Is There a Dr. House in the Machine? Towards Clinical and Diagnostic Reasoning in LLM-based Medical Agents

Azmine Toushik Wasi[1,2♦†], **Abdur Rahman**[1,2♦], **Riashat Islam**[3], **and Md Rizwan Parvez**[4]
[1]Computational Intelligence and Operations Laboratory
[2]Shahjalal University of Science and Technology
[3]Microsoft AI Frontiers [4]Qatar Computing Research Institute
[♦]Equal Contribution [†]**Correspondence:** azmine32@student.sust.edu

Abstract

Large language models (LLMs) are increasingly used in agentic systems that reason over multiple steps, call tools, and interact with clinical knowledge sources. This survey examines whether LLM-based medical agents approximate expert diagnostic reasoning or reproduce its outputs without its causal and epistemic foundations. We organize agent design along five axes: reasoning paradigms, agent architectures, tool use, degree of autonomy, and embodiment. We then analyze learning and adaptation, evaluation, safety, and deployment as cross-cutting concerns that interact with these axes. Drawing on research in machine learning, biomedical informatics, and clinical AI, we examine how design choices shape diagnostic reliability and failure modes, and identify requirements for clinical and medical agents.

1 Introduction

Clinical diagnosis is an iterative process involving hypothesis generation, evidence aggregation, and uncertainty management, which traditional static ML systems often fail to capture, particularly under distribution shifts and rare diseases (Gu et al., 2025; Jin et al., 2021). High-performing medical language models frequently depend on surface correlations rather than grounded clinical inference, leading to brittle behavior, unsafe overconfidence, hallucinations, and reasoning failures (Gu et al., 2025; Huang et al., 2025a; Zuo and Jiang, 2025). These limitations have motivated a transition toward systems emphasizing diagnostic reasoning, verification, and tool-grounded evidence integration (Chen et al., 2025a).

This transition is reflected in the evolution from early rule-based clinical decision support systems to agentic medical AI capable of maintaining state, planning multi-step actions, and dynamically interacting with external tools and knowledge bases (Tsai et al., 2018; Roberts et al., 2018; Mehandru et al., 2024). LLMs have accelerated this shift by enabling flexible reasoning, natural language interaction, and tool orchestration within unified frameworks (Kim et al., 2024a; Wu et al., 2024). Multi-agent architectures further decompose workflows into specialized roles that mirror interdisciplinary clinical teams and mitigate individual model bias (Tang et al., 2024; Kim et al., 2024a), while methods such as Chain-of-Diagnosis, Tree-of-Thought reasoning, Monte Carlo Tree Search, and debate-based reflective frameworks structure diagnostics into interpretable and collaborative reasoning processes (Yao et al., 2023a; Guo et al., 2025b; Zuo et al., 2024). Nevertheless, many systems still rely on linguistic reasoning without robust causal understanding, limiting progress toward expert-level clinical performance (Ke et al., 2024).

In this survey, we provide a comprehensive review of LLM-based medical agents with emphasis on diagnostic reasoning and tool integration across clinical domains, as outlined in Figure 1 and 2. Section 2 introduces clinical reasoning, language models, and agent frameworks. Section 3 presents our five-axis framework: reasoning paradigms, agent architectures, tool use, degree of autonomy, and embodiment. It also summarizes how these design choices interact and shape evaluation and safety. The appendix provides the review methodology and expanded discussions of learning and adaptation, applications, datasets, evaluation, safety, deployment, and system case studies.

2 Foundations

2.1 Large Language Models in Medicine

We distinguish static predictor models from interactive agents. Medical dialogue datasets and transformer-based conversational systems have spurred single-agent AIs that engage directly with patients or clinicians (He et al., 2020). Medical pretraining and instruction tuning support coher-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1632–1667
November 6-10, 2026 ©2026 Association for Computational Linguistics

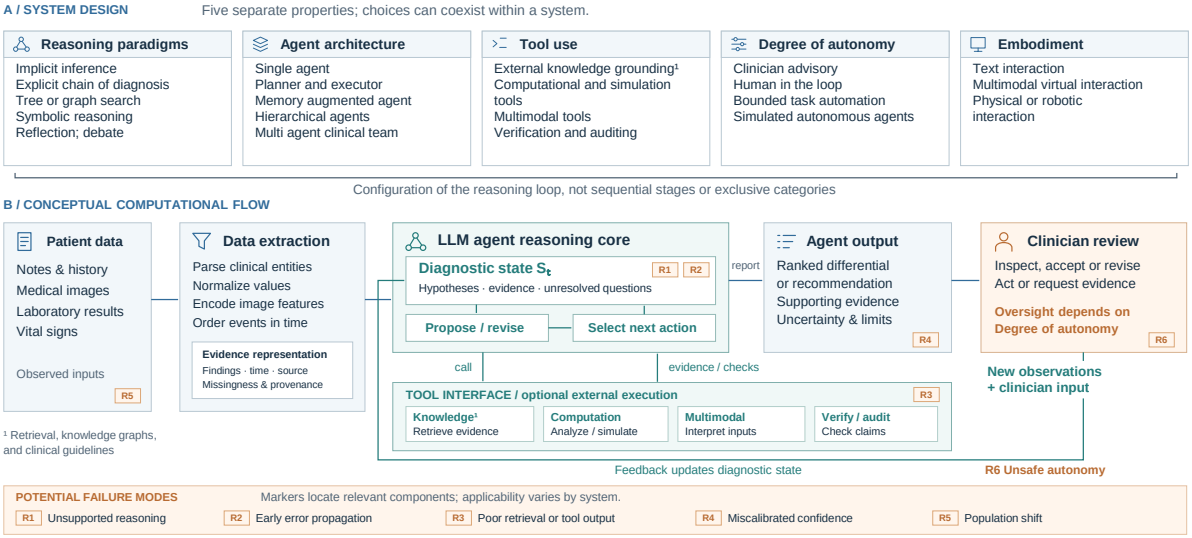


Figure 1: **Conceptual architecture of clinical reasoning in LLM based medical agents.** Five design axes characterize system configurations. Patient data drive diagnostic updating, tool use, recommendations, and uncertainty, with clinician input and new observations refining the state.

ent multi-turn interactions even with limited data. Meanwhile, embodied and web-interaction benchmarks expand agent capabilities beyond text, showing that LLMs are increasingly assessed as parts of interactive, multimodal systems (Yang et al., 2025c; Chezelles et al., 2025).

■ **Transformer Foundations.** LLMs are predominantly based on the Transformer architecture, which uses self-attention to model contextual dependencies in sequences (Vaswani et al., 2023). Given input tokens, attention computes contextualized representations, enabling scalable sequence modeling. This architecture has enabled large-scale pretraining on heterogeneous corpora, including biomedical text (Lee et al., 2019; Alsentzer et al., 2019). In medicine, Transformers have demonstrated strong performance on clinical text understanding, medical QA, and report generation (Jin et al., 2021; Roberts et al., 2018). However, architectural generality does not guarantee domain safety or reasoning fidelity (Gu et al., 2025).

■ **Medical Pretraining and Instruction Tuning.** Medical adaptation of LLMs typically involves domain-specific pretraining or instruction tuning on curated biomedical datasets. Models such as BioBERT and ClinicalBERT leverage masked language modeling on PubMed and EHR text to improve domain vocabulary and semantics (Lee et al., 2019; Alsentzer et al., 2019). Instruction tuning further aligns models with clinical tasks by training on input–output pairs representing diagnostic reasoning or medical dialogue (Jin et al., 2021; Sun et al., 2025). Multi-agent data generation pipelines have been used to scale high-quality medical reasoning datasets while controlling annotation cost (Li et al., 2025a; Du et al., 2025). Despite gains, instruction tuning alone often fails to enforce causal or guideline-consistent reasoning (Liu et al., 2025d).

■ **RLHF and Clinical Alignment.** Reinforcement Learning from Human Feedback (RLHF) aims to align LLM outputs with human preferences and safety constraints. In clinical contexts, reward models may encode preferences for factual accuracy, uncertainty expression, and guideline adherence (Ouyang et al., 2022; Reddy, 2024). Formally, RLHF optimizes a policy π_θ to maximize expected reward $\mathbb{E}_{y\sim\pi_\theta}[R(y)]$ under human-evaluated feedback (Ouyang et al., 2022). While RLHF reduces harmful outputs, studies show that it can mask uncertainty and encourage confident but incorrect answers (Huang et al., 2025a; Gu et al., 2025). Agentic verification and tool grounding are increasingly viewed as complementary to RLHF for clinical alignment (Liu et al., 2025c; Mukherjee et al., 2024).

■ **Limitations of Pure Language Modeling.** Despite impressive fluency, pure language modeling remains fundamentally limited for clinical diagnosis. LLMs lack direct grounding in patient physiology, causal mechanisms, and real-time clinical context (Jin et al., 2021). Empirical evaluations reveal that models can generate plausible but incorrect explanations unsupported by evidence, a phenomenon exacerbated in long reasoning chains

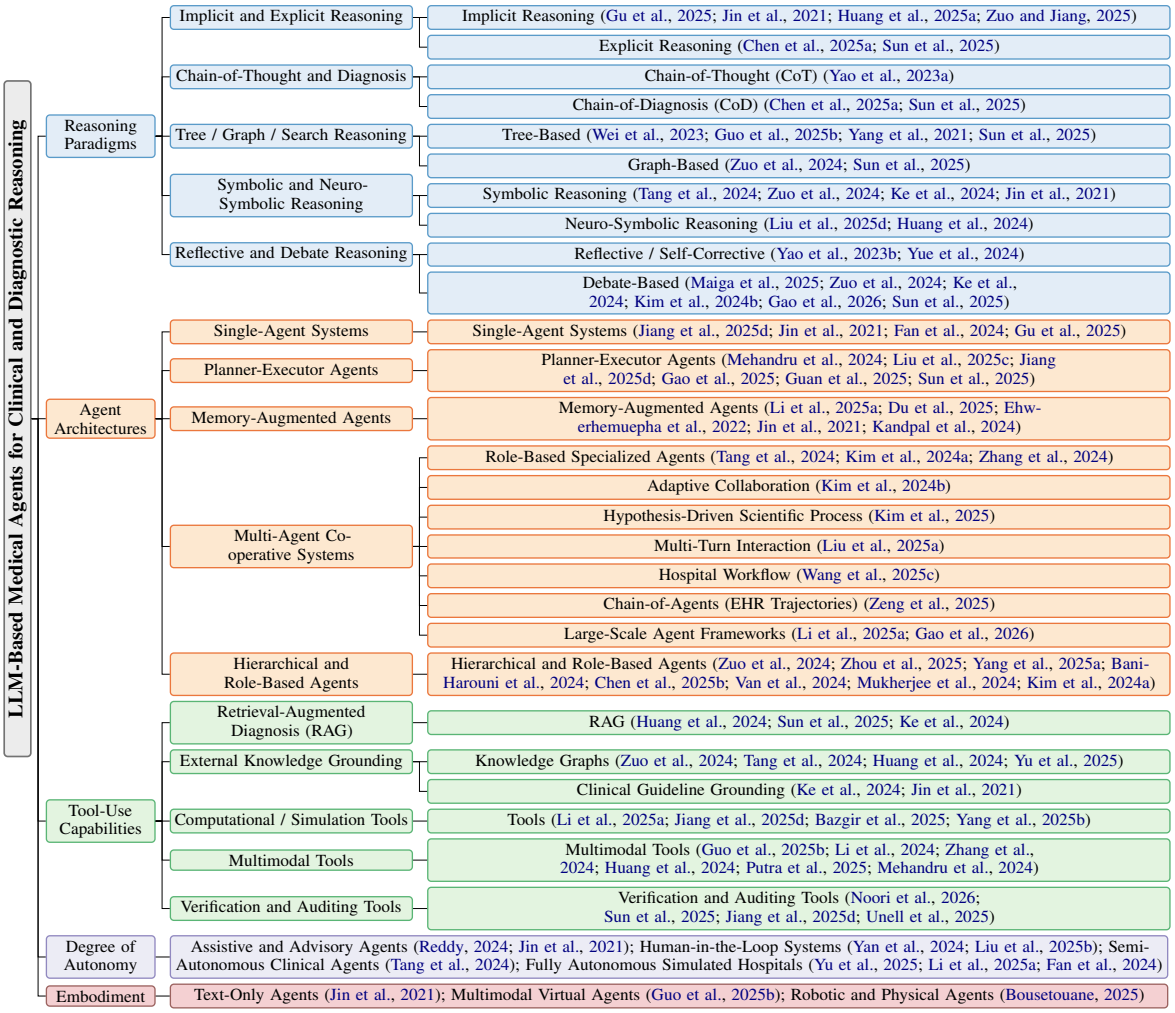


Figure 2: LLM-based medical agents organized along five major axes: reasoning paradigms, agent architectures, tool-use capabilities, degree of autonomy, and embodiment. The hierarchy reveals how diagnostic reasoning strategies interact with system architectures, external tools, autonomy levels, and physical embodiment.

(Li et al., 2023; Zuo and Jiang, 2025). Furthermore, next-token prediction does not inherently enforce logical consistency or probabilistic coherence (Chen et al., 2025a; Sun et al., 2025). These limitations motivate architectures that integrate external tools, symbolic reasoning, and structured decision processes (Li et al., 2024; Zuo et al., 2024).

2.2 From Models to Agents

■ **Definitions: Tools, Agents, Workflows.** A *tool* is an external function or resource that provides information or computation beyond the model's parameters, such as retrieval systems, calculators, or simulators. An *agent* is an autonomous system that maintains state, selects actions, invokes tools, and updates its strategy based on observations (Russell and Norvig, 2009; Mehandru et al., 2024). A *workflow* defines the structured interaction between reasoning, tool use, and environment feedback over time (Kim et al., 2024a). In medicine, workflows often span data retrieval, hypothesis generation, verification, and documentation (Tang et al., 2024; Jiang et al., 2025d). Formalizing these distinctions clarifies why diagnostic competence emerges from systems, not isolated models (Liu et al., 2025c; Zhuang et al., 2025).

■ **ReAct, Plan-and-Execute, and Reflective Loops.** Agentic reasoning frameworks commonly interleave thought, action, and observation. ReAct-style agents explicitly alternate reasoning traces with tool calls, enabling grounded inference (Yao et al., 2023b). Plan-and-Execute architectures first generate a high-level plan $P = \{a_1, \ldots, a_k\}$ and then execute actions sequentially with feedback (Mehandru et al., 2024). Reflective loops introduce self-critique functions $f_{\text{reflect}}(y)$ that evaluate and revise outputs (Yao et al., 2023b). In medical agents, these loops support diagnostic verification,

guideline checking, and error correction (Yue et al., 2024; Ke et al., 2024). Empirical results show that such architectures outperform single-pass reasoning on complex diagnostic tasks (Jiang et al., 2025d; Li et al., 2025a).

■ **Why Diagnosis Requires Agents, Not Chatbots?** Diagnosis unfolds over time, requiring memory, hypothesis revision, and adaptive information gathering, capabilities absent in stateless chatbots. Clinical reasoning depends on interaction with evolving patient data, tests, and consultations, which naturally maps to agent–environment loops (Tang et al., 2024; Jiang et al., 2025d). Static QA systems fail to model longitudinal dependencies or decision consequences (Jin et al., 2021; Fan et al., 2024). Agentic systems can explicitly represent diagnostic state $s_t = (H_t, E_t)$ and update it through actions a_t based on new observations (Russell and Norvig, 2009). Simulation studies demonstrate that agent-based medical systems achieve superior robustness and interpretability compared to conversational baselines (Fan et al., 2024; Li et al., 2025a). Consequently, meaningful progress toward expert-level diagnostic reasoning necessitates agents rather than chatbots.

3 LLM-Based Medical Agents

Here, we introduce a framework (Figure 2) that organizes LLM-based medical agents along five key axes. These five axes capture the main dimensions that determine how LLM-based medical agents reason, operate, and interact within clinical environments. Together they separate cognitive strategy (reasoning), system organization (architecture), external capability integration (tool use), operational control (autonomy), and interaction modality (embodiment), enabling systematic comparison of heterogeneous medical AI systems. Unlike prior surveys that primarily provide descriptive overviews of LLM-based medical systems and applications (Wang et al., 2025b; Jin et al., 2021; Roberts et al., 2018) (§C), this work adopts a reasoning-centric and diagnostic-first framework that systematically compares paradigms, agent architectures, tool integration, autonomy, and safety mechanisms to analyze how design choices shape failure modes, reliability, and clinical impact.

