracy).

As the field progresses, this taxonomy may evolve. New dimensions might include regulatory classification (FDA-approved vs experimental) or learning approach (fully pre-trained vs continually learning online, which introduces another set of concerns in medicine). But the proposed categories should remain relevant, as they are based on fundamental aspects of healthcare delivery and AI system design.

VII. DISCUSSION: TRENDS, OPPORTUNITIES, AND LIMITATIONS

The recent surge of research in LLM-based healthcare agents reveals both exciting opportunities and sobering limitations. A clear trend is the push toward generalist medical AI systems that can perform a wide array of tasks (often with minimal task-specific tuning) by virtue of broad foundational knowledge. This is exemplified by efforts like Med-PaLM and GPT-4's evaluations on medical exams [54], [81]. The allure is that a single model that might answer a patient's question one minute, help a doctor with a diagnosis the next, and then summarize a research article, adapting to each context. However, achieving reliability across such varied tasks is a major challenge.

A. Breadth vs. Depth Tradeoffs

One observation is that current LLM agents often excel at breadth over depth. They know something about many topics (as shown by passing board exams in multiple specialties [53]), but in any given narrow niche, a specialist (or specialist model) might do better. For example, GPT-4 can interpret basic chest X-ray findings, but a dedicated radiology CNN tuned on thousands of images might catch subtle cues GPT-4 misses [44].

This suggests a near-term strategy of hybrid systems: use LLMs for what they're best at (language, broad reasoning, adapting to new tasks) and use specialty AI or algorithms for well-defined subtasks (image recognition, pharmacokinetic calculations, etc.), with the LLM agent orchestrating the whole. In the taxonomy, this corresponds to tool-integrated

agents or multi-agent systems combining strengths. In practice, we might see hospital AI systems where an LLM agent calls on a suite of "sub-AI". For example, an ECG interpretation model, a dermatology image classifier, and then composes a summary for the clinician. This modular approach could marry the depth of narrow AI with the flexibility of LLMs.

B. Data Quality, Missing Modalities, and Real-World Variability

A major challenge in building and deploying LLM-based agents for healthcare is the quality, completeness, and consistency of the data they rely on. Medical data in real-world clinical settings is often fragmented, inconsistently recorded, and incomplete across modalities. Electronic health records may lack recent lab results, omit medications, or contain clinical notes with vague or ambiguous language. Some patient records might not include diagnostic imaging or pathology findings that are essential for comprehensive decision-making.

These limitations affect both model training and downstream inference. During training, most domain-specific language models are built using curated datasets that include published literature, question-answer pairs, and structured records from public sources. While these are valuable, they rarely capture the irregularities, noise, and missing information that are common in real clinical workflows. When a model trained on idealized data is applied to messy, incomplete inputs at the point of care, performance can degrade significantly.

Robustness to missing or partial inputs is especially important for LLM agents that aim to support clinical tasks. A triage agent, for example, might be asked to make a recommendation based on limited history, incomplete lab work, or sparse vital signs. If the model is not explicitly trained to handle such scenarios, it may produce unsafe or misleading outputs. Worse, it might express undue confidence without recognizing the gaps in information.

To address these issues, researchers are exploring multiple strategies. One approach involves masked training, where certain inputs are intentionally removed during training to simulate incomplete records. This helps the model learn to operate under uncertainty and avoid overconfident predictions. Another strategy is to integrate retrieval tools that allow agents to search structured databases or request additional data from connected systems when a key piece of information is missing. Retrieval-augmented generation and multi-agent coordination can also help distribute responsibilities across specialized modules.

In the training phase, techniques such as synthetic data augmentation and multi-source data fusion are used to expose models to a wider range of input conditions. Some groups are leveraging federated learning to train across diverse healthcare institutions, which helps improve generalization to varied patient populations and documentation styles, without compromising privacy.

Ultimately, clinical environments are dynamic and imperfect. Designing LLM agents that can recognize when data is missing, ask clarifying questions, flag uncertainties, or defer to human oversight is essential for safe deployment. Building this kind of robustness is not a simple matter of scale or computation. It requires thoughtful alignment between model design, data engineering, and clinical context.

