# The Landscape of Medical Agents: A Survey

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Medical Agents are an emerging class of agentic systems deployed in clinical settings that operate over multimodal, longitudinal data, maintain internal state, plan and adapt sequences of actions, and interact with clinical information systems under governance constraints. They extend traditional medical artificial intelligence (MedAI) beyond narrow diagnostic and predictive models toward workflow-centric architectures that address persistent challenges such as administrative burden, fragmented workflows, and workforce strain. In this paper, we (i) propose a functional definition and three-level developmental roadmap for Medical Agents, linking architectural capabilities (planning, memory, tool use, long-horizon control) to degrees of workflow integration and autonomy; (ii) map representative deployments across hospital departments and tasks, including domain-specific agents and multi-agent hospital simulations; and (iii) synthesize cross-cutting challenges in safety, robustness, fairness, evaluation, and governance, outlining research directions for advancing capabilities under clinical constraints and achieving system-level impact. We argue that Medical Agents should be treated as emerging infrastructure for learning health systems, whose value will be measured less by benchmark accuracy than by reliable restructuring of clinical workflows.

## 1 Introduction

### 1.1 The Rise of General AI Agents: From Static Models to Autonomous Systems

Artificial Intelligence is undergoing an architectural shift from static, query-bound LLMs to dynamic, goal-directed agents built around a Perception–Cognition–Action loop [76]. Rather than treating the LLM as a passive knowledge engine, agentic systems couple it with memory, planning, and tool use so that it can pursue multi-step objectives over extended time horizons. In this survey we focus on how this general agentic paradigm is instantiated in medicine, where safety, regulation, and organisational constraints sharply shape how agents can be designed, deployed, and governed. A complementary empirical review by Vatsal et al. maps 49 healthcare agent studies onto a seven-dimensional taxonomy spanning cognitive capabilities, knowledge management, interaction patterns, adaptation, safety, framework typology, and core tasks, surfacing clear asymmetries: external knowledge integration is commonly realised ($\sim$76%) whereas event-triggered activation ($\sim$92% absent) and drift detection ($\sim$98% absent) remain largely unaddressed, indicating that current agents excel at retrieval-grounded advising but require stronger adaptation and compliance platforms [273].

The contemporary AI landscape is defined by a shift from self-contained, passive LLMs to integrated, active agents. Standalone LLMs operate as stateless one-shot predictors, whereas agents wrap these models in scaffolding for persistent state, planning, and tool-mediated interaction with external systems [126; 321]. The foundational architecture enabling this leap toward autonomy is a continuous operational cycle, a core **Agentic Loop** comprising three distinct but interconnected stages: **Perception, Cognition, and Action**. This framework is closely related to what prior work has variously described as sense–plan–act, observe–reason–act, or perception–action loops, and is not an arbitrary engineering choice; it is deeply rooted in principles from cognitive science and neuroscience, modeling the modular structure of intelligent biological systems that must navigate and interact with a dynamic world [255].

Perception constitutes the agent's sensory interface, its channel for receiving information from the external world. Its primary function is to capture raw environmental inputs and transform them into structured,

meaningful representations that can ground subsequent reasoning in real-world observations. This is an increasingly sophisticated process that handles diverse and often multimodal data streams. An agent must seamlessly integrate textual inputs from user commands or lengthy documents, visual data from cameras or screen captures, structured signals from API responses, and, in the case of embodied agents, continuous streams of physical sensor readings from systems like GPS or LIDAR. However, perception is more than mere input processing. The most critical process within this stage is **semantic interpretation**: the translation of this raw data into an actionable model of the current state of the world. A failure in this stage, that is misinterpreting the sentiment of a user email, failing to extract the correct value from a JSON response, or misclassifying an object in a visual field, would provide a flawed "ground truth" to the cognitive engine. The subsequent decisions, while logically sound based on the flawed model, would be disconnected from reality, thereby compromising the agent's effectiveness and safety.