The five axes describe complementary system dimensions, not mutually exclusive paper categories. A system has a position on each axis, so a work may span several axes, which are conceptually distinct but not statistically independent. Boundary cases are assigned by primary function, with secondary interactions noted. Learning and adaptation, evaluation, safety, and deployment are cross cutting concerns rather than additional dimensions. Although the labels overlap with general agent taxonomies, the framework is medically specific: Axis I links reasoning to clinical cognition, Axis II addresses accountability across care teams, Axis III covers guideline grounding, clinical data formats, and protected health information, Axis IV connects autonomy with oversight, regulation, and medico legal responsibility, and Axis V captures the sensory and safety demands of clinical environments and vulnerable patients. These constraints help relate agent design to clinical failure modes.

3.1 Clinical Reasoning Foundations

The five axes of our framework (§3) are grounded in established theories of clinical cognition. Hypothetico-deductive reasoning describes how clinicians generate and iteratively refine ranked diagnostic hypotheses as new evidence emerges (Yang et al., 2021). Computationally, this resembles a search over a hypothesis space $\mathcal{H}$ where observations o_t update posterior probabilities $P(h \mid o_{1:t})$, making tree- and graph-based paradigms (Axis I) analogs of hypothesis pruning and hierarchical architectures (Axis II) analogs of staged reasoning. Bayesian diagnostic inference formalizes diagnosis as probabilistic updating, where priors $P(d)$ are revised using likelihoods $P(e \mid d)$ to yield posteriors $P(d \mid e) \propto P(e \mid d), P(d)$ (Yang et al., 2021), motivating uncertainty quantification and calibrated confidence estimation across all axes. Dual-process theory distinguishes fast System 1 pattern matching from slower System 2 deliberation (Kahneman, 2011); within our framework, implicit LLM reasoning approximates System 1, whereas explicit chain-of-diagnosis, search-based, and debate-based methods approximate System 2 for rare, conflicting, or high-stakes cases. Together, these theories justify the five axes by linking reasoning paradigms to cognitive strategy, architectures to clinical role structure, tool use to Bayesian evidence integration, autonomy to medico-legal constraints, and embodiment to available sensory modalities.

3.2 Axis I: Reasoning Paradigms

■ **Implicit and Explicit Reasoning.** Implicit reasoning relies on end-to-end language generation

where diagnoses emerge without explicit intermediate steps, often excelling on benchmarks due to scale but lacking transparency and faithfulness guarantees (Gu et al., 2025; Jin et al., 2021). Explicit reasoning externalizes inference steps for inspection and intervention (Chen et al., 2025a), offering greater interpretability favored in clinical settings despite verbosity and potential spurious rationales (Sun et al., 2025). Implicit methods are generally more robust to noise but prone to hallucination, especially under distribution shifts (Huang et al., 2025a; Zuo and Jiang, 2025). *In practice, implicit reasoning often performs well in large-scale benchmark settings, while explicit reasoning is preferred in clinical workflows where traceability, auditability, and human oversight are critical.*

■ **Chain-of-Thought and Chain-of-Diagnosis.** Chain-of-Thought (CoT) structures inference as linear reasoning steps, improving complex medical QA and diagnostics (Yao et al., 2023a). Chain-of-Diagnosis (CoD) further aligns these steps with clinical diagnostic phases, enhancing interpretability and calibration by modeling disease confidence distributions (Chen et al., 2025a; Sun et al., 2025). However, linear chains struggle with branching hypotheses and risk error accumulation, motivating exploration of non-linear reasoning. *These approaches are effective for structured diagnostic tasks but may degrade when multiple competing hypotheses must be explored simultaneously.*

■ **Tree-, Graph-, and Search-Based Reasoning.** Tree- and graph-based methods treat diagnosis as search over hypotheses, enabling parallel path exploration and pruning (Wei et al., 2023; Guo et al., 2025b), echoing hypothetico-deductive clinical reasoning (Yang et al., 2021). Graph reasoning supports relational inference, especially when combined with knowledge graphs (Zuo et al., 2024). Despite improved robustness, these methods are computationally intensive and require carefully designed rewards to avoid degenerate behavior, limiting real-time use (Sun et al., 2025). *Consequently, search-based reasoning is particularly suited to complex differential diagnosis scenarios where hypothesis branching is necessary.*

■ **Symbolic, Programmatic, and Neuro-Symbolic Reasoning.** These approaches integrate medical rules, logic, and probabilistic models with LLMs to constrain inference (Tang et al., 2024; Zuo et al., 2024), combining neural flexibility with symbolic rigor (Liu et al., 2025d; Huang et al., 2024). Symbolic grounding reduces hallucination in critical tasks (Ke et al., 2024) but suffers brittleness and maintenance challenges (Jin et al., 2021), leading to growing preference for hybrid designs. *Hybrid neuro-symbolic systems therefore offer a promising compromise between interpretability, reliability, and adaptability in safety-critical medical contexts.*

■ **Reflective, Self-Corrective, and Debate-Based Reasoning.** Reflective reasoning adds iterative self-evaluation and revision, using single or auxiliary models to refine diagnoses (Yao et al., 2023b; Yue et al., 2024). Debate-based frameworks extend this to multi-agent hypothesis challenges, reducing anchoring bias and improving error detection (Maiga et al., 2025; Zuo et al., 2024; Ke et al., 2024). MDAgents enable moderated group critique to enhance accuracy (Kim et al., 2024b), while TOOLUNIVERSE orchestrates complex tool-augmented debate pipelines for verification beyond single-pass LLMs (Gao et al., 2026). Yet, increased computation and risk of incorrect consensus highlight the need for grounding and expert oversight (Sun et al., 2025). *Such collaborative reasoning strategies are particularly valuable in high-uncertainty cases where independent verification and critique can mitigate individual model biases.*

Clinical reasoning paradigms involve trade-offs between efficiency, interpretability, robustness, and computational cost: linear methods such as CoT prioritize simplicity and speed, whereas search-, debate-, and symbolic-based approaches emphasize hypothesis exploration, error correction, and structured knowledge integration, leading many medical agents to combine multiple paradigms for balanced performance. Unlike general-domain reasoning, medical reasoning must satisfy physiological causality, strict time constraints, and asymmetric clinical risks between false negatives and false positives, making calibrated and auditable explicit reasoning essential in high-stakes settings. In clinical use, reasoning must respect physiological causality, time constraints, and the asymmetric costs of false-negative and false-positive decisions.

3.3 Axis II: Agent Architectures

■ **Single-Agent Systems.** Single-agent systems encapsulate reasoning, planning, and tool use within a single LLM instance, making them simple to deploy and evaluate in early clinical prototypes (Jiang et al., 2025d). However, they struggle with long-horizon reasoning, interdisciplinary tasks, and self-critique (Jin et al., 2021). Empirical studies show that although single agents perform competitively

on static benchmarks, their performance degrades in interactive or longitudinal diagnostic settings (Fan et al., 2024; Gu et al., 2025). *Overall, single-agent designs offer simplicity and efficiency but lack the collaborative reasoning required for complex clinical workflows.*

■ **Planner-Executor Agents.** Planner–executor agents decompose diagnosis into explicit planning and sequential execution stages (Mehandru et al., 2024). This separation improves controllability and enables structured multi-step reasoning in tasks such as medication analysis and diagnostic workflows (Liu et al., 2025c; Jiang et al., 2025d; Gao et al., 2025). Systems like PharmaData-Agent demonstrate how explicit planning supports transparent clinical data analysis pipelines (Guan et al., 2025). However, planning errors may propagate through execution, and static plans can become outdated as new patient information emerges (Sun et al., 2025). *These architectures improve interpretability and control but require adaptive replanning mechanisms to remain reliable in dynamic clinical environments.*

■ **Memory-Augmented Agents.** Memory-augmented agents maintain representations of patient history, prior cases, or long-term clinical knowledge to support longitudinal reasoning and continuity of care (Li et al., 2025a; Du et al., 2025; Ehwerhemuepha et al., 2022). Empirical analyses show that memory improves contextual consistency and reduces redundant questioning across encounters (Jin et al., 2021). However, persistent memory introduces risks of stale or biased information accumulation and increases privacy concerns when sensitive data is stored or reused across reasoning cycles (Kandpal et al., 2024). *Designing principled memory update and forgetting policies is therefore essential for balancing personalization with safety.*

■ **Multi-Agent Cooperative Systems.** Multi-agent architectures distribute reasoning across specialized agents aligned with clinical roles such as physician, radiologist, or pharmacist (Tang et al., 2024; Kim et al., 2024a; Zhang et al., 2024). Systems such as MDAgents and MoAgent demonstrate how collaborative reasoning, hypothesis testing, and multimodal integration can improve diagnostic accuracy and robustness (Kim et al., 2024b, 2025). Other frameworks model hospital workflows, deep research pipelines, or longitudinal EHR reasoning using chains of interacting agents with shared memory (Liu et al., 2025a; Wang et al., 2025c; Li et al., 2025b; Zeng et al., 2025). Large-scale frameworks such as Agent Hospital and TOOLUNIVERSE further show emergent capabilities from coordinated agent populations and tool ecosystems (Li et al., 2025a; Gao et al., 2026). Empirical evaluations consistently report improved performance on complex diagnostic tasks compared with single-agent systems (Jiang et al., 2025d; Dutta and Hsiao, 2024), partly because structured discussion and critique reduce individual model bias (Ke et al., 2024; Kim et al., 2024c). However, these systems introduce coordination overhead and risks such as communication failure, inconsistent beliefs, or flawed consensus (Maiga et al., 2025; Sun et al., 2025). *Thus, multi-agent designs improve robustness and collaboration but require careful coordination and communication mechanisms.*

■ **Hierarchical and Role-Based Agents.** Hierarchical architectures organize diagnostic workflows using layered coordination, where high-level agents manage specialized sub-agents across diagnostic and therapeutic stages (Zuo et al., 2024; Zhou et al., 2025; Yang et al., 2025a). Systems such as MEGDA illustrate how modular agent roles support structured clinical data analysis and task decomposition (Bani-Harouni et al., 2024). Benchmarks like MedBrowseComp reveal that such architectures improve multi-hop reasoning but still expose substantial gaps in clinical knowledge integration (Chen et al., 2025b). Role separation can also enhance safety by restricting high-risk communication pathways in patient-facing agents (Zhou et al., 2025; Van et al., 2024; Mukherjee et al., 2024). However, poorly designed hierarchies or role assignments may propagate coordination errors and degrade diagnostic outcomes (Kim et al., 2024a). *Adaptive and context-aware hierarchical coordination therefore remains an open research challenge.*

Overall, agent architectures determine how clinical reasoning is coordinated across tasks. Single-agent and planner-executor systems are efficient and interpretable but struggle with complex collaborative reasoning, whereas memory-augmented, multi-agent, and hierarchical designs better support longitudinal context, interdisciplinary expertise, and structured workflows at the cost of higher coordination complexity. Architectural performance depends on alignment with task uncertainty and workflow constraints: search- or debate-based systems are effective for multi-hypothesis reasoning, while linear or single-agent systems are better suited to structured, low-branching tasks due to their effi-

ciency and controllability. Unlike general-domain AI, medical multi-agent systems must also respect established clinical hierarchies and accountability chains involving physicians, radiologists, pharmacists, and specialized teams such as tumor boards or rapid response units, each with distinct consensus and escalation protocols. In clinical settings, multi-agent designs must also preserve accountability across roles such as physicians, radiologists, and pharmacists, with clear procedures for consensus, escalation, and human review.

3.4 Axis III: Tool-Use Capabilities

Tool-use capabilities ground LLM reasoning in external resources. We organize these capabilities into three categories differentiated by their functional role: *external knowledge grounding*, *computational and simulation tools*, *multimodal tools*, and *verification and auditing tools*. The first category is itself divided into three subtypes based on knowledge representation, since retrieval-augmented systems, knowledge graphs, and guideline grounding share the same core challenges—retrieval quality, grounding faithfulness, and update maintenance—and differ primarily in how knowledge is encoded rather than in their functional intent.

■ **External Knowledge Grounding.** RAG, knowledge graphs, and guideline grounding all anchor LLM outputs in external evidence sources independent of model parameters, but share failure modes including stale knowledge, retrieval errors, and grounding-faithfulness failures where conclusions do not follow from retrieved evidence (Sun et al., 2025; Ke et al., 2024).

Subtype 1. Unstructured Representation: Retrieval-Augmented Diagnosis (RAG). RAG retrieves guidelines, literature, or patient records during inference to improve grounding and performance on rare or guideline-driven cases (Huang et al., 2024; Sun et al., 2025). However, accuracy remains highly sensitive to retrieval quality, and irrelevant documents can mislead reasoning (Ke et al., 2024). TC-RAG (Jiang et al., 2025a) further blurs the boundary between retrieval and memory-augmented agents by jointly managing retrieval state and memory.

Subtype 2. Structured Representation: Knowledge Graph Grounding. Knowledge graphs support relational reasoning over comorbidities, drug interactions, and diagnostic hierarchies (Zuo et al., 2024; Tang et al., 2024; Huang et al., 2024), and systems such as AIPatient integrate them with EHR data to simulate clinical scenarios (Yu et al., 2025). Although large medical KGs are difficult to maintain (Jin et al., 2021), hybrid systems such as HyKGE (Jiang et al., 2025b), which combine LLM-generated hypotheses with KG verification, substantially reduce hallucinations compared to pure RAG or KG lookup.

Subtype 3. Normative Representation: Clinical Guideline Grounding. Guideline grounding enforces adherence to protocols such as NCCN or AHA/ACC recommendations, functioning as a normative constraint layer rather than a standard retrieval mechanism (Ke et al., 2024). Systems such as CancerGUIDE (Unell et al., 2025) combine guideline grounding with disagreement estimation to produce safer treatment recommendations with calibrated confidence.

■ **Computational and Simulation Tools.** Computational tools provide quantitative reasoning capabilities such as dosage calculation, risk scoring, and physiological modeling, while simulation environments enable testing of diagnostic strategies before deployment (Li et al., 2025a). These tools can significantly improve reliability but introduce integration complexity and require rigorous validation to ensure clinical realism (Jiang et al., 2025d). For example, GRASP converts Quantitative Systems Pharmacology models into executable reasoning pipelines using graph-based agent coordination (Bazgir et al., 2025). Similarly, zero-shot LLM agents have been used for automated inverse radiotherapy planning, achieving expert-level plan quality without task-specific training (Yang et al., 2025b). *Such tools extend LLM reasoning beyond text but demand careful verification before clinical use.*

■ **Multimodal Tools (Imaging, Pathology, Audio).** Multimodal tools extend agent perception to imaging, pathology slides, and speech signals, enabling integrated clinical analysis across heterogeneous data sources (Guo et al., 2025b; Li et al., 2024; Zhang et al., 2024; Huang et al., 2024). Systems such as multitask multimodal agents combine vision, speech, and retrieval-based reasoning to support interactive clinical decision support (Putra et al., 2025). These capabilities are particularly important for radiology and pathology workflows. However, multimodal fusion increases computational cost and introduces alignment challenges across modalities (Mehandru et al., 2024). *Achieving robust multimodal grounding remains an important research direction for medical agents.*

■ **Verification, Citation, and Auditing Tools.** Verification tools check agent outputs against trusted sources or logical constraints, improving transparency and regulatory compliance (Noori et al., 2026). Citation mechanisms enhance traceability but do not guarantee correctness without independent verification (Sun et al., 2025). Auditing tools provide post-hoc analysis of reasoning processes, which is essential for debugging and safety monitoring in clinical systems (Jiang et al., 2025d). For example, CancerGUIDE integrates disagreement estimation and guideline grounding to produce NCCN-aligned treatment recommendations with calibrated confidence (Unell et al., 2025). *Verification and auditing components are therefore increasingly considered mandatory for safe clinical deployment.*

Overall, tool-use capabilities extend LLM agents beyond pure language reasoning by grounding decisions in external knowledge, structured medical resources, computational models, and multimodal data. Retrieval systems, knowledge graphs, and multimodal tools improve factual accuracy and diagnostic coverage but also introduce dependencies on tool reliability, data quality, and system integration, making verification and auditing essential for trustworthy deployment. Unlike general-domain agents, medical tool use must satisfy normative constraints requiring consistency with evidence-based guidelines, since violations may constitute patient safety risks. Clinical tools must also handle specialized formats such as DICOM and HL7 FHIR while complying with regulatory frameworks governing protected health information, which constrain storage, logging, and data transmission. Tool use grounds LLM reasoning in external resources. We organize these capabilities into four categories by function: external knowledge grounding, computational and simulation tools, multimodal tools, and verification and auditing tools. External knowledge grounding has three subtypes based on how knowledge is represented: retrieval-augmented generation for unstructured sources, knowledge graphs for structured relations, and clinical guidelines for normative requirements.