C. Evaluation Challenges

Another trend is the importance of evaluation on real-world use cases. Many early papers used proxies like exam questions or synthetic cases to test LLMs [133], [134]. While useful, these don't capture the complexity of live settings. Emerging work is starting to do clinical simulations or prospective studies like the JAMA randomized clinical trials with physicians [57].

These often show a drop in performance compared to benchmark tests, underlining the classic reality gap between lab and practice. As LLM agents move toward deployment, we need rigorous clinical trials and user studies to truly measure impact on outcomes. For example, does a documentation agent actually save physicians' time and reduce errors? Does a triage agent safely reduce wait times in the emergency room? We also need usability research: even a high-performing agent might fail if it doesn't integrate well into workflows or if users don't trust it.

D. Factual Reliability and Safety

Hallucination and factuality remain fundamental issues, with recent surveys providing comprehensive taxonomies of hallucination types and challenges in LLMs, particularly in high-stakes domains like healthcare [135]. As noted, tool-use and retrieval help, but hallucinations haven't been fully eliminated. For instance, an LLM might cite non-existent articles or give outdated treatment advice confidently [45]. Approaches like citing sources [136] and including uncertainties, like having the agent express uncertainty when the information is not clear, are being explored.

Encouragingly, some evidence suggests domain-tuned models hallucinate less about their domain. Singhal et al. reported that Med-PaLM's answers were not only more accurate but also had a lower rate of "potentially harmful" content compared to the base model [10]. Still, absolute safety is elusive. Even human doctors make errors, but we hold AI to a higher initial standard if it's to be widely used. Recent work shows that LLMs often lack metacognitive capabilities needed for reliable clinical reasoning, which further complicates their deployment in high-stakes settings [137].

E. Bias and Fairness

If an agent is used for patient-facing advice or as a decision aid, we must ensure it works equitably for all populations. There's concern that LLMs might perpetuate biases present in training data (like underrepresenting certain groups or suggesting different treatments based on race/gender inappropriately) [138]. Studies in other domains (like GPT-3 showing different sentiments toward different demographic names) raise red flags. Medical LLM evaluations should include testing on cases representing diverse patient backgrounds to detect biases. Moreover, an agent might need additional conditioning

or data to handle language dialects or cultural contexts of patients (a chatbot might misunderstand a non-native speaker's description of symptoms, for example).

F. Privacy and Data Protection

Privacy is another limitation/opportunity dimension. LLMs can inadvertently leak information from training data, and when used on patient data, could expose sensitive info if not properly controlled. Solutions include on-premise deployment (keeping the model within the health provider's secure environment), differential privacy techniques in training (to prevent memorizing individual records), and careful access control. Interestingly, some see opportunity in using LLMs to detect privacy issues – e.g., an agent that scans output to ensure no patient identifiers are being revealed inappropriately (like in a research paper draft).

G. Regulatory and Ethical Considerations

Regulation and ethical guidelines for LLM agents are still catching up. Bodies like the World Health Organization have issued cautionary statements about generative AI in healthcare, emphasizing that patient safety, accountability, and transparency are non-negotiable. There's an open question of how to attribute responsibility when an AI agent is involved. If a doctor follows an AI's suggestion and an error occurs, who is liable? Likely, the human is, at present. This means AI agents should be used in a way that the human user knows or can know the basis of the suggestion (hence the interest in explainability of LLM decisions, even if partial). Some research into explainable LLMs is looking at letting the model output not just answers but reasoning and evidence [43], which could help a clinician justify or reject the agent's output.

H. Integration and Implementation

Integration with electronic health systems is both a technical and social challenge. Technically, connecting an LLM to an EHR API is doable (as prototypes show [61], [139]), but making it robust against the messy, inconsistent nature of real EHR data (typos, fragmented records, different schemas) is hard. Socially, clinicians may resist if an AI agent adds complexity or even if it saves time, but they fear medico-legal consequences. Effective integration likely needs end-user input in design and gradual rollout.