The cognitive core is where the LLM controller operates, augmented by sophisticated submodules that enable coherent, goal-directed behavior over time. This is the agent's "brain," responsible for deliberation, strategy, and memory. Planning, a key cognitive function, grants the agent the ability to think steps ahead. The most common technique is **Hierarchical Task Decomposition**, whereby the agent recursively breaks down abstract objectives (e.g., "Conduct a market analysis for our new product") into a concrete sequence of executable sub-tasks (e.g., 1. Identify top 3 competitors; 2. For each competitor, search for recent financial reports; 3. Synthesize findings into a summary; 4. Identify market trends from tech journals; 5. Draft a final report). Crucially, this process must be dynamic. Real-world execution is fraught with uncertainty. Therefore, advanced agents continuously engage in **Dynamic Replanning and Resilience**. They monitor the outcomes of their actions and adapt their plans in response to errors, unexpected outputs, or new information, a capacity for self-correction and strategic adjustment that is essential for navigating complexity. This resilience is built upon a multi-layered memory architecture that liberates the agent from the LLM's finite context window [211; 277]. Mirroring human cognitive structures, this system includes a short-term **Working Memory** for immediate context; a long-term **Episodic Memory**, often implemented with vector databases, storing specific past experiences to learn from mistakes; a **Semantic Memory** for generalized, context-independent knowledge; and a **Procedural Memory** that encodes optimized workflows for recurring tasks. Finally, a **Decision Arbitration** submodule acts as the central orchestrator. It evaluates the inputs from perception and memory against the agent's goals, weighs competing plans, manages uncertainty, and selects the optimal next action, thereby translating complex deliberation into a single, intended behavior.

Action is the final phase of the loop, where the agent executes its decisions and interacts with its environment. **Tools** serve as the agent's "hands and eyes," extending its capabilities beyond the LLM's internal computations into the real world. In digital environments, this is primarily achieved through **API calls** and **Function Calling**, which empower the agent to access real-time information from the web, execute code for data analysis, query proprietary databases, and interact with virtually any external software system [219]. The sophistication of an agent is often a direct function of its **Tool Orchestration** capabilities. This involves not only selecting the right tool but also meticulously preparing its parameters (often by transforming the output of a previous tool), managing complex dependency chains, and implementing robust error-handling and retry logic. For example, a single high-level goal might require orchestrating a search query, then passing those results to a data extraction tool, whose output is then used to parameterize a code execution environment, which in turn writes to a file that is finally attached to an email. The scope of action is vast and growing, ranging from these intricate digital tasks like automated software debugging and financial modeling to direct physical actuation in the domains of robotics and autonomous vehicles.

### 1.2 The Clinical Imperative: Beyond Diagnostic Tools

Having outlined the evolution of general-purpose agentic systems above, we now specialize this view to *medical agents*, discussing how these capabilities reflect in clinical tasks, workflows, and safety-critical environments.

#### 1.2.1 From Diagnostic Tools to Medical Agents: A Strategic Clinical Imperative

The global healthcare system is under structural strain from unsustainable workload, inefficient workflows, and escalating clinician burnout. Current generations of medical AI have delivered important diagnostic

gains, but most deployed systems remain narrow, task-specific "point solutions" around the EHR that introduce extra streams of alerts and scores for clinicians to interpret rather than removing work. Health systems therefore face a strategic inflection point: moving from fragmented, single-purpose tools to integrated, autonomous **Medical Agents** that manage multi-step clinical and administrative workflows, integrate multimodal longitudinal data, surface context-aware recommendations, and execute routine steps such as drafting notes or populating order sets under robust governance.