3.5 Axis IV: Degree of Autonomy

■ **Assistive and Advisory Agents.** Assistive agents provide recommendations while leaving decision-making and actions to clinicians, preserving human authority in the clinical loop. These systems are currently the most common in real-world healthcare deployments because they align well with regulatory requirements and clinical governance structures (Reddy, 2024). However, their effectiveness depends heavily on clinician trust, usability, and integration with existing workflows (Jin et al., 2021). *Thus, assistive agents offer a pragmatic pathway for early clinical adoption but provide limited automation benefits.*

■ **Human-in-the-Loop Systems.** Human-in-the-loop systems explicitly incorporate clinician feedback during reasoning or decision-making to balance safety and automation (Yan et al., 2024). Such designs can improve diagnostic reliability by enabling intervention when model uncertainty or errors arise. However, repeated interaction may increase clinician cognitive load and disrupt clinical workflows (Liu et al., 2025b). *These systems therefore represent a compromise between automation and safety, but their success depends on carefully designed interaction protocols.*

■ **Semi-Autonomous Clinical Agents.** Semi-autonomous agents perform constrained operational tasks such as alert generation, documentation assistance, or structured data entry (Tang et al., 2024). By automating routine activities, they can reduce administrative burden and improve clinical efficiency. Nevertheless, continuous monitoring and safeguards remain necessary to mitigate potential safety risks. *Semi-autonomous agents therefore provide practical efficiency gains while maintaining human oversight for high-stakes decisions.*

■ **Fully Autonomous (Simulated) Hospitals.** Fully autonomous environments simulate complete clinical workflows in which agents independently perform diagnosis, treatment planning, and coordination. Platforms such as AIPatient provide controlled environments for testing large-scale medical agent ecosystems (Yu et al., 2025). While such simulations enable rapid experimentation and training, their translation to real clinical settings remains uncertain due to safety, regulatory, and ethical constraints (Li et al., 2025a; Fan et al., 2024). *Consequently, fully autonomous systems currently serve primarily as research testbeds rather than deployable clinical solutions.*

Overall, the autonomy spectrum reflects trade-offs between safety, efficiency, and clinical responsibility. Lower-autonomy systems prioritize reliability and regulatory feasibility, while higher-autonomy systems promise greater efficiency but introduce substantial safety and governance challenges. Most current clinical deployments therefore adopt hybrid

models that combine automation with structured human oversight. In medicine, autonomy therefore describes not only how many actions an agent can take, but also which decisions require clinician review and who remains accountable for the outcome.

3.6 Axis V: Embodiment

■ **Text-Only Agents.** Text-only agents dominate due to simplicity and scalability within existing LLM frameworks. They rely on textual interaction with clinical data and EHRs, enabling strong knowledge-based reasoning. However, lacking sensory grounding limits their effectiveness for perception-heavy tasks such as imaging or physiological signal interpretation (Jin et al., 2021). *Thus, they are suited for knowledge-centric reasoning but weak in perception-driven clinical settings.*

■ **Multimodal Virtual Agents.** Multimodal agents combine text, imaging, and structured records to enable richer clinical reasoning and improved performance in tasks such as radiology and decision support (Guo et al., 2025b). However, integrating heterogeneous modalities and managing computational overhead remain key challenges. *Nonetheless, multimodal integration is increasingly essential for realistic clinical AI systems.*

■ **Robotic and Physical Agents.** Robotic agents extend LLM reasoning into physical environments for tasks such as surgical assistance, rehabilitation, and monitoring. These systems integrate reasoning with sensing and actuation but require strict safety validation, regulatory approval, and hardware reliability (Bousetouane, 2025). *They represent a promising yet early-stage frontier in medical AI deployment.*

Overall, embodiment shapes how agents perceive and interact with clinical environments. While text-only systems dominate due to scalability, multimodal and embodied agents enable richer interaction and real-world deployment. Clinical embodiment additionally involves safety constraints for vulnerable populations (e.g., pediatric, unconscious, immunocompromised patients) and environmental constraints such as infection control and sensor noise, extending beyond standard robotics. Embodiment (Axis V) describes the agent's native interaction modality with users and its environment. A text-only agent can call an imaging model as a tool while remaining text-only in its own interface; the two properties are recorded independently.

3.7 Cross-Axis Interactions

The five axes represent distinct but interacting design choices. Table 7 summarizes predicted mechanisms and failure modes for selected axis pairs, supported by examples or empirical evidence where available. These interactions make the framework analytical rather than a set of independent labels: a design choice may reduce one failure mode while creating another. Appendix K shows how individual systems occupy positions across multiple axes.

3.8 Evaluation, Safety, and Reliability

Medical agents require system level evaluation beyond accuracy or exact match. Assessment should cover reasoning faithfulness, tool use, grounding, calibration, robustness across populations and settings, expert review, and interactive or longitudinal effects on clinical workflows. Safety failures may arise from unsupported outputs, incorrect tool calls, misapplied retrieval, propagated planning errors, persistent memory, or correlated multi agent errors. Reliable deployment therefore requires calibrated uncertainty, appropriate clinician oversight, and post deployment monitoring.

4 Conclusions and Future Directions

Across the five axes, current medical agents increasingly reproduce the visible structure of diagnostic reasoning, but the evidence does not show that they reproduce its causal and epistemic foundations. **Reasoning:** explicit chains and search can represent hypothesis refinement, but remain vulnerable to correlational errors and poor calibration under distribution shift. **Architecture:** multiple agents can distribute work and critique one another, but coordination does not guarantee independent judgment or clinical accountability. **Tools:** retrieval, guidelines, and computational models can improve grounding, yet tool errors and misapplied evidence can produce grounded hallucinations. **Autonomy:** human oversight remains necessary for high-stakes decisions because benchmark performance does not establish safe independent action. **Embodiment:** multimodal systems broaden the evidence available to an agent, but their performance depends on perception quality and modality alignment. These findings support system-level evaluation that measures causal reasoning, calibration, tool reliability, and clinical outcomes rather than benchmark accuracy alone.

Limitations

As a survey and conceptual synthesis, this work does not introduce new models, algorithms, datasets, or clinical experiments, nor does it provide controlled empirical comparisons or ablation studies. Consequently, quantitative claims about performance, error propagation, or safety are derived from previously published studies rather than generated through new experiments in this manuscript. Direct head-to-head evaluations across paradigms are therefore beyond the scope of this review and would require dedicated experimental studies.

While we aim to provide comprehensive coverage and structured comparison, the rapidly evolving landscape of LLM-based medical agents implies that some recent developments may not be fully represented at the time of writing. Moreover, certain detailed implementation-level optimizations, deployment engineering decisions, and regulatory validation procedures fall outside the scope of this framework, which focuses on conceptual organization and cross-paradigm synthesis rather than system-level certification.

Our framework is intended to organize, compare, and contextualize existing research along orthogonal design axes, rather than to replace empirical validation studies. Any limitations in predictive claims should thus be interpreted as inherent to survey-based research, which synthesizes prior evidence rather than generating new experimental results.

Potential Risks

This work is an analytical survey and does not introduce new models, algorithms, datasets, or clinical decision-making systems. Consequently, it poses no direct risks to patient safety, data privacy, or clinical workflows. Any potential risk of misinterpretation, such as overestimating the readiness of medical agents, is mitigated by our explicit emphasis on limitations, safety concerns, and the gap between benchmark success and real-world reliability. We do not advocate autonomous deployment and consistently frame medical agents as assistive or experimental technologies. By highlighting open challenges, evaluation shortcomings, and regulatory constraints, this work aims to reduce, rather than exacerbate, the risk of premature or inappropriate adoption. Overall, the contribution is informational and risk-neutral, serving to guide safer future research and deployment.

References

Jason Abaluck, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh. 2016. The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review*, 106(12):3730–64.

Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic Arabic medical dialogues using advanced multi-agent LLM techniques. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 11–26, Bangkok, Thailand. Association for Computational Linguistics.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings. *Preprint*, arXiv:1904.03323.

Li Lianjie Anthony, Antonio Bandeira, Niroj Bhandari, Rahul Gorijavolu, Aaron Teo Rui Kang, Jasmine Xiaojin Zhang, Alvin Aung Aung Hein, and March I-Cheng Chen. 2025. Demo: Medbot, a practical tool built by clinicians for clinicians to leverage AI agents to enhance clinical practice. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Felipe Arias Russi, Carolina Salazar Lara, and Ruben Manrique. 2025. A multi-agent framework with diagnostic feedback for iterative plain language summary generation from cochrane medical abstracts. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 87–104, Suzhou, China. Association for Computational Linguistics.

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

Omid Bazgir, Mohammad Jafarnejad, Vineeth Manthapuri, and Ilia Rattsev. 2025. GRASP: Graph reasoning agents for systems pharmacology with human-in-the-loop. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Fouad Bousetouane. 2025. Physical ai agents: Integrating cognitive intelligence with real-world action. *Preprint*, arXiv:2501.08944.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. *Preprint*, arXiv:2012.07805.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2025a. CoD, towards an interpretable medical agent using chain of

diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14345–14368, Vienna, Austria. Association for Computational Linguistics.

Shan Chen, Pedro José Ferreira Moreira, Yuxin Xiao, Samuel Schmidgall, Jeremy L. Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle Bitterman. 2025b. Medbrowsecomp: Benchmarking medical deep research and computer use. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. 2025. The browsergym ecosystem for web agent research. *Preprint*, arXiv:2412.05467.

Avisha Das, Salih Selek, Alia R. Warner, Xu Zuo, Yan Hu, Vipina Kuttichi Keloth, Jianfu Li, W. Jim Zheng, and Hua Xu. 2022. Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 285–297, Dublin, Ireland. Association for Computational Linguistics.

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

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haohao Ying. 2025. Llms can simulate standardized patients via agent coevolution. *Preprint*, arXiv:2412.11716.

Radhika Dua, Young Joon Fred Kwon, Siddhant Dogra, Daniel Freedman, Diana Ruan, Motaz Nashawaty, Danielle Rigau, Daniel Alexander Alber, Kang Zhang, Kyunghyun Cho, and Eric Karl Oermann. 2025. Clinically grounded agent-based report evaluation: An interpretable metric for radiology report generation. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

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

Amalie Dyda, Michael Purcell, Stephanie Curtis, Emma Field, Priyanka Pillai, Kieran Ricardo, Haotian Weng, Jessica C. Moore, Michael Hewett, Graham Williams, and Colleen L. Lau. 2021. Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. *Patterns*, 2(12):100366.

Louis Ehwerhemuepha, Kimberly Carlson, Ryan Moog, Ben Bondurant, Cheryl Akridge, Tatiana Moreno, Gary Gasperino, and William Feaster. 2022. Cerner real-world data (crwd) - a de-identified multicenter electronic health records database. *Data in Brief*, 42:108120.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray.

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

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

Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. 2025. Txagent: An ai agent for therapeutic reasoning across a universe of tools. *Preprint*, arXiv:2503.10970.

Shanghua Gao, Richard Zhu, Pengwei Sui, Zhenglun Kong, Sufian Aldogom, Yepeng Huang, Ayush Noori, Reza Shamji, Krishna Parvataneni, Theodoros Tsiligkaridis, and Marinka Zitnik. 2026. Tooluniverse: An open platform for democratizing ai scientists. *Preprint*, arXiv:2509.23426.

Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel CF Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, Rui Wang, Lei Song, Guanghui Qin, Naoto Usuyama, Cliff Wong, Hao Cheng, HoHin Lee, Praneeth Sanapathi, Sarah Hilado, Tristan Naumann, Javier Alvarez-Valle, Jiang Bian, Mu Wei, Khalil Malik, Lidong Zhou, Jianfeng Gao, Eric Horvitz, Matthew P. Lungren, Doug Burger, Eric Topol, Hoifung Poon, and Paul Vozila. 2025. The illusion of readiness in health ai. *Preprint*, arXiv:2509.18234.

Zihan Guan, Hanyin Wang, Zhongliang Zhou, Qiaohui Zhou, Peining Tao, and Junshui Ma. 2025. Demo: Pharmadata-agent: A specialized agent for pharmaceutical data analysis. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao

Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025a. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning.

Guangfu Guo, Xiaoqian Lu, and Yue Feng. 2025b. MedVRAgent: A framework for medical visual reasoning-enhanced agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18602–18616, Suzhou, China. Association for Computational Linguistics.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. Meddialog: Two large-scale medical dialogue datasets. *Preprint*, arXiv:2004.03329.

Van Hoang, Eoin Rogers, and Robert Ross. 2024. How can client motivational language inform psychotherapy agents? In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 23–40, St. Julians, Malta. Association for Computational Linguistics.

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

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

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

Akshay Jagatap, Srujana Merugu, and Prakash Mandayam Comar. 2025. RxLens: Multi-agent LLM-powered scan and order for pharmacy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 822–832, Albuquerque, New Mexico. Association for Computational Linguistics.

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xinyu Ma, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025a. TC–RAG: Turing–complete RAG's case study on medical LLM systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11400–11426, Vienna, Austria. Association for Computational Linguistics.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025b. HyKGE: A hypothesis knowledge graph enhanced RAG framework for accurate and reliable medical LLMs responses. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11856, Vienna, Austria. Association for Computational Linguistics.

Xun Jiang, Feng Li, Han Zhao, Jiahao Qiu, Jiaying Wang, Jun Shao, Shihao Xu, Shu Zhang, Weiling Chen, Xavier Tang, Yize Chen, Mengyue Wu, Weizhi Ma, Mengdi Wang, and Tianqiao Chen. 2025c. Long term memory: The foundation of ai self-evolution. *Preprint*, arXiv:2410.15665.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025d. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9):AIdbp2500144.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1).

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A. Choquette-Choo, and Zheng Xu. 2024. User inference attacks on large language models. *Preprint*, arXiv:2310.09266.

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

Jun Hyeong Kim, Seokhyun Moon, Seonghwan Kim, Junhyeok Jeon, Taein Kim, Jisu Seo, Songmi Kim, and Woo Youn Kim. 2025. Moagent: A hypothesis-driven multi-agent framework for drug mechanism of action discovery. In *NeurIPS 2025 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*.

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

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

Yubin Kim, Chanwoo Park, Hyewon Jeong, Cristina Grau-Vilchez, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Cynthia Breazeal, and Hae Won Park. 2024c. A demonstration of adaptive collaboration of large language models for medical decision-making. *Preprint*, arXiv:2411.00248.

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

Moustafa Laymouna, Yuanchao Ma, David Lessard, Tibor Schuster, Kim Engler, and Bertrand Lebouché. 2024. Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *J Med Internet Res*, 26:e56930.

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

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2025a. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Preprint*, arXiv:2405.02957.

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

Yuan Li, Matthew Pan, Claire Liu, Hengjin Zhu, Xijing Wang, and Yingtao Luo. 2025b. Demo: Orchestrating large language model agents and resources for medical deep research. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Benjamin Liu, Sejong Kim, Drona Thoka, Varun Puttagunta, Kaylin Sheng, Mark Li, Kiran Nijjer, Adnan Ahmed, Thi Uyen Hanh Le, Sai Chidvilas Gudiboina, Ali Ugur, and Kevin Zhu. 2025a. Multi-turn LLM systems for diagnostic decision-making: Considerations, biases, and challenges. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Benjamin Liu, Dillon Mehta, Rishi Malhotra, Adam Zobian, Yong Ying Tan, Samir Chopra, Daniella Rand, Natalie Pang, Abhiram Gudimella, Raghav Thallapragada, Derek Jiu, Prisha Shah, and Kevin Zhu. 2025b. Examining the vulnerability of multi-agent medical systems to human interventions for clinical reasoning. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xin Liu, Zhengyang Wang, Xianfeng Tang, Shiyang Li, Xiang He, Ruijie Wang, Bing Yin, Xiao Gu, Lei Clifton, and David A. Clifton. 2025c. DrAgent: Empowering large language models as medical agents for multi-hop medical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15656–15668, Suzhou, China. Association for Computational Linguistics.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang SU, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. 2025d. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. *Preprint*, arXiv:2412.01605.

Yungeng Liu, Zan Chen, Yu Guang Wang, and Yiqing Shen. 2024. Autoproteinengine: A large language model driven agent framework for multimodal automl in protein engineering. *Preprint*, arXiv:2411.04440.