I. Future Directions

Looking ahead, we see benchmarking and leaderboards emerging for LLM agents on specific medical tasks. For example, the MEDIQA challenge series has started including tasks like summarization and question answering on medical texts, where multiple teams (often using LLMs) compete [64]. Similarly, there may be future leaderboards for "AI clinical assistant" performance, possibly run by NIH or other bodies, where standardized patient cases are used to evaluate end-to-end agent performance. These benchmarks can drive progress and shine light on which approaches work best.

Finally, it's worth noting that the pace of improvement in LLMs is rapid. What was cutting-edge in 2021 [81] was eclipsed by 2023 (GPT-4 near-expert on many tasks). If this pace continues, which is uncertain as scaling laws may hit limits, many current "limitations" might be partially addressed by simply more capable models. For instance, a model that's 10× its current size with better training might hallucinate far less and understand complex clinical nuance more.

However, some issues, like the need for up-to-date knowledge, integration with specific hospital data, will not be solved by scale alone. They need system-level solutions like tools, fine-tuning, etc. Moreover, we must be mindful that bigger is not always accessible: not every hospital can run a 540B-parameter model. The open-source community is working on smaller, specialized models (like LLaMA-derived med models), which might democratize access. The literature already has examples of distilled models that approach larger model performance on specific tasks, which could be crucial for widespread use, especially in resource-limited healthcare settings globally [140].

In conclusion, LLM-based agents in healthcare are at a frontier: the capability is here, as demonstrated by myriad studies, but real-world utility at scale is just beginning to be tested. The trend is clearly toward more integration (with tools, data, and workflows) and more specialization (fine-tuning for domains, multi-agent collaboration). The potential benefits, like improved efficiency, reduced errors, increased patient engagement, and accelerated research, are enormous, which explains the optimism and high pace of research. At the same time, the limitations, like accuracy, bias, trust, and safety, are equally monumental challenges that we must address through interdisciplinary effort (AI researchers, clinicians, ethicists, and engineers working together). The literature surveyed provides a foundation and early milestones on this journey. Going forward, expect to see iterative improvements: today's "assistant" agents becoming more reliable co-pilots in healthcare, perhaps one day culminating in AI agents that are as ubiquitous and trusted in medicine as stethoscopes and medical calculators, always with the aim of enhancing human expertise, not replacing it.

VIII. Conclusions

LLM-based agents represent a transformative development in healthcare AI, with the capability to understand and generate medical language, reason over clinical information, and interact with both humans and digital systems. Since 2018, research has rapidly expanded from proof-of-concept Q&A bots to a broad spectrum of applications: assisting in medical imaging interpretation, automating clinical documentation, providing decision support, engaging with patients, and accelerating drug discovery.

This survey has reviewed key peer-reviewed contributions in these areas, emphasizing how specific LLM architectures (GPT-3/4, domain-tuned models like Med-PaLM and BioGPT [9], [10]) and agent designs (prompting strategies, tool integration, memory, multi-agent collaboration) are being utilized. We also drew parallels to LLM-agent innovations in education,

robotics, and science, suggesting how techniques like role-playing, real-world grounding, and autonomous experimentation can be adapted to solve healthcare challenges.

A few recurring themes emerged. First, no single approach fits all in healthcare. Successful implementations tailor the agent to the task, often combining an LLM with external knowledge sources or constraints for reliability. Second, human oversight and partnership are crucial. The highest-performing systems treat the LLM agent as an assistant to amplify human capabilities, not a replacement. Third, evaluation and safety must be rigorously addressed before deployment in clinical settings, given the high stakes. Many studies, while optimistic, urge caution and further validation [97].

Our proposed taxonomy categorizes medical LLM agents by function, autonomy, and tooling, which can aid in systematically thinking about future developments. For instance, a fully autonomous diagnostic agent with tool use and multimodal input might be a future goal, but our review shows we are not there yet. Current agents that diagnose operate under supervision and often without direct image inputs, focusing on text data and relying on human confirmation. On the other hand, documentation agents are already nearing practical usability and may soon become common in hospitals [65]. Patient-facing agents are perhaps the most double-edged: they can empower patients with information, but also pose risks if they give incorrect advice, so any deployment will need careful guardrails and likely a human ”backstop” (like quick access to a nurse or doctor).