**Structural failures of contemporary clinical workflows.** Clinical practice today is dominated by high administrative and cognitive load. A substantial share of physician time is absorbed by documentation, EHR navigation, inbox management, and coordination tasks [249], much of it mediated through fragmented interfaces that push work into after-hours "pajama time" [14; 232]. This creates an **administrative–economic feedback loop**: low-value tasks drive burnout and turnover, reduce available clinical capacity, and ultimately undermine care quality. Against this backdrop, agentic AI should be viewed not as a peripheral enhancement but as part of a broader strategy to redesign workflows and systematically remove avoidable burden.

In parallel, there is growing evidence that strategic deployment of AI-driven automation can yield material savings at the system level [86]. Various analyses suggest that intelligently applied automation could reduce healthcare expenditures by a significant margin, particularly if focused on labor-intensive, repetitive tasks that do not intrinsically require human judgment. Importantly, the next generation of AI must not only maintain or improve on the diagnostic accuracy gains achieved by earlier systems, for example by reducing missed radiological or pathological diagnoses, but also address the friction that surrounds diagnosis, follow-up, and care coordination. The economic case for Medical Agents therefore rests on a dual mandate: sustaining clinical quality while structurally reducing the operational drag that currently absorbs clinician time.

**Limits of traditional point-solution MedAI.** Traditional MedAI has achieved important but bounded successes, primarily through deep learning models designed for singular, well-defined tasks. In imaging, for example, algorithms have reduced missed diagnoses in specific pulmonary conditions and improved the detection of malignancies in digital histopathology [159]. These systems serve as powerful pattern recognizers and have demonstrated that well-curated data and targeted models can measurably improve certain aspects of diagnostic performance.

However, these achievements also define the limits of the prevailing paradigm. Most current systems function as **point solutions**: they answer narrowly scoped questions such as "Is there a tumor on this scan?" or "What is the estimated risk score for this patient?" without engaging with the broader clinical context or the longitudinal nature of care. They do not coordinate downstream actions, manage the sequence of follow-up tests, or reconcile their outputs with competing priorities such as comorbidities, patient preferences, or resource constraints. Moreover, many of these models suffer from structural limitations: they lack standardized deployment patterns, require frequent retraining and recalibration, and are challenging to maintain over time within complex hospital IT environments. As a result, their contributions remain local rather than system-transforming.

A central technical limitation of the current MedAI landscape is the failure to integrate outputs across tools in a reusable, machine-actionable form. For instance, in medical imaging, annotations generated by AI systems are often stored in proprietary or non-standard formats that do not support simple reuse as input for subsequent models or downstream analytics [66]. This design choice accumulates into a form of **technical debt of fragmented AI implementation**: every new system must rebuild its own data infrastructure rather than leveraging a shared, interoperable substrate.

This fragmentation mirrors long-standing issues in the broader health IT ecosystem, where EHR platforms frequently maintain patient information in siloed databases that inhibit longitudinal, cross-setting analysis. Clinicians rarely interact with a unified, whole-person view of the patient; instead, they navigate through multiple interfaces and partial records. When AI tools are deployed as additional, isolated modules, they risk becoming yet another layer of fragmentation that clinicians must manually synthesize. From an architectural perspective, the problem is not only whether a model makes accurate predictions, but whether its outputs can be seamlessly integrated into an end-to-end workflow in which they trigger and shape subsequent steps.

The design intent behind many traditional AI tools is to reduce clinician workload by automating "tiring and repetitive" subtasks such as morphological analysis, lesion detection, or filtering images that do not warrant review. In practice, however, the current deployment model often introduces a paradox. These systems generate annotations, risk scores, or predictive labels that are presented to clinicians as advisory outputs. They do not, by default, interpret those outputs within the context of the entire patient trajectory or execute the corresponding workflow steps.

As a result, clinicians must still integrate disparate AI outputs with clinical guidelines, institutional protocols, and patient-specific factors, and then manually initiate orders, update documentation, and communicate with the care team. Each additional tool typically adds new alerts, dashboards, or notifications into an already crowded EHR environment, exacerbating alert fatigue. In this sense, traditional point solutions tend to reduce low-level manual effort while leaving, or even increasing, the high-level cognitive effort required for synthesis and decision-making. This paradox highlights the need for a new architectural approach in which AI systems are designed not merely to compute but to coordinate, orchestrate, and execute within workflows.