Junjie Luo, Yihong Guo, Anqi Liu, Ritu Agarwal, and Guodong Gordon Gao. 2025. PAME-AI: Patient messaging creation and optimization using agentic AI. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Abdine Maiga, Anoop Shah, and Emine Yilmaz. 2025. Error detection in medical note through multi agent debate. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 124–135, Viena, Austria. Association for Computational Linguistics.

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

Andreas Geert Motzfeldt, Joakim Edin, Casper L. Christensen, Christian Hardmeier, Lars Maaløe, and Anna Rogers. 2025. Code like humans: A multi-agent solution for medical coding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22612–22627, Suzhou, China. Association for Computational Linguistics.

Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, Cezanne Bianco, Swapnil Sharma, Rae Lasko, Michelle Voisard, Sanchay Harneja, Darya Filippova, Gerry Meixiong, Kevin Cha, Amir Youssefi, Meyhaa Buvanesh, Howard Weingram, Sebastian Bierman-Lytle, Harpreet Singh Mangat, Kim Parikh, Saad Godil, and Alex Miller. 2024. Polaris: A safety-focused llm constellation architecture for healthcare. *Preprint*, arXiv:2403.13313.

Ayush Noori, Aaron E. Boussina, Hai Ho Bich, James Anibal, Julia Maslinski, Manuel Burger, Martin Faltys, Adam Rodman, Alan Karthikesalingam, Alessandro Blasimme, Annelia Itwaru, Ben Kaplan, Bilal A. Mateen, Christopher A. Longhurst, Daniel Yang, Dave deBronkart, Effy Vayena, Fedor Sergeev, Gauden Galea, Ha Thi Hai Duong, Harold F. Wolf III, Jacob Waxman, Joerg C. Schefold, Joshua C. Mandel, Juliana Rotich, Kenneth D. Mandl, Lily Poursoltan, Maryam Mustafa, Melissa Miles, Nigam H. Shah, Noa Dagan, Pavan Bodanki, Peter Lee, Philipp Koralus, Prathamesh Parchure, Prem Timsina, Ran D. Balicer, Robert Korom, Scott Mahoney, Seth Hain, Tien Yin Wong, Trevor Mundel, Vivek Natarajan, Ankit Sakhuja, Benjamin Glicksberg, C. Louise Thwaites, Gunnar Rätsch, Karandeep Singh, David A. Clifton, Isaac S. Kohane, and Marinka Zitnik. 2026. A global log for medical ai. *Preprint*, arXiv:2510.04033.

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. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Himanshu Gautam Pandey, Akhil Amod, and Shivang Kumar. 2024. Advancing healthcare automation: Multi-agent system for medical necessity justification. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 39–49, Bangkok, Thailand. Association for Computational Linguistics.

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

Muhamad Iqbal Januadi Putra, Raka Admiral Abdurrahman, and Vincent Alexander. 2025. Multitask asynchronous bidirectional multimodal agent for personalized treatment companions. In *NeurIPS 2025 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*.

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

Kirk Roberts, Yuqi Si, Anshul Gandhi, and Elmer V. Bernstam. 2018. A framenet for cancer information in clinical narratives: Schema and annotation.

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

Stuart Russell and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, United States.

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

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2025. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Joshua Strong, Qianhui Men, and Alison Noble. 2025. Trustworthy and practical ai for healthcare: A guided deferral system with large language models. *Preprint*, arXiv:2406.07212.

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

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, and Yu Rong. 2025. ReasonMed: A 370K multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26446–26467, Suzhou, China. Association for Computational Linguistics.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. 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. Association for Computational Linguistics.

Shu-Chiao Tsai, Hooman Aghaebrahimi Samani, Yu-Wei Kao, Ke-Ning Zhu, and Brian Jalaian. 2018. Design and development of interactive intelligent medical agent. *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 210–215.

Alyssa Unell, Noel C Codella, J. Samuel Preston, Peniel Argaw, Wen wai Yim, Zelalem Gero, Cliff Wong, Eric Horvitz, Amanda Hall, Rachel Ruican Zhong, Jiachen Li, Shrey Jain, Mu Wei, Matthew P. Lungren, and Hoifung Poon. 2025. CancerGUIDE: Cancer guideline understanding via internal disagreement estimation. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

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

Jelte van Waterschoot, Iris Hendrickx, Arif Khan, Esther Klabbers, Marcel de Korte, Helmer Strik, Catia Cucchiarini, and Mariët Theune. 2020. Bliss: An agent for collecting spoken dialogue data about health and well-being. In *International Conference on Language Resources and Evaluation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Junda Wang, Zonghai Yao, Lingxi Li, Junhui Qian, Zhichao Yang, and hong yu. 2025a. Chatthero: A language agent for recovery support. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025b. A survey of LLM-based agents in medicine: How far are we from baymax? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359, Vienna, Austria. Association for Computational Linguistics.

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. 2025c. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. page 2250–2261.

Azmine Toushik Wasi and Md. Iqramul Hoque. 2025. Position: Adjacent technologies are the key enablers of scalable and safe clinical MLLM deployment. In *NeurIPS 2025 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*.

Azmine Toushik Wasi and Mst Rafia Islam. 2025. Perspective: Lessons from cybersecurity for biological AI safety and regulation. In *NeurIPS 2025 Workshop on Regulatable ML*.

Azmine Toushik Wasi, Mst Rafia Islam, and Rahatun Nesa Priti. 2025. Position: Without global governance, AI-enabled biodesign tools risk dangerous proliferation. In *NeurIPS 2025 Workshop on Biosecurity Safeguards for Generative AI*.

Azmine Toushik Wasi, Mahir Absar Khan, Rahatun Nesa Priti, Abdur Rahman, Mst Rafia Islam, and Miskat Tahmin Ali. 2026. Clinician perspectives on generative ai integration in bangladesh's healthcare workflows: A brief study. In *Proceedings of the 2026 ACM Interactive Health Conference*, IH '26, New York, NY, USA. Association for Computing Machinery.

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

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

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

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

Dingkang Yang, Jinjie Wei, Mingcheng Li, Jiyao Liu, Lihao Liu, Ming Hu, Junjun He, Yakun Ju, Wei Zhou, Yang Liu, and Lihua Zhang. 2025a. Medaide: Information fusion and anatomy of medical intents via llm-based agent collaboration. *Preprint*, arXiv:2410.12532.

Dongrong Yang, Xin Wu, Yibo Xie, Xinyi Li, Qiuwen Wu, Jackie Wu, and Yang Sheng. 2025b. Zero-shot

large language model agents for fully automated radiotherapy treatment planning. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. 2025c. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *Preprint*, arXiv:2502.09560.

Yang Yang, Hongxing Huo, Jingchi Jiang, Xuemei Sun, Yi Guan, Xitong Guo, Xiang Wan, and Shengping Liu. 2021. Clinical decision-making framework against over-testing based on modeling implicit evaluation criteria. *Journal of Biomedical Informatics*, 119:103823.

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

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

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Jingxian He, Wenyue Hua, Mingyu Jin, Guang Chen, Yang Zhou, Zhao Li, Trisha Gupte, Ming-Li Chen, Zahra Azizi, Qi Dou, Bryan P. Yan, Yongfeng Zhang, Yanqiu Xing, Themistocles L. Danielle S. Bitterman, Themistocles L. Assimes, Xin Ma, Lin Lu, and Lizhou Fan. 2025. Simulated patient systems powered by large language model-based ai agents offer potential for transforming medical education. *Preprint*, arXiv:2409.18924.

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

Sihang Zeng, Yujuan Fu, Sitong Zhou, Zixuan Yu, Lucas Jing Liu, Jun Wen, Matthew Thompson, Ruth Etzioni, and Meliha Yetisgen. 2025. Traj-coa: Patient trajectory modeling via chain-of-agents for lung cancer risk prediction. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.

Luyao Zhang, Jianhua Shu, Jili Hu, Fangfang Li, Junjun He, Peng Wang, and Yiqing Shen. 2024. Exploring the potential of large language models in radiological imaging systems: Improving user interface design and functional capabilities. *Electronics*, 13(11):2002.

Yucheng Zhou, Lingran Song, and Jianbing Shen. 2025. MAM: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25319–25333, Vienna, Austria. Association for Computational Linguistics.

Yinghao Zhu, Yifan Qi, Zixiang Wang, Lei Gu, Dehao Sui, Haoran Hu, Xichen Zhang, Ziyi He, Junjun He, Liantao Ma, and Lequan Yu. 2025. Healthflow: A self-evolving ai agent with meta planning for autonomous healthcare research. *Preprint*, arXiv:2508.02621.

Yangyang Zhuang, Wenjia Jiang, Jia-Yu Zhang, Ze Yang, Joey Tianyi Zhou, and Chi Zhang. 2025. Learning to be a doctor: Searching for effective medical agent architectures. *Proceedings of the 33rd ACM International Conference on Multimedia*.

Micaela Maria Zucchelli, Nicola Matteucci Armandi Avogli Trotti, Andrea Pavan, Laura Piccardi, and Raffaella Nori. 2025. The dual process model: the effect of cognitive load on the ascription of intentionality. *Frontiers in Psychology*, Volume 16 - 2025.

Kai-Wen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2024. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis.

Kaiwen Zuo and Yirui Jiang. 2025. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *Preprint*, arXiv:2412.18947.

A Review Methodology

A.1 Review Objective

This review follows a structured methodology designed to support broad coverage and transparent synthesis of research on LLM-based medical agents. Because this work is a survey rather than an empirical study, methodological validity is defined in terms of systematic search procedures, explicit inclusion criteria, cross-paradigm classification, and evidence-aware interpretation. The objective is not to generate new experimental results, but to organize and critically synthesize existing findings within a unified conceptual framework. All methodological decisions were guided by the aim of maximizing coverage across reasoning paradigms, agent architectures, tool-use capabilities, autonomy levels, and embodiment modalities.

A.2 Literature Identification

Searches were conducted from January through April 2025 in arXiv, PubMed, IEEE Xplore, and the ACM Digital Library, and in the proceedings of NeurIPS, ICML, ICLR, ACL, EMNLP, and AAAI. The query combined Group A (medical agent OR clinical agent OR diagnostic agent OR LLM agent healthcare) with Group B (reasoning OR tool use OR multimodal OR safety OR hallucination OR uncertainty OR benchmark OR deployment). We also used backward and forward citation tracking from seed papers. The complete set of included studies is listed in Supplementary Table 3.

A.3 Study Selection and PRISMA-Style Flow Summary

To ensure transparency and reproducibility, we followed a structured multi-stage screening process inspired by PRISMA reporting standards. The procedure included identification of records across databases, removal of duplicates, title and abstract screening, full-text assessment, and final inclusion based on predefined eligibility criteria. Study selection followed explicit inclusion and exclusion principles to maintain methodological consistency. Included works had to propose, evaluate, or systematically analyze LLM-based medical agents with relevance to diagnostic reasoning, tool integration, safety mechanisms, autonomy design, or multimodal interaction. Empirical studies were prioritized when available, particularly those reporting benchmark evaluations, retrospective analyses, or real-world deployment outcomes, while conceptual frameworks were included when they introduced novel architectural or reasoning paradigms that shaped subsequent research. Table 4 summarizes the number of records at each stage of the review process, and Table 5 details the eligibility criteria applied during screening.

A.4 Literature Mapping

Following identification and screening, we mapped each included study onto five design axes: reasoning paradigms, agent architectures, tool use, degree of autonomy, and embodiment. These axes describe conceptually distinct design choices, but they are not mutually exclusive categories or statistically independent variables. Each system can take a value on every axis. Learning and adaptation, discussed in Appendix D, is treated as a cross-cutting concern that can modify systems along multiple axes. Evaluation, safety, and deployment are also analyzed as cross-cutting concerns. Table 6 summarizes the five axes.

A.5 Evidence Synthesis

Evidence synthesis was conducted using structured qualitative comparison rather than quantitative meta-analysis, reflecting the heterogeneity of tasks, datasets, and evaluation protocols across studies. Each work was interpreted according to its empirical validation level, distinguishing between controlled benchmarks, simulation-based evaluations, retrospective analyses, and real-world clinical deployments. This stratification allowed systematic assessment of evidence strength while preserving methodological diversity. Claims regarding robustness, safety, or generalization were interpreted in relation to the reported evaluation context rather than extrapolated beyond it. This evidence-aware approach ensures that comparisons remain grounded in study design characteristics.

A.6 Bias Assessment

Bias assessment focused on recurring limitations in the literature, including hallucination susceptibility, error propagation in multi-step reasoning, overconfidence under distribution shift, and challenges in external validation. Particular attention was given to differences between static benchmark performance and longitudinal or interactive clinical scenarios. Where relevant, discrepancies between simulated environments and real-world deployment were highlighted to prevent overinterpretation of benchmark gains. The methodological

Table 1: A unified framework of LLM-based medical agents organized along six orthogonal axes. The taxonomy highlights design trade-offs across reasoning, architecture, tool use, autonomy, embodiment, and learning, enabling systematic comparison of heterogeneous systems.

Axis	Section	Subsection	Related Works
Reasoning Paradigms	Implicit and Explicit Reasoning	Implicit Reasoning	(Gu et al., 2025; Jin et al., 2021; Huang et al., 2025a; Zuo and Jiang, 2025)
		Explicit Reasoning	(Chen et al., 2025a; Sun et al., 2025)
	Chain-of-Thought and Chain-of-Diagnosis	Chain-of-Thought (CoT)	(Yao et al., 2023a)
		Chain-of-Diagnosis (CoD)	(Chen et al., 2025a; Sun et al., 2025)
	Tree-, Graph-, and Search-Based Reasoning	Tree-Based Reasoning	(Wei et al., 2023; Guo et al., 2025b; Yang et al., 2021; Sun et al., 2025)
		Graph-Based Reasoning	(Zuo et al., 2024; Sun et al., 2025)
	Symbolic, Programmatic, and Neuro-Symbolic Reasoning	Symbolic Reasoning	(Tang et al., 2024; Zuo et al., 2024; Ke et al., 2024; Jin et al., 2021)
		Neuro-Symbolic Reasoning	(Liu et al., 2025d; Huang et al., 2024)
	Reflective, Self-Corrective, and Debate-Based Reasoning	Reflective and Self-Corrective Reasoning	(Yao et al., 2023b; Yue et al., 2024)
		Debate-Based Reasoning	(Maiga et al., 2025; Zuo et al., 2024; Ke et al., 2024; Kim et al., 2024b; Gao et al., 2026; Sun et al., 2025)
Agent Architectures	Single-Agent Systems	-	(Jiang et al., 2025d; Jin et al., 2021; Fan et al., 2024; Gu et al., 2025)
	Planner–Executor Agents	-	(Mehandru et al., 2024; Liu et al., 2025c; Jiang et al., 2025d; Gao et al., 2025; Guan et al., 2025; Sun et al., 2025)
	Memory-Augmented Agents	-	(Li et al., 2025a; Du et al., 2025; Ehwerhemuepha et al., 2022; Jin et al., 2021; Kandpal et al., 2024)
	Multi-Agent Cooperative Systems	Role-Based Specialized Agents	(Tang et al., 2024; Kim et al., 2024a; Zhang et al., 2024)
		Adaptive Collaboration	(Kim et al., 2024b)
		Hypothesis-Driven Scientific Process	(Kim et al., 2025)
		Multi-Turn Interaction	(Liu et al., 2025a)
		Hospital Workflow	(Wang et al., 2025c)
		Deep Research	(Li et al., 2025b)
		Chain-of-Agents with Shared Memory (EHR Trajectories)	(Zeng et al., 2025)
		Large-Scale Agent Frameworks	(Gao et al., 2026)
		Performance and Benchmarking	(Jiang et al., 2025d; Dutta and Hsiao, 2024)
		Cooperative Reasoning	(Ke et al., 2024; Kim et al., 2024c)
		Other	(Maiga et al., 2025; Sun et al., 2025; Yue et al., 2024)
	Hierarchical and Role-Based Agents	-	(Zuo et al., 2024; Zhou et al., 2025; Yang et al., 2025a; Bani-Harouni et al., 2024; Chen et al., 2025b; Zhou et al., 2025; Van et al., 2024; Mukherjee et al., 2024)
Tool-Use Capabilities	Retrieval-Augmented Diagnosis (RAG)	-	(Huang et al., 2024; Sun et al., 2025; Ke et al., 2024)
	Knowledge Graph and Guideline Grounding	Knowledge Graph Grounding	(Zuo et al., 2024; Tang et al., 2024; Huang et al., 2024; Yu et al., 2025)
		Guideline Grounding	(Ke et al., 2024; Jin et al., 2021)
	Computational and Simulation Tools	-	(Li et al., 2025a; Jiang et al., 2025d; Bazgir et al., 2025; Yang et al., 2025b)
	Multimodal Tools (Imaging, Pathology, Audio)	-	(Guo et al., 2025b; Li et al., 2024; Zhang et al., 2024; Huang et al., 2024; Putra et al., 2025; Mehandru et al., 2024)
	Verification, Citation, and Auditing Tools	-	(Noori et al., 2026; Sun et al., 2025; Jiang et al., 2025d; Unell et al., 2025)
Degree of Autonomy	Assistive and Advisory Agents	-	(Reddy, 2024; Jin et al., 2021)
	Human-in-the-Loop Systems	-	(Yan et al., 2024; Liu et al., 2025b)
	Semi-Autonomous Clinical Agents	-	(Tang et al., 2024)
	Fully Autonomous (Simulated) Hospitals	-	(Yu et al., 2025; Li et al., 2025a; Fan et al., 2024)
Embodiment	Text-Only Agents	-	(Jin et al., 2021)
	Multimodal Virtual Agents	-	(Guo et al., 2025b)
	Robotic and Physical Agents	-	(Bousetouane, 2025)