In terms of trends, the community is moving toward specialized excellence (fine-tuning or prompting LLMs to reach expert-level performance on specific tasks [42]) while also exploring generalist agents that can multitask and coordinate multiple abilities [24], [60]. This dual trajectory is likely to continue. There is also increasing collaboration between academic, industry, and clinical partners, as seen in some large studies, like multi-center evaluations of ChatGPT on medical exams [81], [95] or interdisciplinary teams developing oncology assistants [86]. Such collaboration will be key to addressing the remaining gaps.

In closing, LLM-based agents in healthcare are poised to become valuable teammates in clinical and research settings. The surveyed literature provides early evidence of their potential, from drafting high-quality clinical summaries to offering diagnostic suggestions and answering complex medical questions. At the same time, these studies illuminate the limitations that must be overcome through careful design, thorough validation, and a commitment to keeping human experts in the loop.

If these conditions are met, LLM agents could significantly enhance healthcare delivery: improving efficiency, expanding access to medical knowledge, and perhaps even raising the standard of care by reducing human error and complementing human expertise with tireless machine intelligence. The next few years will be crucial in translating the promising results in publications into real-world systems that safely and effectively benefit both clinicians and patients. The medical AI community should continue to critically evaluate, iterate, and ethically deploy these technologies, ensuring that this new generation of AI agents truly serves the goals of medicine. To better patient outcomes, public health, and support for the clinicians who dedicate themselves to care.

REFERENCES

[1] Jiarui Ye and Hao Tang. Multimodal large language models for medicine: A comprehensive survey, 2025.

[2] Manar Aljohani, Jun Hou, Sindhura Kommu, and Xuan Wang. A comprehensive survey on the trustworthiness of large language models in healthcare, 2025.

[3] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax?, 2025.

[4] Wasif Khan, Seowung Leem, Kyle B. See, Joshua K. Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine, 2025.

[5] Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. A survey on medical large language models: Technology, application, trustworthiness, and future directions, 2024.

[6] SUN Lei, WANG An'an, SONG Yimin, DONG Jing, LIU Xiaoli, LIANG Hong, LI Lixuan, SONG Xinyu, FAN Yong, JIA Zhilong, LI Tao, and ZHANG Zhengbo. Applications, challenges, and prospects of large language models in the field of clinical medicine. *ACADEMIC JOURNAL OF CHINESE PLA MEDICAL SCHOOL*, 45:1–11, 2024.

[7] 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, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine 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, Ben 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, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha 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, Johannes 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, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz 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, Mateusz 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 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario 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 Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll 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, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[9] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 09 2022.

[10] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[11] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023.

[12] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, August 2024.

[13] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025.

[14] Tugba Akinci D'Antonoli, Arnaldo Stanzione, Christian Bluethgen, Federica Vernuccio, Lorenzo Ugga, Michail E Klontzas, Renato Cuocolo, Roberto Cannella, and Burak Koçak. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and Interventional Radiology*, 30(2):80, 2024.

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

[16] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting

in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[17] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, 2024.

[18] Feiyuan Zhang, Dezhi Zhu, James Ming, Yilun Jin, Di Chai, Liu Yang, Han Tian, Zhaoxin Fan, and Kai Chen. Dh-rag: A dynamic historical context-powered retrieval-augmented generation method for multi-turn dialogue, 2025.

[19] Alex J Goodell, Simon N Chu, Dara Rouholiman, and Larry F Chu. Large language model agents can use tools to perform clinical calculations. *npj Digital Medicine*, 8(1):163, 2025.

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

[21] Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. Medagent-pro: Towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow, 2025.

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

[23] Fang Zeng, Zhiliang Lyu, Quanzheng Li, and Xiang Li. Enhancing llms for impression generation in radiology reports through a multi-agent system. *arXiv preprint arXiv:2412.06828*, 2024.

[24] Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration, 2025.