To crystallize this contrast, it is useful to compare traditional MedAI tools with agentic architectures along several dimensions.

A concise comparison between traditional MedAI point solutions and agentic Medical Agents can be made along five axes: primary goal (single-task prediction versus multi-step workflow execution), data handling (single-modality versus multimodal longitudinal reasoning), workflow impact (additional alerts versus automation of routine steps), autonomy (no autonomy versus semi-autonomous execution under oversight), and adaptiveness (static models versus governed, data-driven updates).

### 1.2.2 Why Healthcare Demands Agentic Architectures

The case for agentic AI in healthcare is grounded in the intrinsic complexity of clinical data and decision-making. Modern medicine operates on multimodal, temporally structured, and safety-critical data; static single-task models struggle to fuse heterogeneous inputs, reason over evolving trajectories, and expose their reasoning in clinically meaningful ways. Agentic systems, by contrast, are explicitly designed to integrate structured EHR fields, unstructured notes, streaming sensor data, and guidelines into a unified, interpretable view of the patient, and to provide explainable, workflow-aligned recommendations that respect the high-stakes, norm-governed nature of care [247; 245; 230].

### 1.2.3 Designing Medical Agents and Agentic Workflows

We use the terms "Medical Agents" and "agentic AI workflows" to refer to architectures in which agentic components are explicitly embedded in clinical processes and bounded by governance and oversight. Compared with traditional MedAI, these systems are (i) autonomy-enhancing, executing multi-step tasks within well-specified workflows rather than emitting one-off predictions; (ii) reasoning-centred, synthesising heterogeneous inputs to make context-sensitive decisions; (iii) planning-enabled, updating plans as new data arrive; and (iv) workflow-optimising, coordinating processes to reduce friction and delay.

In practice, health systems typically deploy modular "swarms" of task-specific agents for documentation support, triage, resource allocation, trial recruitment, and other functions, coordinated by an orchestration layer rather than a single monolithic model [124]. A hybrid AI–human model remains essential: clinicians initiate and supervise tasks and retain authority over high-consequence decisions, while agents handle routine execution, maintain continuity across handoffs, and reclaim time from high-volume administrative work such as intelligent documentation, prior authorisation, quality measurement, and coding and billing.

**Domain-specific applications of Medical Agents.** Beyond administrative burden, Medical Agents are increasingly relevant across several core domains of healthcare delivery and biomedical science. These domains illustrate how agentic architectures can be tailored to distinct workflows while sharing common infrastructure for perception, cognition, and action.

**Clinical diagnosis and decision support.** In diagnostic workflows, Medical Agents can operate as continuous companions that monitor multimodal patient data, maintain and update differential diagnoses,

and highlight early warning signals of deterioration. Agents can synthesize imaging findings, laboratory trajectories, vital sign trends, and free-text clinical notes into structured diagnostic hypotheses, propose targeted follow-up tests, and flag guideline-relevant steps that may have been overlooked. By embedding these capabilities directly in the electronic health record, agents can provide real-time, context-aware decision support at the moment of ordering or documentation, while preserving the clinician as the final arbiter of care [263].

**Drug discovery and development.** In pharmaceutical research and development, Medical Agents can orchestrate complex, data-intensive workflows that span target identification, preclinical evidence synthesis, trial design, and post-marketing surveillance. For early-stage discovery, agents can autonomously scan biomedical literature, omics databases, and real-world evidence to prioritize targets and mechanisms, generate structured summaries of prior attempts, and surface safety signals from related compounds. During clinical development, agents can support protocol authoring, eligibility criteria refinement, site selection, and patient recruitment by integrating data from registries, clinical trial management systems, and electronic health records. In later phases, agents can monitor safety and effectiveness signals across heterogeneous data sources, supporting adaptive trial designs and pharmacovigilance activities [47; 81; 330].