Table 2: Diagnostic Tasks and Applications of LLM-Based Medical Agents

Task	Description	References
Diagnostic Decision Support	Mature application evolving from QA to agentic frameworks integrating longitudinal context, tool use, and hypothesis tracking. Outperforms single-pass models in complex cases but amplifies errors if uncritically accepted. Mostly assistive with regulatory limits.	(Jin et al., 2021; Jiang et al., 2025d; Liu et al., 2025c; Tang et al., 2024; Zeng et al., 2025; Unell et al., 2025; Huang et al., 2025a; Reddy, 2024; Yan et al., 2024)
Differential Diagnosis and Hypothesis Ranking	Enumerates and ranks multiple plausible conditions, outperforming traditional classifiers. Multi-agent frameworks enhance complex benchmark performance. Challenges include reduced precision and poor calibration without priors.	(Chen et al., 2025a; Guo et al., 2025b; Wei et al., 2023; Kim et al., 2024b; Jiang et al., 2025d; Fan et al., 2024; Gu et al., 2025)
Medication Safety and Treatment Planning	Drug interaction checks, dosage calculation, and planning augmented by knowledge graphs and guideline retrieval. Reflective verification improves accuracy, but pharmacological errors persist. Autonomy is limited; human confirmation required.	(Zuo et al., 2024; Tang et al., 2024; Ke et al., 2024; Liu et al., 2025d; Jin et al., 2021; Reddy, 2024; Yan et al., 2024)
Patient-Facing Conversational Agents	Enhance triage, education, and symptom assessment with natural language. Improve accessibility but risk unsafe reassurance and delayed escalation. Tool-augmented agents with escalation outperform free-form chatbots. Mostly low-risk informational deployment.	(Roberts et al., 2018; Jin et al., 2021; Laymouna et al., 2024; Wang et al., 2025a; Huang et al., 2025a; He et al., 2020; van Waterschoot et al., 2020; Das et al., 2022; Sun et al., 2025; Reddy, 2024; Yan et al., 2024; Zuo and Jiang, 2025)
Clinical Workflow Automation	Automates documentation, triage routing, test recommendations, reducing clinician workload. Task-specific agents address variability and summaries; frameworks automate authorization and medication reconciliation. Errors may cascade, requiring verification and audit.	(Tang et al., 2024; Jiang et al., 2025d; Kim et al., 2024a; Roberts et al., 2018; Li et al., 2025b; Lee et al., 2024; Sudarshan et al., 2024; Pandey et al., 2024; Jagatap et al., 2025; Jin et al., 2021; Noori et al., 2026; Motzfeldt et al., 2025; Arias Russi et al., 2025)
Training, Simulation, and Medical Education	Supports clinician training with scalable, repeatable complex case exposure. Acts as standardized patients or tutors for engagement and feedback. Enables experimentation but limited fidelity restricts transferability. Multi-agent copilots enhance interactive learning.	(Li et al., 2025a; Fan et al., 2024; Laymouna et al., 2024; Sun et al., 2025; Jiang et al., 2025d; Mehandru et al., 2024; Jin et al., 2021; Wei et al., 2024)
Remote Monitoring and Early Warning Systems	Analyze streaming physiological and self-reported data for early deterioration detection. Integrates multimodal data; improves sensitivity over rule-based systems but suffers from false positives. Calibration and constrained autonomy critical.	(Tang et al., 2024; Li et al., 2024; Jiang et al., 2025d; Sun et al., 2025; Jin et al., 2021; Huang et al., 2025a)

Table 3: Primary Literature Sources Considered in This Review

Source Category	Examples
Preprint Repositories	arXiv, bioRxiv, medRxiv
Biomedical Databases	PubMed, MEDLINE
Computer Science Databases	IEEE Xplore, ACM Digital Library
Conference Proceedings	NeurIPS, ICML, ICLR, ACL, EMNLP, AAAI, KDD
Clinical and Medical Journals	Nature Medicine, The Lancet Digital Health, JAMIA, NEJM AI, BMJ, JAMA, NPJ Digital Medicine, Lancet AI
Interdisciplinary AI Journals	Nature Machine Intelligence, IEEE Transactions on Medical Imaging, Artificial Intelligence in Medicine

Table 4: PRISMA-style flow summary of study selection.

Stage	**Records**
Records identified through database searching	256
Additional records identified through other sources	313
Total records before deduplication	569
Duplicates removed	291
Records screened by title and abstract	278
Records excluded after screening	131
Full-text articles assessed for eligibility	147
Full-text articles excluded with reasons	60
Studies included in qualitative synthesis	**87**

synthesis therefore emphasizes structural patterns across paradigms rather than isolated performance rankings. This approach supports a systems-level understanding of reliability and failure modes in medical agent design.

Finally, this review prioritizes transparency and extensibility. The taxonomy is designed to accommodate emerging reasoning paradigms, architectural innovations, and regulatory frameworks without structural modification. Although comprehensive, the rapidly evolving nature of the field implies that some recent developments may not be fully captured at the time of writing. The methodology reflects established standards for high-level scientific review, focusing on systematic organization, comparative analysis, and conceptual synthesis rather than primary experimentation. This design enables the framework to function as a stable reference model for future research in LLM-based medical agents.

B Clinical Diagnostic Reasoning

This section summarizes foundational theories that underlie medical decision-making and diagnostic inference in clinical practice. We outline classical reasoning paradigms, including hypothetico-deductive, probabilistic, and dual-process models, which inform both human expertise and computational diagnostic systems. These frameworks provide the conceptual basis for evaluating how LLM-based agents structure inference, manage uncertainty, and mitigate error. By grounding our taxonomy in established clinical reasoning theory, we ensure alignment between agent design principles and normative standards of medical cognition. Understanding these foundations is essential for interpreting the strengths and limitations of modern agentic diagnostic systems.

B.1 Hypothetico-Deductive Reasoning

Hypothetico-deductive reasoning models clinical diagnosis as an iterative cycle of hypothesis generation, evidence collection, and hypothesis refinement. Given a hypothesis set $\mathcal{H} = \{h_1, \ldots, h_n\}$ and clinical evidence E, clinicians aim to maximize explanatory coverage while minimizing unnecessary testing and cognitive effort (Abaluck et al., 2016). This process emphasizes active information seeking, where each diagnostic test is selected for its ability to discriminate among competing hypotheses rather than to confirm a single favored explanation (Jin et al., 2021; Chen et al., 2025a). Expert performance is characterized by efficient hypothesis pruning, strategic test ordering, and adaptive updating under uncertainty, rather than by memorization of disease patterns (Ke et al., 2024). Modern LLM-based medical agents increasingly operationalize this paradigm by structuring reasoning as explicit iterative refinement, often implemented through search-based or tree-structured inference mechanisms (Wei et al., 2023; Li et al., 2024). This alignment between classical clinical methodology and agentic search architectures provides a conceptual bridge between human diagnostic strategy and computational reasoning systems.

B.2 Bayesian and Probabilistic Diagnosis

Bayesian diagnostic reasoning formalizes disease inference as probabilistic updating of latent variables using observed evidence. For a disease variable D and evidence E, inference follows Bayes' rule, expressed as $P(D \mid E) = \frac{P(E|D)P(D)}{P(E)}$. This

Criterion	Include	Exclude
System type	Proposes, evaluates, or systematically analyzes an LLM-based medical agent.	No LLM component; purely rule-based or traditional machine-learning system.
Relevance	Addresses diagnostic reasoning, tool integration, safety, autonomy, or multimodal interaction in medicine.	Non-medical domain, or a general-purpose agent with no clinical application.
Depth	Full paper of at least four pages with a system description or evaluation.	Workshop abstract under four pages without evaluation.
Evidence	Reports benchmark evaluation, retrospective analysis, simulation, or deployment outcomes. Include conceptual work only when it describes a relevant system or paradigm.	Purely theoretical work without a system description or evaluation.

Table 5: Eligibility criteria used to select studies for the review. Confirm that these criteria match the screening procedure actually applied.

Design axis	Comparative focus
I. Reasoning	Implicit and explicit inference; linear, search-based, symbolic, and reflective strategies.
II. Architecture	Single-agent, planner-executor, memory-augmented, hierarchical, and multi-agent systems.
III. Tool use	External knowledge grounding, computation and simulation, multimodal tools, and verification and auditing.
IV. Autonomy	Assistive and advisory systems through bounded human-in-the-loop and simulated autonomous systems.
V. Embodiment	Text-only, multimodal virtual, and robotic or physical interaction.

Table 6: Five design axes used to describe LLM-based medical agents. Learning and adaptation, evaluation, safety, and deployment are treated as cross-cutting concerns.

formulation explicitly integrates prior probabilities with likelihood functions to produce posterior disease estimates. Classical probabilistic graphical models and decision-theoretic approaches provided principled frameworks for uncertainty modeling but were historically limited by computational scalability and data requirements in complex clinical environments (Jin et al., 2021; Pollard et al., 2018). Contemporary LLM-based agents approximate Bayesian reasoning either implicitly, through internally represented confidence distributions, or explicitly, by outputting calibrated disease probabilities alongside explanations (Chen et al., 2025a; Sun et al., 2025). Despite these advances, calibration remains a persistent challenge, with many models exhibiting systematic overconfidence and misaligned posterior estimates, particularly under distribution shift (Huang et al., 2025a; Gu et al., 2025). Improving probabilistic grounding and uncertainty quantification is therefore central to aligning AI-driven diagnosis with formal statistical decision theory.

B.3 Dual-Process Theory in Medicine

Dual-process theory conceptualizes clinical reasoning as the interaction between fast, intuitive processes (System 1) and slower, deliberative processes (System 2) (Zucchelli et al., 2025). System 1 enables rapid pattern recognition based on experience and similarity matching, supporting efficient preliminary assessments in time-sensitive contexts. System 2 provides analytical oversight, enabling structured hypothesis testing, counterfactual reasoning, and correction of intuitive errors. Diagnostic errors frequently occur when heuristic-driven System 1 judgments dominate without sufficient System 2 verification, or when analytical processes fail to override biased initial impressions. LLM-based systems exhibit analogous dynamics, where next-token prediction resembles fast associative inference, while structured reasoning mechanisms emulate deliberative multi-step analysis (Chen et al., 2025a). Reflective and iterative agent designs further operationalize this duality by alternating between rapid generation and verification phases, thereby approximating cognitive switching between intuitive and analytical modes (Yao et al., 2023b; Yue et al., 2024). This correspondence provides a theoretical basis for designing agents that balance efficiency with structured self-evaluation.

B.4 Failure Modes and Cognitive Biases

Clinical reasoning is susceptible to systematic cognitive biases, including anchoring, availability bias, premature closure, and confirmation bias, which can lead to persistent diagnostic errors (Jin et al., 2021). These biases reflect predictable deviations

Table 7: Predicted interactions between pairs of design axes. Mechanisms are analytical predictions; cited studies provide examples or empirical support where available.

Axis pair	Predicted constraint or failure mode	Mechanism and evidence
I × II	Linear reasoning in a single-agent architecture has limited error recovery; debate requires multiple agents.	Each step in a linear chain conditions later steps, while another agent can challenge an intermediate claim. CoD and MDAgents provide representative examples (Chen et al., 2025a; Kim et al., 2024a).
I × III	Retrieval can ground explicit reasoning but adds dependence on retrieval quality; implicit reasoning without external checks can hallucinate.	A relevant source can constrain unsupported claims, while a poor or misapplied source can support an incorrect conclusion (Sun et al., 2025; Ke et al., 2024).
I × IV	Debate with bounded human oversight allows correction; full autonomy removes that fallback.	Moderator review can constrain consensus (Kim et al., 2024a).
I × V	Text-only systems cannot reason over unobserved images or signals; multimodal inputs introduce alignment errors.	The available input modality limits what evidence can enter the reasoning process. Use simulated text-only systems as examples only where the source supports this characterization.
II × III	Memory and retrieval can make evidence provenance unclear.	Joint management of persistent memory and retrieved state can blur whether a conclusion came from the patient record, an external source, or an earlier model output.
II × IV	Full autonomy in multi-agent systems increases coordination risk without human review.	Communication failures and conflicting beliefs can propagate when no clinician can intervene. (Maiga et al., 2025; Sun et al., 2025).
II × V	Physical actions can propagate coordination errors across agent layers.	Actuation adds safety and rollback constraints beyond those of text-only interaction.
III × IV	Autonomous tool calls increase exposure to prompt injection and data leakage.	Tool calls cross system boundaries, allowing incorrect or adversarial outputs to propagate without review (Ke et al., 2024; Reddy, 2024).
III × V	End-to-end performance can be limited by the weakest multimodal tool.	A perception or segmentation error can affect downstream reasoning. This is a structural prediction about tool dependence.
IV × V	Greater autonomy raises safety demands when an agent can act physically.	Physical actions can have immediate consequences that are harder to reverse than text recommendations.

from normative probabilistic reasoning and often arise under time pressure, uncertainty, or information overload. LLM-based diagnostic systems demonstrate structurally similar failure patterns, including hallucinated content, spurious correlations, and overconfident outputs that are weakly grounded in evidence (Huang et al., 2025a; Zuo and Jiang, 2025). Such behaviors can propagate across multi-step reasoning pipelines, particularly in autonomous or multi-agent configurations without adequate verification mechanisms. Reflective reasoning, structured debate, and cross-model critique have been proposed as mitigation strategies, promoting explicit error detection and hypothesis re-evaluation (Maiga et al., 2025; Ke et al., 2024). Understanding these parallels between human cognitive vulnerabilities and model-level failure modes is essential to prevent amplification of bias when scaling diagnostic AI systems in clinical settings.

C Prior Work and Research Landscape

Several recent surveys have examined the emergence of LLMs and LLM-based agents in healthcare, reflecting growing interest in agentic medical AI. Early reviews primarily focused on medical language models as task-specific predictors or question-answering systems, emphasizing benchmark performance on clinical text understanding, exam-style QA, and report generation (Jin et al., 2021; Roberts et al., 2018). While these works established foundational capabilities, they largely treated diagnosis as a static inference problem and offered limited discussion of reasoning structure, tool use, or longitudinal decision-making.

More recent surveys have expanded scope to agent-based systems. Notably, Wang et al. (2025b) provide a broad overview of LLM-based agents in medicine, covering system architectures, clinical planning, medical reasoning frameworks, and external capacity enhancement. Their survey comprehensively catalogs applications such as clinical decision support, documentation, training simu-

lations, and healthcare service optimization, and highlights key challenges including hallucination, multimodal integration, implementation barriers, and ethical concerns. This work represents an important step toward unifying agent architectures and applications within healthcare. However, existing surveys, including Wang et al. (2025b), primarily adopt a *descriptive* perspective, organizing the field around system components and use cases without deeply interrogating the nature of diagnostic reasoning itself. In particular, prior reviews rarely distinguish between implicit pattern matching and explicit clinical reasoning, nor do they systematically analyze how different agentic design choices influence failure modes, safety, or clinical reliability. Evaluation is often discussed at a high level, with limited emphasis on reasoning faithfulness, error propagation, or workflow-level impact. Several concrete agent frameworks and domain applications, AutoPE for protein engineering, MedRAX for multimodal radiology, and RxLens for pharmacy workflows, illustrate the breadth of agentic solutions spanning bench science to bedside automation (Liu et al., 2024; Fallahpour et al., 2025; Jagatap et al., 2025).