[25] Junde Wu, Jiayuan Zhu, and Yuyuan Liu. Agentic reasoning: Reasoning llms with tools for the deep research. *arXiv preprint arXiv:2502.04644*, 2025.

[26] David Bani-Harouni, Nassir Navab, and Matthias Keicher. Magda: Multi-agent guideline-driven diagnostic assistance, 2024.

[27] Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: Advancing rare disease care through llm-empowered multi-disciplinary team, 2025.

[28] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[29] Cyril Zakka, Joseph Cho, Gracia Fahed, Rohan Shad, Michael Moor, Robyn Fong, Dhamanpreet Kaur, Vishnu Ravi, Oliver Aalami, Roxana Daneshjou, et al. Almanac copilot: Towards autonomous electronic health record navigation. *arXiv preprint arXiv:2405.07896*, 2024.

[30] Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. Rationale-guided retrieval augmented generation for medical question answering. *arXiv preprint arXiv:2411.00300*, 2024.

[31] Chengrui Wang, Qingqing Long, Meng Xiao, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. Biorag: A rag-llm framework for biological question reasoning. *arXiv preprint arXiv:2408.01107*, 2024.

[32] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129, 2024.

[33] A. J. Goodell, S. N. Chu, D. Rouholiman, and L. F. Chu. Large language model agents can use tools to perform clinical calculations. *NPJ Digital Medicine*, 8(1):163, 2025.

[34] Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Seunghyun Won, and Edward Choi. Ehrnoteqa: A patient-specific question answering benchmark for evaluating large language models in clinical settings. *Preprint*, 2024.

[35] Roboam R Aguirre, Orlando Suarez, Mailenys Fuentes, and Marcos A Sanchez-Gonzalez. Electronic health record implementation: a review of resources and tools. *Cureus*, 11(9), 2019.

[36] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[37] Will E Thompson, David M Vidmar, Jessica K De Freitas, John M Pfeifer, Brandon K Fornwalt, Ruijun Chen, Gabriel Altay, Kabir Manghnani, Andrew C Nelsen, Kellie Morland, et al. Large language models with retrieval-augmented generation for zero-shot disease phenotyping. *arXiv preprint arXiv:2312.06457*, 2023.

[38] Songsoo Kim, Donghyun Kim, Hyun Joo Shin, Seung Hyun Lee, Yeseul Kang, Sejin Jeong, Jaewoong Kim, Miran Han, Seong-Joon Lee, Joonho Kim, Jungyon Yum, Changho Han, and Dukyong Yoon. Large-scale validation of the feasibility of gpt-4 as a proofreading tool for head ct reports. *Radiology*, 314(1):e240701, 2025. PMID: 39873601.

[39] Yucheng Shi, Peng Shu, Zhengliang Liu, Zihao Wu, Quanzheng Li, Tianming Liu, Ninghao Liu, and Xiang Li. Mgh radiology llama: A llama 3 70b model for radiology, 2024.

[40] Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, Sekeun Kim, Jiang Hu, Haixing Dai, Lin Zhao, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Tianming Liu, Quanzheng Li, and Xiang Li. Radiology-llama2: Best-in-class large language model for radiology, 2023.

[41] Zhengliang Liu, Aoxiao Zhong, Yiwei Li, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Peng Shu, Cheng Chen, Sekeun Kim, Haixing Dai, Lin Zhao, Lichao Sun, Dajiang Zhu, Jun Liu, Wei Liu, Dinggang Shen, Xiang Li, Quanzheng Li, and Tianming Liu. Radiology-gpt: A large language model for radiology, 2024.

[42] Roman Johannes Gertz, Alexander Christian Bunck, Simon Lennartz, Thomas Dratsch, Andra-Iza Iuga, David Maintz, and Jonathan Kottlors. Gpt-4 for automated determination of radiologic study and protocol based on radiology request forms: A feasibility study. *Radiology*, 307(5):e230877, 2023. PMID: 37310247.

[43] Zhaoyi Sun, Hanley Ong, Patrick Kennedy, Liyan Tang, Shirley Chen, Jonathan Elias, Eugene Lucas, George Shih, and Yifan Peng. Evaluating gpt-4 on impressions generation in radiology reports. *Radiology*, 307(5):e231259, 2023. PMID: 37367439.