**Rehabilitation and longitudinal care management.** In rehabilitation medicine and chronic disease management, Medical Agents can function as coordinators of long-horizon care plans that extend beyond single encounters or inpatient stays. Agents can aggregate data from physical therapy notes, patient-reported outcomes, wearable sensors, and home monitoring devices to track recovery trajectories, adherence to exercise regimens, and functional milestones. Based on this longitudinal view, they can recommend adjustments to therapy intensity, prompt clinicians when goals are not being met, and generate personalized education materials or exercise schedules for patients. In integrated care teams, agents can also help coordinate tasks among physicians, therapists, nurses, and social workers, reducing the risk that critical follow-ups or referrals are lost in handoffs [131].

**Medical education and training.** In medical education, Medical Agents can serve as adaptive tutors and simulation orchestrators that personalize training to the needs of individual learners. Agents can generate realistic virtual patients, construct case-based learning scenarios aligned with curricular objectives, and provide stepwise feedback on diagnostic reasoning, documentation, and order entry. By analyzing learner interactions over time, agents can identify recurrent misconceptions, recommend targeted resources, and scaffold increasingly complex cases. In clinical environments, agents can also assist with just-in-time teaching by surfacing concise, guideline-aligned explanations and evidence summaries at the bedside or within the EHR interface, supporting continual professional development without disrupting workflow [268; 190].

### 1.2.4 Making Medical Agents Trustworthy: Governance, Ethics, and Regulation

Because Medical Agents operate in safety-critical, adaptive settings, governance must be treated as a continuous operational function rather than a one-off approval step. Four pillars are central [23] : **Accountability**, with clear ownership of AI-enabled decisions and failure management; **Transparency**, via explanations and audit trails that allow clinicians and regulators to understand recommendations; **Fairness**, through monitoring and mitigation of performance disparities across populations; and **Safety**, via ongoing testing, monitoring, and fail-safes to prevent patient harm.

These principles translate into concrete obligations: formal AI governance plans, robust logging and performance monitoring to detect drift, privacy-preserving data practices and informed consent that explain how agents shape care, and contractual expectations of vendors around bias assessment, monitoring, and incident response. Under such structures, Medical Agents can be integrated into the core infrastructure of care delivery while maintaining trust among clinicians, patients, and regulators.

## 1.3 Defining the Medical Agent: A Developmental Roadmap

The successful integration of AI into healthcare requires moving beyond isolated, task-specific models toward autonomous, workflow-centric **Medical Agents**. Such agents represent a convergence of two key trends: the

architectural shift from static LLMs to dynamic, agentic systems, and the pressing need within healthcare for solutions that address administrative burden, workforce strain, and fragmented care delivery.

In this survey, we propose a *developmental roadmap* to frame the evolution of Medical Agents. This roadmap is not a rigid maturity model but a conceptual guide that outlines how these systems can progress through increasing levels of capability, workflow integration, and governance. It provides a shared vocabulary for clinicians, engineers, and policymakers to situate emerging technologies and anticipate future challenges. The evolution can be understood across three interconnected levels.

### 1.3.1 Level 1: Knowledge-Centric Assistance

At the foundational level, Medical Agents function as **knowledge-centric assistants**. Their core purpose is to augment the clinician's access to information, acting as sophisticated, interactive medical encyclopedias that can synthesize and rephrase information on demand.