In contrast, this survey adopts a *reasoning-centric and diagnostic-first perspective*. We position diagnostic reasoning, not model architecture or task category, as the primary organizing principle.Our taxonomy organizes LLM-based medical agents along five design axes: reasoning paradigms, agent architectures, tool use, autonomy, and embodiment. Learning and adaptation is discussed separately as a cross-cutting concern, while evaluation, safety, and deployment are analyzed across the framework. This framework enables direct comparison across systems that may appear similar at the application level but differ substantially in their diagnostic assumptions and risk profiles. Furthermore, while prior surveys typically treat safety, ethics, and deployment as peripheral considerations, our work integrates these aspects into the core analysis. We emphasize how hallucinations, overconfidence, bias, and error propagation emerge differently across reasoning and agentic paradigms, and why many failures are systemic rather than model-specific. Finally, we critically examine the gap between benchmark success and real clinical impact, arguing that progress toward expert-level diagnostic agents requires advances in causal reasoning, uncertainty quantification, and evaluation beyond static benchmarks. Taken together, this survey complements existing reviews by moving from a catalog of systems toward a principled analysis of *how and why* LLM-based medical agents reason, fail, and succeed in diagnostic contexts. By reframing the field through the lens of diagnostic reasoning and tool-mediated inference, we aim to provide a foundation for more rigorous, safe, and clinically grounded agent design.

D Learning and Adaptation of Medical Agents

In this section, we explore key challenges and emerging strategies in the learning and adaptation of medical AI agents. We discuss the trade-offs between static and continually learning systems, privacy-preserving collaborative training, and techniques for safe autonomous improvement.

✠ **Static vs. Continually Learning Agents.** Static medical agents offer the advantage of easier validation and regulatory certification since their behavior remains fixed once deployed, reducing risks related to unexpected changes in performance (Sun et al., 2025). However, static models may degrade over time as clinical guidelines evolve, new diseases emerge, or population health trends shift, limiting their long-term utility (Jin et al., 2021; Reddy, 2024). Continually learning agents address this by updating their knowledge and adapting to new data, offering responsiveness to changing medical environments and personalized care improvements (Guo et al., 2025a). Yet, this flexibility introduces significant challenges in ensuring model stability, safety, and consistent decision-making, as unchecked updates can amplify errors or biases (Strong et al., 2025; Gu et al., 2025). Balancing adaptability with rigorous oversight remains a critical open problem for medical AI deployment.

✠ **Federated and Privacy-Preserving Learning.** Federated learning frameworks enable collaborative model training across multiple healthcare institutions without exposing sensitive patient data, addressing privacy concerns essential for real-world medical applications (Dutta and Hsiao, 2024). By aggregating locally trained model updates, federated approaches can leverage diverse data distributions, potentially improving generalization across populations and settings (Roberts et al., 2018; Guo et al., 2025b). Despite these benefits, communication bandwidth limitations, heterogeneous data quality, and differences in clinical protocols pose major hurdles for effective model convergence and

consistency (Tang et al., 2024). Moreover, ensuring privacy against adversarial attacks and data leakage remains a significant concern, demanding robust cryptographic and differential privacy techniques (Dutta and Hsiao, 2024). Continued research is needed to balance privacy, accuracy, and system scalability in federated medical AI.

✠ **Experience Replay and Self-Evolution.** Experience replay allows medical agents to iteratively improve diagnostic reasoning and decision-making by revisiting prior cases or simulated experiences, supporting reinforcement learning approaches in controlled environments (Li et al., 2025a). This method can accelerate learning of complex multi-step clinical workflows, enabling agents to refine strategies and correct mistakes without risking patient safety (Jin et al., 2021; Huang et al., 2025b). However, autonomous self-evolution raises critical concerns regarding alignment with clinical goals, interpretability, and safety guarantees, as unchecked learning could lead to unpredictable or harmful behaviors (Zhuang et al., 2025). Ongoing research aims to integrate human-in-the-loop oversight and verification mechanisms to ensure that continuous improvement aligns with medical standards and ethical constraints (Strong et al., 2025). Successfully harnessing experience replay and self-evolution could unlock adaptive, resilient medical agents capable of long-term deployment.

E Diagnostic Tasks and Applications

Here, we outline diagnostic tasks and clinical applications of LLM-based medical agents (Table 2) with current progress and limitations.

✠ **Diagnostic Decision Support.** Diagnostic decision support is the most mature application of LLM-based medical agents, evolving from early medical question answering systems to agentic frameworks that integrate longitudinal context, tool use, and hypothesis tracking (Jin et al., 2021; Jiang et al., 2025d). These agentic decision support systems outperform single-pass models in complex or ambiguous cases (Liu et al., 2025c; Tang et al., 2024; Zeng et al., 2025; Unell et al., 2025), but they can also amplify errors when outputs are accepted uncritically (Huang et al., 2025a; Reddy, 2024). Consequently, real-world deployments remain assistive rather than autonomous (Yan et al., 2024), with regulatory and medico-legal constraints continuing to limit adoption despite strong benchmark performance (Jin et al., 2021).

✠ **Differential Diagnosis and Hypothesis Ranking.** Differential diagnosis focuses on enumerating and ranking multiple plausible conditions, where LLM-based agents outperform traditional classifiers by reasoning over symptom combinations, disease prevalence (Chen et al., 2025a), and structured hypothesis spaces (Guo et al., 2025b; Wei et al., 2023). Multi-agent and collaborative frameworks such as MDAgents further improve performance on complex benchmarks through structured hypothesis discussion and expert critique (Kim et al., 2024b). However, comparative studies show that while explicit hypothesis ranking improves recall for rare diseases, it can reduce precision and suffers from poor calibration without reliable priors (Jiang et al., 2025d; Fan et al., 2024), underscoring the need for probabilistic grounding and clinician oversight (Gu et al., 2025).

✠ **Medication Safety and Treatment Planning.** Medication safety tasks, including drug interaction checks, dosage calculations, and treatment planning under comorbidities, benefit from agents augmented with knowledge graphs and guideline retrieval, which reduce hallucinations compared to pure LLMs (Zuo et al., 2024; Tang et al., 2024). Reflective verification and symbolic constraints improve accuracy in these tasks (Ke et al., 2024; Liu et al., 2025d), yet subtle pharmacological errors persist, especially in cases like renal impairment or polypharmacy (Jin et al., 2021; Reddy, 2024), so most systems limit autonomy and require human confirmation (Yan et al., 2024), keeping treatment planning a high-risk domain.

✠ **Patient-Facing Conversational Agents.** These agents enhance triage, education, and symptom assessment through natural language, improving accessibility and health literacy (Roberts et al., 2018; Jin et al., 2021), with systems like ChatThero offering stressor-aware, memory-persistent multi-session support beyond baseline LLMs (Laymouna et al., 2024; Wang et al., 2025a). However, evaluations reveal risks of unsafe reassurance and delayed escalation (Huang et al., 2025a), even as therapy- and telemedicine-oriented models (e.g., BLISS) show potential for behavioral interventions (He et al., 2020; van Waterschoot et al., 2020; Das et al., 2022). Tool-augmented agents with escalation rules outperform free-form chatbots (Sun et al., 2025; Reddy, 2024), and regulatory guidance stresses guardrails and transparency, restricting deployment mostly to low-risk informational roles (Yan et al., 2024; Zuo and Jiang, 2025).

✠ **Clinical Workflow Automation.** Agentic systems automate documentation, triage routing, and test recommendations by maintaining task state and integrating with hospital systems, significantly reducing clinician workload for note drafting and orders (Tang et al., 2024; Jiang et al., 2025d; Kim et al., 2024a; Roberts et al., 2018; Li et al., 2025b). Task-specific agents tackle documentation variability (e.g., Sporo AI Scribe (Lee et al., 2024)) and generate patient-friendly summaries from technical reports (Sudarshan et al., 2024), while broader frameworks automate prior authorization (Pandey et al., 2024), pharmacy cart assembly, and OCR-based medication reconciliation (Jagatap et al., 2025). Yet, automation errors can cascade through workflows (Jin et al., 2021), highlighting the need for verification and auditability (Noori et al., 2026), and adoption varies across institutions due to integration challenges despite promising domain-specialized agents (Motzfeldt et al., 2025; Arias Russi et al., 2025).

✠ **Training, Simulation, and Medical Education.** Simulation-based agents support clinician training and diagnostic evaluation by providing scalable, repeatable exposure to complex or rare cases (Li et al., 2025a; Fan et al., 2024; Laymouna et al., 2024), with agents acting as standardized patients or tutors to boost engagement and personalized feedback (Sun et al., 2025; Jiang et al., 2025d). Simulations permit greater autonomy and experimentation than real-world settings (Mehandru et al., 2024), though limited fidelity restricts transferability (Jin et al., 2021), positioning these tools as complements rather than substitutes. Multi-agent copilots like MEDCO enable interactive, case-based learning that mirrors clinical teamwork and accelerates expertise via multi-role scenarios (Wei et al., 2024).

✠ **Remote Monitoring and Early Warning Systems.** Remote monitoring agents analyze streaming physiological and self-reported data to detect early signs of deterioration. Tool-augmented agents can integrate wearable data, lab trends, and symptom reports (Tang et al., 2024; Li et al., 2024). Early studies suggest improved sensitivity for adverse event detection compared to rule-based systems (Jiang et al., 2025d; Sun et al., 2025). However, false positives remain common and can overwhelm clinicians (Jin et al., 2021). Robust calibration and thresholding are critical for deployment (Huang et al., 2025a). As with other high-risk domains, autonomy is tightly constrained.

F Datasets and Benchmarks

In this section, we review the primary datasets and benchmarks (Table 8) used to evaluate medical AI agents, covering diverse modalities such as clinical text, medical imaging, and multimodal records. We highlight their roles in assessing diagnostic reasoning, tool use, and interactive capabilities within healthcare contexts.

✠ **Clinical Text and EHR Datasets.** Clinical text datasets derived from electronic health records form the backbone of medical language modeling. Resources such as MIMIC have enabled large-scale research but are limited by de-identification and institutional bias (Johnson et al., 2016; Jin et al., 2021). Agent-based systems often rely on synthetic patient records to avoid privacy issues (Sun et al., 2025; Jiang et al., 2025d). A diverse ecosystem of datasets underpins agent development: foundational clinical corpora (eICU, i2b2) support monitoring and alerting agents (Pollard et al., 2018; Romanov and Shivade, 2018), while large medical dialogue corpora enable training of conversational AIs (He et al., 2020). Concurrently, agent-centric benchmarks and environments (EMBODIEDBENCH, BrowserGym, EHRFlowBench) provide standardized testbeds for evaluating interactive, embodied, and web-enabled agents across long-horizon tasks (Yang et al., 2025c; Chezelles et al., 2025; Zhu et al., 2025). These resources highlight that evaluating agentic behavior requires simulation platforms and longitudinal EHR-style inputs rather than only static QA tests. However, synthetic data may fail to capture real-world complexity (Jiang et al., 2025c). Dataset shift remains a persistent challenge (Pollard et al., 2018). These limitations constrain generalization claims.

✠ **Medical QA and Exam Benchmarks.** Medical QA benchmarks, including exam-style datasets, are widely used to evaluate diagnostic knowledge (Jin et al., 2021; Roberts et al., 2018). While useful for baseline comparison, such benchmarks poorly reflect real diagnostic workflows (Chen et al., 2025a). Agents often achieve high scores through pattern recognition rather than reasoning (Gu et al., 2025). Recent benchmarks incorporate reasoning traces and tool use to address these gaps (Jiang et al., 2025d; Sun et al., 2025). Nonetheless, exam performance does not equate to clinical competence.

✠ **Multimodal Diagnostic Datasets.** Multimodal datasets combine text with imaging, pathology, or audio data. These datasets enable evaluation

Table 8: Datasets and Benchmarks for Evaluating LLM-Based Medical Agents

Category	**Description**	**References**
Clinical Text and EHR Datasets	Primary datasets from EHRs such as MIMIC, eICU, and i2b2 underpin medical language modeling and agent development. Synthetic data is common for privacy but risks losing complexity. Agent-centric benchmarks and environments support interactive, embodied evaluation. Dataset shift and institutional bias remain challenges.	(Johnson et al., 2016; Jin et al., 2021; Sun et al., 2025; Jiang et al., 2025d; Pollard et al., 2018; Romanov and Shivade, 2018; Yang et al., 2025c; Chezelles et al., 2025; Zhu et al., 2025; Jiang et al., 2025c)
Medical QA and Exam Benchmarks	Widely used for baseline diagnostic knowledge evaluation, including medical exams. Limitations include reliance on pattern recognition and poor reflection of real clinical workflows. Recent benchmarks add reasoning traces and tool use for improved assessment.	(Jin et al., 2021; Roberts et al., 2018; Chen et al., 2025a; Gu et al., 2025; Jiang et al., 2025d; Sun et al., 2025)
Multimodal Diagnostic Datasets	Combine text with imaging, pathology, and audio for multimodal agent evaluation, especially in radiology and pathology. Challenges include annotation cost, modality imbalance, and alignment difficulties.	(Guo et al., 2025b; Li et al., 2024; Mehandru et al., 2024; Jiang et al., 2025d; Jin et al., 2021)
Agent-Centric and Interactive Benchmarks	Assess agent capabilities in planning, memory, and multi-step tool use in simulated clinical tasks. More realistic than static QA but often rely on simplified environments. Generalization to real hospitals remains unproven.	(Jiang et al., 2025d; Fan et al., 2024; Mehandru et al., 2024; Sun et al., 2025; Jin et al., 2021)
Simulation-Based Evaluation Environments	Enable stress testing and long-horizon evaluation via simulated clinical environments such as Agent Hospital. Crucial for training and safety validation, though fidelity and realism are ongoing challenges.	(Li et al., 2025a; Mehandru et al., 2024; Jiang et al., 2025d; Jin et al., 2021)
Limitations of Existing Benchmarks	Current benchmarks prioritize accuracy over safety, robustness, and workflow impact. Many lack clinician input or real-world constraints. Benchmark saturation motivates new evaluation paradigms.	(Zuo and Jiang, 2025; Reddy, 2024; Sun et al., 2025; Chen et al., 2025a; Jiang et al., 2025d)

of multimodal agents in radiology and pathology (Guo et al., 2025b; Li et al., 2024). However, annotation cost and modality imbalance limit dataset scale (Mehandru et al., 2024). Benchmarking multimodal agents remains challenging due to alignment issues (Jiang et al., 2025d). As a result, most evaluations remain modality-specific (Jin et al., 2021).

✠ **Agent-Centric and Interactive Benchmarks.** Agent-centric benchmarks evaluate planning, memory, and tool use over multiple steps. Examples include simulated clinical environments and task-oriented benchmarks (Jiang et al., 2025d; Fan et al., 2024). These benchmarks better reflect agent capabilities than static QA tasks (Mehandru et al., 2024). However, they often rely on simplified environments (Sun et al., 2025). Generalization to real hospitals remains unproven (Jin et al., 2021).

✠ **Simulation-Based Evaluation Environments.** Simulation environments enable stress testing and long-horizon evaluation of medical agents. They support safe experimentation with autonomy and failure modes (Li et al., 2025a; Mehandru et al., 2024). Simulations are increasingly used for training and benchmarking (Jiang et al., 2025d). Simulation platforms such as Agent Hospital and virtual patient simulacra allow agents to gain procedural experience at scale, enabling rapid iteration and emergent skill acquisition that would be impractical with real patients (Li et al., 2025a). We view these simulated hospitals and educational environments as critical intermediate steps for validating agent safety, interaction protocols, and longitudinal reasoning prior to clinical pilots. Fidelity and realism remain key challenges (Jin et al., 2021). Nevertheless, simulation is indispensable for progress.

✠ **Limitations of Existing Benchmarks.** Current benchmarks emphasize accuracy over safety, robustness, and workflow impact (Zuo and Jiang, 2025). Many lack clinician involvement or real-world constraints (Reddy, 2024). Benchmark saturation is increasingly observed (Sun et al., 2025). There is a growing consensus that new evaluation paradigms are required (Chen et al., 2025a; Jiang et al., 2025d).

F.1 Representative Quantitative Performance on Standard Benchmarks

Table 9 maps five representative systems to widely recognized medical benchmarks. Because these systems use different base models, tool configurations, and benchmark subsets, direct cross-system numerical comparison is not meaningful; the table is intended to situate each system's reported performance within its specific architectural context as defined by the taxonomy.