[44] S. Ziegelmayer, A. Marka, N. Lenhart, N. Nehls, S. Reischl, F. Harder, A. Sauter, M. Makowski, M. Graf, and J. Gawlitza. Evaluation of gpt-4's chest x-ray impression generation: A reader study on performance and perception. *Journal of Medical Internet Research*, 25:e50865, 2023.

[45] Mahmud Omar, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. Overconfident ai? benchmarking llm self-assessment in clinical scenarios. *medRxiv*, 2024.

[46] Jeremy Qin, Bang Liu, and Quoc Dinh Nguyen. Enhancing healthcare LLM trust with atypical presentations recalibration. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2520–2537, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[47] Shaun S. Lim, Thanh D. Phan, Michelle Law, Gerard S. Goh, Henry K. Moriarty, Mark W. Lukies, Thomas Joseph, and Warren Clements. Non-radiologist perception of the use of artificial intelligence (ai) in diagnostic medical imaging reports. *Journal of Medical Imaging and Radiation Oncology*, 66(8):1029–1034, Dec 2022.

[48] Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Karan Singhal, Shekoofeh Azizi, Tao Tu, Mike Schaekermann, Rhys May, Roy Lee, SiWai Man, Zahra Ahmed, Sara Mahdavi, Yossi Matias, Joelle Barral, Ali Eslami, Danielle Belgrave, Vivek Natarajan, Shravya Shetty, Pushmeet Kohli, Po-Sen Huang, Alan Karthikesalingam, and Ira Ktena. Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation, 2023.

[49] Nathan W. Sterling, Fiona Brann, Samuel O. Frisch, and Jordan D. Schrager. Patient-readable radiology report summaries generated via large language model: Safety and quality. *Journal of Patient Experience*, 11, 2024.

[50] Jonathan Kottlors, Grischa Bratke, Philip Rauen, Christoph Kabbasch, Thorsten Persigehl, Marc Schlamann, and Simon Lennartz. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology*, 308(1):e231167, 2023.

[51] Su Hwan Kim, Severin Schramm, Jonas Wihl, Philipp Raffler, Marlene Tahedl, Julian Canisius, Ina Luiken, Lukas Endrös, Stefan Reischl, Alexander Marka, Robert Walter, Mathias Schillmaier, Claus Zimmer, Benedikt Wiestler, and Dennis M. Hedderich. Boosting llm-assisted diagnosis: 10-minute llm tutorial elevates radiology residents' performance in brain mri interpretation. *medRxiv*, 2024.

[52] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, and Michael Ingrisch. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European Radiology*, 34(5):2817–2825, 2024.

[53] Satheesh Krishna, Nishaant Bhambra, Robert Bleakney, and Rajesh Bhayana. Evaluation of reliability, repeatability, robustness, and con-

fidence of gpt-3.5 and gpt-4 on a radiology board–style examination. *Radiology*, 311(2):e232715, 2024. PMID: 38771184.

[54] Rajesh Bhayana, Robert R. Bleakney, and Satheesh Krishna. Performance of chatgpt on a radiology board-style examination. *Radiology*, 308(1):e231040, 2023.

[55] Tugba A. D'Antonoli, Arnaldo Stanzione, Christian Bluethgen, Federica Vernuccio, Lorenzo Ugga, Michail E. Klontzas, Renato Cuocolo, Roberto Cannella, and Burak Koçak. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*, 30(2):80–90, March 2024.

[56] Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, Ulrich Keilholz, Ulf Leser, and Damian T. Rieke. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689, 11 2023.

[57] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 10 2024.

[58] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

[59] G. S. Hill, J. L. Fischer, N. L. Watson, C. A. Riley, and A. M. Tolisano. Assessing the quality of artificial intelligence-generated patient counseling for rhinosinusitis. *International Forum of Allergy & Rhinology*, 14(10):1634–1637, October 2024. Epub 2024 Jun 18.

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

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

[62] Sajan B. Patel and Kyle Lam. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108, 2023.

[63] Robert Porter, Adam Diehl, Benjamin Pastel, J. Henry Hinnefeld, Lawson Nerenberg, Pye Maung, Sebastien Kerbrat, Gillian Hanson, Troy Astorino, and Stephen J. Tarsa. Llmd: A large language model for interpreting longitudinal medical records, 2024.

[64] John Giorgi, Augustin Toma, Ronald Xie, Sondra S. Chen, Kevin R. An, Grace X. Zheng, and Bo Wang. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models, 2023.

[65] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142, 2024.

[66] Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions, 2024.

[67] Lingxuan Zhu, Weiming Mou, and Rui Chen. Can the chatgpt and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *Journal of Translational Medicine*, 21(1):269, 2023.

[68] Jiayuan Zhu and Junde Wu. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. *arXiv preprint arXiv:2502.07143*, 2025.

[69] Y.H. Yeo, J.S. Samaan, W.H. Ng, P.S. Ting, H. Trivedi, A. Vipani, W. Ayoub, J.D. Yang, O. Liran, B. Spiegel, and A. Kuo. Assessing the performance of chatgpt in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clinical and Molecular Hepatology*, 29(3):721–732, Jul 2023.

[70] V. Sorin, E. Klang, M. Sklair-Levy, I. Cohen, D. B. Zippel, N. Balint Lahat, E. Konen, and Y. Barash. Large language model (chatgpt) as a support tool for breast tumor board. *NPJ Breast Cancer*, 9(1):44, May 2023.

[71] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1):210, 2023.

[72] Ye Wang, Honggang Zhao, Simone Sciabola, and Wenlu Wang. cmolgpt: A conditional generative pre-trained transformer for target-specific de novo molecular generation. *Molecules*, 28(11):4430, May 2023.

[73] Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Yingzhou Lu, and Yue Zhao. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration, 2025.

[74] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024.

[75] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025.

[76] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024.

[77] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

[78] Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, and Christian Lovis. Prompt engineering paradigms for medical applications: scoping review and recommendations for better practices, 2024.

[79] Declan Grabb. The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence*, 6(0), 2023.

[80] Douglas B. Flora and Nikhil G. Thaker. Designing prompts for generative artificial intelligence in clinical oncology contexts. *AI in Precision Oncology*, 1(1):19–21, 2024.

[81] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):1–12, 02 2023.

[82] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems, 2023.

[83] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[84] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[85] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

[86] Reza Khanmohammadi, Ahmed I Ghanem, Kyle Verdecchia, Ryan Hall, Mohamed Elshaikh, Benjamin Movsas, Hassan Bagher-Ebadian, Indrin Chetty, Mohammad M. Ghassemi, and Kundan Thind. Iterative prompt refinement for radiation oncology symptom extraction using teacher-student large language models, 2024.

[87] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In

H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[88] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

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

[90] Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. Llm-based medical assistant personalization with short- and long-term memory coordination, 2024.

[91] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.

[92] Jialin Wang and Zhihua Duan. Agent ai with langgraph: A modular framework for enhancing machine translation using large language models, 2024.

[93] Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning, 2024.

[94] Emre Sezgin. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digital Health*, 9:20552076231186520, Jul 2023.

[95] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 10 2024.

[96] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[97] Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. Medhalu: Hallucinations in responses to healthcare queries by large language models, 2024.

[98] Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. Bias patterns in the application of llms for clinical decision support: A comprehensive study, 2024.

[99] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. Simulating classroom education with llm-empowered agents, 2024.

[100] Fiammetta Caccavale, Carina L. Gargalo, Krist V. Gernaey, and Ulrich Krühne. Towards education 4.0: The role of large language models as virtual tutors in chemical engineering. *Education for Chemical Engineers*, 49:1–11, 2024.

[101] Hyein Seo, Taewook Hwang, Jeesu Jung, Hyeonseok Kang, Hyuk Namgoong, Yohan Lee, and Sangkeun Jung. Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences*, 15(2), 2025.

[102] Bahar Radmehr, Adish Singla, and Tanja Käser. Towards generalizable agents in text-based educational environments: A study of integrating rl with llms. *arXiv preprint arXiv:2404.18978*, 2024.