G Evaluation Methodologies

Evaluating medical AI agents is inherently complex, as traditional metrics often fail to capture critical aspects of clinical reasoning, communication, and decision-making. Effective evaluation therefore requires a multi-faceted approach that assesses not only predictive accuracy, but also reasoning fidelity, tool-use correctness, robustness under realistic clinical conditions, and real-world workflow impact. Table 10 outlines the principal evaluation dimensions and representative methodologies used to assess LLM-based medical agents.

✠ **Accuracy-Centric Metrics.** Metrics such as accuracy, F1 score, and exact match remain standard for quantifying output correctness but provide limited insight into an agent's clinical reasoning quality (Jin et al., 2021). Surface-level metrics like ROUGE and BLEU inadequately reflect meaningful clinical outcomes, such as treatment appropriateness or diagnostic reliability (Gu et al., 2025; Reddy, 2024). High accuracy can mask systematic errors and overconfidence, especially in safety-critical scenarios. Thus, while useful, accuracy metrics must be complemented by deeper evaluations of clinical performance (Chen et al., 2025a).

✠ **Reasoning Faithfulness and Interpretability.** Interpretability is essential for clinician trust and regulatory compliance. Faithfulness metrics aim to verify that reasoning traces genuinely support final conclusions, yet automatic measurement remains challenging (Li et al., 2023; Sun et al., 2025). ICARE is an interpretable, clinically grounded LLM-based evaluation framework for radiology report generation that distinguishes hallucinations from omissions and aligns closely with clinician assessments (Dua et al., 2025). Frameworks such as Chain-of-Diagnosis (CoD) improve transparency by explicitly modeling diagnostic pathways and uncertainty distributions (Jiang et al., 2025d). Human evaluation often remains necessary to assess whether explanations align with clinical reasoning (Jin et al., 2021).

✠ **Tool-Use Correctness and Grounding.** Medical agents increasingly rely on external tools and multimodal data, necessitating evaluation of correct tool invocation and output interpretation (Mehandru et al., 2024). Incorrect grounding in tools or

Table 9: Representative quantitative performance on standard medical benchmarks. Taxonomy codes: I = Reasoning Paradigm, II = Architecture, III = Tool Access, IV = Autonomy, V = Embodiment. Cross-system numerical comparison should be interpreted with caution due to heterogeneous base models, tool configurations, and evaluation conditions.

System	Base Model	Taxonomy (Axes I-V)	Benchmark	Reported Performance
MDAgents (Kim et al., 2024b)	GPT-4 / Claude / Llama	I: Debate; II: Multi-Agent; III: RAG + KG; IV: HITL; V: Text	10 medical benchmarks (incl. DDXPlus, PathVQA, PMC-VQA)	Best on 7/10 benchmarks; +4.2%-11.8% vs. prior SOTA ($p<0.05$)
CoD (Chen et al., 2025a)	GPT-4 class	I: Explicit Chain; II: Single-Agent; III: None; IV: Assistive; V: Text	DxBench (1,148 cases, 461 diseases); Dxy; Muzhi	SOTA on DxBench; 9,604-disease coverage; entropy-reduced confidence distributions
Agent Hospital (Li et al., 2025a)	GPT-4 class	I: Reflective; II: Multi-Agent + Memory; III: EHR Sim.; IV: Full Auto (Sim.); V: Text	MedQA (USMLE)	SOTA among medical agent methods on MedQA; self-evolving without manual labeling
MedRAX (Fallahpour et al., 2025)	GPT-4o / Claude	I: ReAct; II: Single Planner-Executor; III: Multimodal (MedSAM, CheXagent, et al.); IV: Assistive; V: Multimodal	ChestAgentBench (2,500 queries; 675 expert-validated cases)	SOTA vs. open-source and proprietary models; no additional training required
DeepSeek-R1 (Faray de Paiva et al., 2025)	DeepSeek-R1 (671B, open)	I: Long-Chain RL Reasoning; II: Single-Agent; III: None; IV: Assistive; V: Text	USMLE Steps 1-3; MedQA; MMLU-Pro (health)	USMLE: 92%; MedQA: 94.3%; MMLU-Pro health: 84%; below GPT-o1 (95% USMLE) but above Llama-3.1-405B (83%)

databases is a major source of failure, underscoring the need for verification-aware architectures (Ke et al., 2024; Noori et al., 2026). Realistic evaluation frameworks must simulate complex interactions among text, images, and structured data (Tang et al., 2024). This ensures agents can safely leverage specialized diagnostic tools while minimizing hallucinations.

✠ **Robustness, Stress Testing, and Distribution Shift.** Robustness evaluations assess how well agents maintain performance under noise, ambiguity, or data distribution shifts (Pollard et al., 2018). Medical AI systems often exhibit sharp degradation when confronted with underrepresented populations or rare conditions, as seen in BiasMedQA (Huang et al., 2025a). Stress testing with clinician-guided scenarios helps reveal hidden failure modes overlooked by benchmarks (Jiang et al., 2025d). Continuous monitoring post-deployment is vital to detect drift and ensure ongoing reliability (Reddy, 2024). Standard language metrics fail to capture clinically salient failure modes; agent evaluation requires stress testing under noisy, longitudinal, and multimodal conditions (Jin et al., 2021; Schmidgall et al., 2025). Benchmarks that incorporate multimodal inputs and interactive workflows better reveal degradation under distribution shift (Liu et al., 2025d; Fan et al., 2024), while datasets exposing demographic imbalance (e.g., BiasMedQA) demonstrate sharp performance drops for underrepresented groups (Schmidgall et al., 2024). We therefore advocate combined quantitative and expert-in-the-loop evaluation pipelines.

✠ **Human and Clinician-in-the-Loop Evaluation.** Incorporating clinicians into the evaluation loop provides crucial qualitative insights beyond automated metrics (Yan et al., 2024). Such studies, though resource-intensive and limited in scale, remain essential for assessing real-world usability, safety, and workflow integration (Jin et al., 2021). Frameworks like SurgBox and MEDCO simulate interactive environments where agents assist clinicians, enabling assessment of collaboration quality and decision support effectiveness (Reddy, 2024; Chen et al., 2025a). Hybrid evaluation paradigms combining human oversight with automated measures are gaining traction.

✠ **Outcome-Oriented and Workflow-Level Metrics.** Outcome-based evaluations focus on an agent's impact on clinical workflows and patient outcomes rather than isolated task performance (Tang et al., 2024). Longitudinal and multi-step

assessments capture adaptive planning and complex decision sequences inherent to real healthcare delivery (Jiang et al., 2025d). Benchmarks like EHRFlowBench and simulation platforms such as Agent Hospital enable end-to-end evaluation of diagnostic and therapeutic workflows (Chen et al., 2025a). While such studies remain rare, they are critical for demonstrating true clinical value and readiness for deployment.

H Safety, Ethics, and Reliability

✠ **Hallucinations and Overconfidence.** Across nearly all evaluations, hallucination emerges as the dominant failure mode of LLM-based medical agents, yet its manifestation varies significantly by system design. Pure language models exhibit hallucinations primarily due to reliance on statistical plausibility rather than evidentiary grounding, whereas agentic systems often hallucinate through incorrect tool invocation or misinterpretation of retrieved evidence (Huang et al., 2025a; Ke et al., 2024). Comparative studies show that retrieval-augmented and guideline-grounded agents reduce hallucination frequency but introduce a new class of "grounded hallucinations," where incorrect conclusions are confidently justified using misaligned sources (Sun et al., 2025; Zuo and Jiang, 2025). Overconfidence further compounds this issue, as most models lack calibrated uncertainty estimates and default to decisive language even under ambiguity (Gu et al., 2025; Jin et al., 2021). Reflective and multi-agent systems mitigate some errors but do so unevenly, often trading correctness for verbosity and computational cost (Maiga et al., 2025; Jiang et al., 2025d). Collectively, these findings suggest that hallucination is not a singular defect but an emergent property of how reasoning, tools, and confidence signaling interact.

✠ **Error Propagation in Agentic Systems.** While agentic architectures improve reasoning depth, they introduce systemic risks through error propagation across planning, execution, and memory. In planner-executor systems, early-stage planning errors disproportionately affect downstream decisions, especially when plans are not dynamically revised (Mehandru et al., 2024; Sun et al., 2025). Memory-augmented agents improve longitudinal coherence but risk entrenching incorrect assumptions, a failure mode largely absent in stateless models (Li et al., 2025a; Jin et al., 2021). Multi-agent systems distribute reasoning across roles, reducing individual bias, yet empirical studies show that correlated errors and premature consensus can still dominate outcomes (Maiga et al., 2025; Kim et al., 2024a). Comparative benchmark results indicate that agentic systems outperform single-step models on average but fail more catastrophically when they fail (Fan et al., 2024; Jiang et al., 2025d). This asymmetry highlights the need for explicit failure containment mechanisms rather than assuming robustness emerges from complexity.

✠ **Bias, Fairness, and Population Shift.** Bias in medical agents reflects both historical inequities in healthcare data and architectural assumptions embedded in model design. Systems trained predominantly on Western, tertiary-care datasets consistently underperform in low-resource and non-English clinical settings (Pollard et al., 2018; Roberts et al., 2018). Agentic systems amplify this issue by embedding biased priors into planning heuristics and retrieval strategies (Sun et al., 2025; Jin et al., 2021). Comparative evaluations reveal that multilingual and culturally adapted agents improve surface-level interaction but do not fully correct diagnostic disparities (Jiang et al., 2025c; Chen et al., 2025a). Moreover, population shift affects not only model accuracy but also confidence calibration, leading to systematically misleading recommendations (Huang et al., 2025a). Algorithmic bias and dataset skew remain pervasive: empirical studies show that medical LLMs can lose substantial accuracy on underrepresented populations, underscoring the dual need for more diverse training corpora and transparent evaluation practices (Schmidgall et al., 2024). Privacy risks from model inversion and data extraction persist in agentic systems that access clinical data, motivating differential privacy and synthetic data approaches as partial mitigations (Kandpal et al., 2024; Carlini et al., 2021; Dyda et al., 2021; Dou et al., 2024). These findings suggest that fairness cannot be retrofitted through fine-tuning alone but must be addressed at the level of data curation, evaluation, and deployment context.

✠ **Privacy, Data Leakage, and Security.** Compared to traditional clinical decision support, LLM-based agents introduce qualitatively new privacy risks due to memorization, tool integration, and conversational interfaces. Fine-tuned models have been shown to leak sensitive clinical details through prompt inversion and unintended memorization (Kandpal et al., 2024; Jin et al., 2021). Agentic systems further expand the attack surface by in-

teracting with external tools, enabling prompt injection and cross-system data leakage (Ke et al., 2024). Federated and privacy-preserving learning mitigate some risks but complicate debugging, auditing, and performance monitoring (Dutta and Hsiao, 2024; Sun et al., 2025). Comparative analyses indicate that security failures often arise at system boundaries rather than within models themselves (Reddy, 2024). This shifts the safety burden from model training to system-level engineering and governance (Wasi and Islam, 2025).

✠ **Regulatory and Medico-Legal Considerations.** Regulatory frameworks lag behind the rapid evolution of medical agents, resulting in uncertainty around classification, validation, and liability. Most current systems are positioned as assistive tools to avoid the regulatory burden associated with autonomous medical devices (Yan et al., 2024; Reddy, 2024). Kiseleva et al. (2022) conceptualize transparency in healthcare AI as a multilayered system of accountabilities, integrating legal requirements and technical mechanisms to align explainability, auditability, and governance across developers, clinicians, and patients (Wasi et al., 2025). However, increasing autonomy blurs this distinction, particularly when agents recommend tests or treatments (Chen et al., 2025a). Comparative regulatory analyses reveal inconsistencies across jurisdictions in how software-based agents are evaluated (Jin et al., 2021). The absence of standardized evaluation protocols further complicates approval pathways (Jiang et al., 2025d). As agentic systems mature, regulatory alignment will become a bottleneck rather than a peripheral concern.

✠ **Accountability and Clinical Responsibility.** Accountability in agent-assisted diagnosis is fundamentally distributed across clinicians, developers, and institutions. While clinicians retain legal responsibility, empirical studies show that automation bias leads to over-reliance on agent recommendations (Jin et al., 2021; Wasi and Islam, 2025). Transparent reasoning traces improve post-hoc analysis but do not resolve responsibility attribution in real time (Noori et al., 2026). Multi-agent systems further diffuse accountability by fragmenting decision-making across components (Maiga et al., 2025). Comparative ethical analyses suggest that accountability must be embedded institutionally rather than delegated to interfaces or disclaimers (Reddy, 2024). This reframes medical agents as socio-technical systems rather than autonomous decision-makers.

I System Design and Deployment

The design and deployment of medical AI agents involve complex trade-offs between system capabilities, integration feasibility, and operational sustainability. Architectural choices, IT infrastructure, and human factors collectively influence the ultimate clinical impact and adoption of these technologies.

✠ **System Architectures and Software Stacks.** Architectural paradigms for medical agents vary widely, ranging from lightweight orchestration layers built atop single LLMs to intricate distributed systems that integrate retrieval mechanisms, simulation environments, and continuous monitoring modules (Mehandru et al., 2024; Tang et al., 2024). Modular designs, while improving system auditability and facilitating regulatory compliance, introduce additional latency and integration complexity that can hinder clinical responsiveness (Jin et al., 2021). Conversely, monolithic architectures often yield higher efficiency but obscure internal failure modes, complicating error analysis and risk mitigation. Comparative studies reveal that institutional constraints, such as existing IT infrastructure, regulatory environment, and available expertise, often dictate architectural choices more than considerations of optimal system design (Jiang et al., 2025d). This misalignment partially explains the persistent gap between promising benchmark performance and limited real-world clinical impact (Wasi and Hoque, 2025).

✠ **Integration with Electronic Health Records and Clinical IT Systems.** Seamless integration with Electronic Health Records (EHRs) remains one of the foremost challenges limiting medical agent deployment. Despite the existence of interoperability standards, vendor heterogeneity and disparate data schemas pose significant obstacles to reliable, bidirectional data exchange (Roberts et al., 2018). Agentic systems further complicate integration by demanding real-time data synchronization and the capacity to incorporate dynamic clinical inputs and outputs (Tang et al., 2024). Case studies consistently demonstrate that the effort and cost required for successful integration frequently surpass those invested in model development itself (Jin et al., 2021). This underscores the critical importance of robust, flexible IT infrastructure, which often becomes the limiting factor in deployment feasibility (Reddy, 2024).

✠ **Monitoring, Logging, and Auditing.** Sustained clinical safety and efficacy require continuous mon-

itoring to detect performance degradation and concept drift post-deployment. Systems capable of logging detailed reasoning traces and interactions with external tools provide invaluable data for retrospective audits and root-cause analyses; however, such comprehensive logging raises significant concerns regarding data storage, privacy, and security (Noori et al., 2026). Conversely, lightweight logging approaches reduce operational overhead but may fail to capture sufficient diagnostic information to resolve failures effectively (Sun et al., 2025). Comparative analyses advocate for tailoring monitoring strategies to specific clinical tasks and contexts rather than relying on generic frameworks (Jiang et al., 2025d). Agentic systems that call external tools and run multi-step workflows require structured audit trails and reproducible tool invocations to enable post-hoc verification and regulatory compliance (Gao et al., 2026). Tool-orchestration platforms and AutoML-style agent frameworks (e.g., AutoPE) illustrate how modular designs can expose intermediate artifacts for auditing while supporting domain scientists who lack deep ML expertise (Liu et al., 2024). Consequently, auditing is best conceived as a continuous operational process rather than a discrete certification event.

✠ **Model Updating, Drift Management, and Maintenance.** The rapid evolution of medical knowledge creates inherent tension between the need for model stability and the imperative to adapt to new evidence and clinical guidelines. Static models facilitate easier regulatory validation but are vulnerable to performance degradation due to data and concept drift (Pollard et al., 2018). Conversely, continual learning techniques enhance adaptability but risk destabilizing previously validated and trusted behaviors, raising safety concerns (Sun et al., 2025). Emerging hybrid strategies that combine a stable core model with updatable peripheral tool modules present a pragmatic balance between robustness and flexibility (Chen et al., 2025a). Nonetheless, ongoing maintenance, including updating datasets, retraining, and validation, remains a substantial operational burden and a major barrier to sustainable deployment (Jin et al., 2021).