[103] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can, not as i say: Grounding language in robotic affordances, 2022.

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

[105] Nils C. Lehnen, Johannes Kürsch, Barbara D. Wichtmann, Moritz Wolter, Zeynep Bendella, Felix J. Bode, Hanna Zimmermann, Alexander Radbruch, Philipp Vollmuth, and Franziska Dorn. Llama 3.1 405b is comparable to gpt-4 for extraction of data from thrombectomy reports—a step towards secure data extraction. *Clinical Neuroradiology*, 2025. Published online 2025-02-25.

[106] Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms, 2024.

[107] Hengguan Huang, Songtao Wang, Hongfu Liu, Hao Wang, and Ye Wang. Benchmarking large language models on communicative medical coaching: a novel system and dataset, 2024.

[108] Yao Zhang, Zijian Ma, Yunpu Ma, Zhen Han, Yu Wu, and Volker Tresp. Webpilot: A versatile and autonomous multi-agent system for web task execution with strategic exploration. *arXiv preprint arXiv:2408.15978*, 2024.

[109] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[110] Joshua Strong, Qianhui Men, and Alison Noble. Trustworthy and practical ai for healthcare: A guided deferral system with large language models, 2025.

[111] E. Brügge, S. Ricchizzi, M. Arenbeck, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education*, 24:1391, 2024.

[112] Florence X. Doo and Vishwa S. Parekh. Beyond the ajr: Early applications of generative artificial intelligence for radiology report interpretation. *American Journal of Roentgenology*, 223(2):e2330696, 2024. PMID: 38117099.

[113] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments, 2024.

[114] Li Zhenzhu, Zhang Jingfeng, Zhou Wei, Zheng Jianjun, and Xia Yinshui. Gpt-agents based on medical guidelines can improve the responsiveness and explainability of outcomes for traumatic brain injury rehabilitation. *Scientific Reports*, 14(1):7626, 2024.

[115] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2), February 2024.

[116] Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, and Ramesh Jain. Conversational health agents: A personalized llm-powered agent framework, 2024.

[117] Dingkang Yang, Jinjie Wei, Mingcheng Li, Jiyao Liu, Lihao Liu, Ming Hu, Junjun He, Yakun Ju, Wei Zhou, Yang Liu, and Lihua Zhang. Medaide: Information fusion and anatomy of medical intents via llm-based agent collaboration, 2025.

[118] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 79410–79452. Curran Associates, Inc., 2024.

[119] Abhishek Dutta and Yen-Che Hsiao. Adaptive reasoning and acting in medical language agents, 2024.

[120] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents, 2025.

[121] Phuc Phan Van, Dat Nguyen Minh, An Dinh Ngoc, and Huy Phan Thanh. Rx strategist: Prescription verification using llm agents system, 2024.

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

[123] Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang SU, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. Medchain: Bridging the gap between llm agents and clinical practice through interactive sequential benchmarking, 2024.

[124] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL*

2024, pages 599–621, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[125] Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology, 2024.

[126] Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, Junyi Gao, and Liantao Ma. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2250–2261. ACM, April 2025.

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

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

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

[130] 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. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis, 2024.

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

[132] Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6):btae353, 2024.

[133] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021.

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

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

[136] Krishna Subedi. The reliability of llms for medical diagnosis: An examination of consistency, manipulation, and contextual awareness, 2025.

[137] Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16(1):642, 2025.

[138] J. L. Cross, M. A. Choma, and J. A. Onofrey. Bias in medical ai: Implications for clinical decision-making. *PLOS Digital Health*, 3(11):e0000651, 2024.

[139] Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Wenyue Hua, Mingyu Jin, Guang Chen, Yang Zhou, Zhao Li, Trisha Gupte, Ming-Li Chen, Zahra Azizi, Yongfeng Zhang, Themistocles L. Assimes, Xin Ma, Danielle S. Bitterman, Lin Lu, and Lizhou Fan. Aipatient: Simulating patients with ehrs and llm powered agentic workflow, 2024.

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