✠ **User Interfaces for Clinicians and Patients.** The design of user interfaces critically shapes how medical agents influence clinical workflows and decision-making. Comparative studies consistently show that clinicians favor succinct, actionable recommendations accompanied by explicit uncertainty quantification over verbose explanatory outputs, which can increase cognitive load and diminish trust (Jin et al., 2021). For patient-facing applications, interface design must incorporate stringent safeguards to prevent misunderstanding and ensure that information is accessible and non-alarming (Reddy, 2024). PAME-AI leverages a multi-agent, DIKW-based framework to optimize patient messaging, significantly improving engagement and medication adherence at scale (Luo et al., 2025). Medbot is a clinician-designed AI agent platform that automates complex clinical workflows, reducing administrative burden and improving care quality in conditions like sepsis Anthony et al. (2025). Effective interface design is thus fundamental to patient safety and clinician acceptance, underscoring the need for human-centered approaches that carefully balance informativeness and simplicity.

✠ **Barriers to Real-World Adoption.** Despite impressive technical advances, widespread clinical adoption of medical AI agents remains limited. Systematic analyses identify trust deficits, unclear liability frameworks, integration costs, and workflow disruption as dominant impediments (Jin et al., 2021; Reddy, 2024). Furthermore, demonstrations of clear clinical value are often context-specific and fail to generalize across diverse healthcare settings (Tang et al., 2024). These findings suggest that barriers to adoption are frequently rooted less in the technical capabilities of models and more in institutional readiness, regulatory environments, and socio-technical factors (Chen et al., 2025a). Addressing these challenges will require concerted efforts spanning technology development, policy-making, and organizational change management.

✠ **Contextual AI Frameworks for Low-Resource Settings.** Frameworks designed for low-resource settings must address challenges such as limited training data and constrained computational infrastructure, particularly in medical institutions with restricted resources (Liu et al., 2025c). Approaches like DrAgent focus on building LLMs as agents to deliver accurate medical decision-making without requiring the extensive data and computational resources typically needed for fine-tuning competitive medical LLMs (Liu et al., 2025c). Additionally, conversational agents offer a promising direction for increasing healthcare access in low-resource contexts, such as telemedicine (He et al., 2020). Research in automated motivational interviewing demonstrates that both in-context learning (ICL) and instruction fine-tuning (IFT) can successfully train language models for clinical dialogue

tasks even under low-resource settings, helping to automate complex interactions (Hoang et al., 2024). The ability to leverage synthetic data, such as generating Arabic medical dialogues from patient notes regardless of the original language, also helps overcome dataset availability barriers for broader research and application in medical dialogue systems (ALMutairi et al., 2024). To specifically address unique challenges in countries like Bangladesh, the MEDIC-BD (Wasi et al., 2026) framework proposes strategies tailored to local healthcare constraints, aiming to enhance diagnostic accuracy, streamline clinical workflows, and improve patient care through context-aware AI integration. In low-resource healthcare settings, constrained compute and limited labeled data require agent designs that prioritize in-context learning and lightweight inference; systems such as DrAgent exemplify LLM-agent approaches optimized for limited resources (Liu et al., 2025c). Conversational and telemedicine agents have particular promise for expanding access when deployed with careful calibration and data-efficient training strategies (He et al., 2020; Hoang et al., 2024). Synthetic-data augmentation (e.g., cross-language medical dialogue generation) can further reduce dataset barriers in multilingual contexts (ALMutairi et al., 2024).

J More Details on Open Challenges and Future Directions

Despite notable progress in developing AI-powered medical agents, particularly those leveraging large language models (LLMs), several critical challenges continue to impede their safe, reliable, and equitable deployment in clinical practice. These challenges span fundamental reasoning capabilities, uncertainty management, generalization, evaluation methodologies, and real-world integration. Addressing them is essential to advance from promising prototypes to impactful healthcare tools, as summarized in Table 11.

✠ **Causal and Counterfactual Medical Reasoning.** Current medical AI systems predominantly rely on correlational patterns extracted from training data, which limits robustness under distribution shifts and novel clinical contexts (Chen et al., 2025a; Sun et al., 2025). This corresponds to a gap at the intersection of Axis I (Reasoning Paradigms) and Axis III (Tool Use): associative implicit reasoning cannot distinguish between a symptom that causes a disease and one that merely co-occurs with it, and no current tool-use configuration compensates for this. A concrete testable question is: *Can tree-based search agents (Axis I) that invoke explicit causal graphical model tools (Axis III) reduce confounding errors in differential diagnosis compared to RAG-augmented chain-of-thought agents on datasets with documented confounders?* Progress on this question requires causal benchmarks that annotate confounding pathways, which are currently absent from standard medical QA datasets.

✠ **Uncertainty Quantification and Calibration.** Effective deployment of medical agents demands not only accurate predictions but also well-calibrated uncertainty estimates. Existing approaches, such as ensembles, Bayesian inference, and abstention-based methods, improve uncertainty quantification but often at the expense of computational complexity and system interpretability (Gu et al., 2025). Strong et al. (2025) proposed a guided deferral framework for trustworthy human–AI collaboration in healthcare, enabling LLMs to abstain from uncertain predictions while providing structured guidance to clinicians, thereby improving safety and calibration under real-world clinical constraints. Moreover, transparent communication of uncertainty to clinicians and patients is critical to foster trust and support informed decision-making (Jin et al., 2021). Despite advances, poor calibration remains a systemic vulnerability, and developing practical, scalable mechanisms for uncertainty estimation and explanation is an active frontier in AI for medicine.

The concrete testable hypothesis for this challenge is: *In multi-agent debate architectures (Axis II), does the consensus mechanism degrade individual agent calibration—measured by Expected Calibration Error (ECE) before and after debate rounds—and can a guided deferral module (Axis IV) restore pre-debate ECE levels?* Strong et al. (2025) provide a deferral framework applicable to this design, but the interaction between debate-induced overconfidence and deferral-based correction has not been empirically isolated.

✠ **Multimodal and Multilingual Generalization.** Medical agents must generalize across a wide spectrum of data modalities, including clinical text, imaging, laboratory values, and structured EHR data, and be deployable in multilingual, multicultural environments to ensure global health equity (Roberts et al., 2018). Current systems often ex-

hibit English-language bias and perform best in imaging-rich contexts prevalent in high-resource settings. Bridging these gaps requires new datasets, architectures capable of robust multimodal fusion, and language-agnostic representations that can accommodate regional documentation practices and clinical workflows (Guo et al., 2025b). Equity and inclusivity considerations are paramount, necessitating research frameworks that prioritize underrepresented populations and languages.

✠ **Evaluation Beyond Benchmarks.** The saturation of static benchmarks in medical AI research constrains meaningful assessment of real-world clinical performance (Sun et al., 2025). Traditional tasks focusing on question-answering or isolated diagnostic queries fail to capture the complexity of clinical reasoning, longitudinal patient management, and interactive communication critical to healthcare. To overcome these limitations, evaluation methodologies must evolve to incorporate outcome-based metrics, longitudinal studies, and simulation environments that reflect the multimodal, sequential, and dynamic nature of clinical practice (Jiang et al., 2025d; Chen et al., 2025a). Importantly, clinical expert judgment and workflow-level validation should complement automated metrics to ensure relevance and rigor.

✠ **From Simulated to Real Clinical Impact.** While simulations provide safe and scalable platforms for prototyping medical agents, translating performance gains to actual clinical settings remains a formidable challenge (Li et al., 2025a; Du et al., 2025). Integration with heterogeneous EHR systems, compliance with regulatory frameworks such as GDPR and FDA guidelines, and alignment with established clinical workflows are necessary but nontrivial steps (Reddy, 2024). Additionally, clinician acceptance hinges on system transparency, usability, and minimal disruption. Addressing cost-effectiveness and maintenance over time is also critical for sustainable adoption. Looking ahead, embedding agents within physical healthcare environments, such as robotic assistants or continuous patient monitoring systems, represents an ambitious but promising direction.

We see promise in combining reinforcement learning, long-chain reasoning, and inference-time scaling to improve complex diagnostic planning and therapeutic reasoning (Guo et al., 2025a; Faray de Paiva et al., 2025; Huang et al., 2025b). Large-scale simulated hospitals and iterative self-play techniques enable accelerated agent learning and can produce emergent capabilities; nevertheless, ensuring alignment and safety during self-evolution remains a pressing open problem (Li et al., 2025a). Presently, medical agents can approximate fragments of such reasoning through advanced reinforcement learning and long-chain inference strategies, and demonstrate improving performance in simulated hospital environments and benchmark tasks (Jin et al., 2021; Reddy, 2024). Nevertheless, fundamental gaps remain in causal understanding, contextual judgment, and accountability mechanisms. Whether these gaps are surmountable with current paradigms or demand fundamentally new approaches is an open question, one that will shape the trajectory of AI integration into clinical practice.

K Case Studies: Applying the Framework

We walk through four representative systems to demonstrate what the five-axis taxonomy reveals that a descriptive, application-based overview would miss.

MDAgents (Kim et al., 2024b). An application-based grouping places MDAgents in the category medical decision support. The taxonomy instead plots it at the intersection of *Debate-Based Reasoning* (Axis I) and *Multi-Agent Cooperative Architecture* (Axis II), with *External Knowledge Grounding via RAG* (Axis III) and *Bounded Human-in-the-Loop Autonomy* (Axis IV). This intersection immediately explains an empirical result that an application label cannot: the 11.8% average accuracy gain from moderator review combined with external knowledge (Kim et al., 2024b) is structurally expected from the taxonomy because debate-based reasoning reduces anchoring bias (Axis I benefit) while the moderator enforces bounded autonomy (Axis IV constraint), preventing unchecked flawed consensus (the Axis I×Axis II interaction identified in §3.7).

Chain-of-Diagnosis (CoD) (Chen et al., 2025a). An application label would call CoD prompt engineering for diagnosis. The framework maps it to *Explicit Chain Reasoning* (Axis I) and *Single-Agent Planner* (Axis II) with no external tool access (Axis III). This axis profile immediately predicts CoD's primary structural limitation: because it is a single-agent, tool-free, linear explicit reasoner, it has no mechanism for recovering from an early diagnostic error—each step conditions on the previous one, so an incorrect hypothesis at step t propa-

gates deterministically. The taxonomy thus predicts the "linear error propagation" failure mode from architecture alone, before any empirical evaluation.

Agent Hospital (Li et al., 2025a). Application-based grouping categorizes Agent Hospital as a simulation environment. The taxonomy maps it to *Fully Autonomous Simulated* (Axis IV) and *Large-Scale Multi-Agent with Memory Augmentation* (Axis II), with *EHR Simulation Tools* (Axis III) and *Text-Only Embodiment* (Axis V). This profile immediately clarifies why Agent Hospital achieves strong performance on MedQA but is structurally blocked from clinical deployment: Axis IV full autonomy is incompatible with current FDA SaMD classification, and Axis V text-only embodiment means the system cannot perceive the physical or multimodal signals (vital signs, imaging, physical exam) required for real-world diagnosis.

MedRAX (Fallahpour et al., 2025). MedRAX would be categorized descriptively as multimodal radiology AI. The taxonomy maps it to *ReAct-Based Tool Orchestration with Dynamic Tool Selection* (Axis I), *Single-Agent Planner-Executor* (Axis II), *Multimodal Tool Ecosystem including MedSAM, CheXagent, and RoentGen* (Axis III), and *Multimodal Virtual Agent Embodiment* (Axis V). The cross-axis interaction (§3.7, Embodiment×Tool Use) predicts that MedRAX's diagnostic accuracy is bounded by the weakest tool in its chain—a prediction consistent with the observed modality-specific failure modes in segmentation-dependent queries on ChestAgentBench.

Table 10: Evaluation Methodologies for LLM-Based Medical Agents

Methodology	**Evaluation Focus and Characteristics**	**Representative References**
Accuracy-Centric Metrics	Standard metrics such as accuracy, F1, and exact match quantify output correctness but fail to capture clinical reasoning quality or safety. Surface-level language metrics (e.g., ROUGE, BLEU) poorly reflect diagnostic reliability and treatment appropriateness. High scores may mask systematic errors and overconfidence in safety-critical settings, necessitating complementary evaluation dimensions.	(Jin et al., 2021; Gu et al., 2025; Reddy, 2024; Chen et al., 2025a)
Reasoning Faithfulness and Interpretability	Evaluates whether reasoning traces genuinely support conclusions, which is critical for clinician trust and regulatory compliance. Automatic faithfulness measurement remains challenging, often requiring human judgment. Frameworks such as Chain-of-Diagnosis explicitly model diagnostic pathways and uncertainty, while clinically grounded tools like ICARE distinguish hallucinations from omissions and align with clinician assessments.	(Li et al., 2023; Sun et al., 2025; Dua et al., 2025; Jiang et al., 2025d; Jin et al., 2021)
Tool-Use Correctness and Grounding	Assesses correct invocation, sequencing, and interpretation of external tools, databases, and multimodal inputs. Incorrect grounding is a major failure mode, motivating verification-aware architectures and auditing mechanisms. Realistic evaluation requires simulation of complex interactions among text, images, and structured clinical data.	(Mehandru et al., 2024; Ke et al., 2024; Noori et al., 2026; Tang et al., 2024)
Robustness, Stress Testing, and Distribution Shift	Measures performance stability under noise, ambiguity, rare conditions, and demographic imbalance. Agents often degrade sharply under distribution shift, as revealed by stress testing and bias-focused benchmarks. Longitudinal, multimodal, and interactive evaluations better expose clinically salient failure modes than static QA metrics. Continuous monitoring is essential post-deployment.	(Pollard et al., 2018; Huang et al., 2025a; Jiang et al., 2025d; Reddy, 2024; Jin et al., 2021; Schmidgall et al., 2025; Liu et al., 2025d; Fan et al., 2024; Schmidgall et al., 2024)
Human and Clinician-in-the-Loop Evaluation	Integrates clinician feedback to assess usability, safety, trust, and workflow integration beyond automated metrics. Though resource-intensive and limited in scale, human evaluation remains indispensable. Interactive environments such as SurgBox and MEDCO enable assessment of collaboration quality and decision-support effectiveness. Hybrid human–automated paradigms are increasingly adopted.	(Yan et al., 2024; Jin et al., 2021; Reddy, 2024; Chen et al., 2025a)
Outcome-Oriented and Workflow-Level Metrics	Focuses on real-world clinical impact, including workflow efficiency and patient outcomes, rather than isolated task accuracy. Longitudinal and multi-step evaluations capture adaptive planning and end-to-end decision processes. Benchmarks like EHRFlow-Bench and simulation platforms such as Agent Hospital support holistic assessment of diagnostic and therapeutic pipelines.	(Tang et al., 2024; Jiang et al., 2025d; Chen et al., 2025a)

Table 11: Open challenges and future directions for LLM-based medical agents.

Area	**Key Challenges**	**Future Directions**
Causal and Counterfactual Reasoning	Predominantly correlational learning limits robustness under distribution shift and novel clinical scenarios; lack of explicit causal modeling hinders intervention-level reasoning.	Integrating causal inference and counterfactual reasoning frameworks with LLMs; hybrid symbolic–statistical models enabling intervention-aware diagnosis and treatment planning (Chen et al., 2025a; Sun et al., 2025).
Uncertainty Quantification and Calibration	Poorly calibrated confidence estimates undermine safety; existing Bayesian and ensemble methods increase complexity and reduce interpretability.	Scalable uncertainty-aware architectures, abstention and deferral mechanisms for human–AI collaboration, and transparent uncertainty communication to clinicians (Gu et al., 2025; Strong et al., 2025; Jin et al., 2021).
Multimodal and Multilingual Generalization	Performance degrades across modalities, languages, and low-resource settings; strong bias toward English text and imaging-rich environments.	Language-agnostic representations, robust multimodal fusion architectures, and datasets prioritizing underrepresented populations and clinical settings (Roberts et al., 2018; Guo et al., 2025b).
Evaluation Beyond Static Benchmarks	Benchmark saturation; static QA-style tasks fail to capture longitudinal reasoning, interaction, and workflow complexity.	Outcome-oriented, longitudinal, and simulation-based evaluation with clinician-in-the-loop assessment and workflow-level validation (Sun et al., 2025; Jiang et al., 2025d; Chen et al., 2025a).
From Simulation to Clinical Deployment	Gap between simulated performance and real-world impact; challenges in EHR integration, regulation, workflow alignment, and clinician trust.	End-to-end clinical pilots, regulatory-aware system design, usability-centered deployment, and cost-effective maintenance strategies (Reddy, 2024; Li et al., 2025a).
Scaling Reasoning and Autonomy	Long-horizon planning, self-improvement, and autonomy introduce alignment and safety risks; limited causal grounding and accountability.	Reinforcement learning, inference-time scaling, and large-scale simulated hospitals with safety-constrained self-play and oversight mechanisms (Guo et al., 2025a; Li et al., 2025a).