**Capabilities and Functionality.** These agents are primarily reactive, responding to explicit user prompts with information derived from a defined knowledge base. Their value lies not just in retrieval, but in synthesis. Main capabilities include:

- **Information Retrieval and Synthesis:** Answering complex clinical questions by drawing from a vast corpus of curated sources, including medical textbooks, the latest clinical guidelines, and biomedical literature. For example, a query like, "What are the guideline-recommended second-line therapies for triple-negative breast cancer in premenopausal women with BRCA mutations?" would yield a synthesized answer, not just a list of documents.
- **Clinical Summarization:** Generating concise, context-aware summaries of lengthy patient records. An agent could produce a one-page summary of a patient's multi-year history, highlighting major events, active problems, and recent trends, tailored for a specific specialist consultation.
- **Patient Communication:** Drafting patient-friendly explanations of complex diagnoses or treatment plans. This improves health literacy by translating medical jargon into understandable language (e.g., explaining the risks and benefits of a planned procedure).
- **Administrative Communication Drafts:** Assisting with routine communications, such as drafting initial responses to patient portal messages about prescription refills or appointment scheduling, for clinician review and approval.

Technologically, these agents rely heavily on Retrieval-Augmented Generation (RAG) to ground their responses in factual data and minimize hallucination. The quality of the underlying knowledge base and the mechanisms to keep it current is paramount. However, they lack persistent, longitudinal memory of patient states and cannot perform multi-step, proactive reasoning.

**Integration and Governance.** Integration into clinical workflows is typically minimal and often asynchronous. These agents operate as "side-car" applications or plugins, requiring clinicians to manually copy and paste information between the agent and the Electronic Health Record (EHR). While this lowers the barrier to deployment, it creates a disjointed user experience. Governance at this level is **content-centric**, focusing on the quality and safety of the information provided. Key priorities include ensuring factual accuracy through high-quality knowledge sources, providing clear source citations for every claim, robustly mitigating the risk of "hallucinated" information, and transparently communicating the agent's limitations. The clinician is always expected to exercise full independent judgment, using the agent as a reference tool.

### 1.3.2 Level 2: Workflow-Integrated Decision Support

The second level marks a significant evolution, as agents become **workflow-integrated decision support tools** that are deeply embedded within the clinical environment and can reason over live patient data.

**Capabilities and Functionality.** Agents at this level move from being reactive to proactive, analyzing patient data in real time to provide context-aware recommendations. They act as vigilant co-pilots. Key capabilities include:

- **Proactive Monitoring and Risk Stratification:** Continuously monitoring a patient's trajectory (e.g., streaming vital signs, new lab results, nursing notes) to detect early signs of deterioration. For instance, an agent could identify a subtle combination of factors pointing to incipient sepsis and suggest a guideline-concordant bundle of orders.
- **Dynamic Differential Diagnosis Support:** Ingesting a patient's presenting symptoms, history, and initial test results to generate and dynamically update a ranked differential diagnosis, helping to prevent premature diagnostic closure.
- **Personalized Therapeutic Guidance:** Recommending treatment adjustments based on patient-specific factors. An agent might suggest a dose adjustment for a renally-cleared drug in response to a new creatinine result, or highlight a potential drug-gene interaction based on pharmacogenomic data.

This requires a major technical leap, including robust, low-latency data integration (e.g., via FHIR APIs), temporal reasoning models (like Transformers for time-series) that can handle irregularly sampled data, and multimodal fusion techniques to synthesize insights from text, images, and structured EHR data. While they can propose actions, they do not execute them autonomously.

**Integration and Governance.** At this level, agents are integrated directly into the EHR, surfacing recommendations and insights "in-flow" as part of the ambient user experience. Governance must mature to become **clinically-focused**, addressing the direct influence these agents have on decisions. This involves rigorous, ongoing validation on local, representative patient data to monitor for model drift and ensure reliability and equity. It also demands strong explainability (XAI) features that allow clinicians to quickly understand the rationale behind a recommendation (i.e., "Why is the agent suggesting this?"). Questions of liability and regulatory classification become critical, as these tools are often considered Software as a Medical Device (SaMD) requiring formal evidence of safety and effectiveness.

#### 1.3.3 Level 3: Semi-Autonomous Workflow Execution

At the highest level of development, Medical Agents become **semi-autonomous systems that orchestrate and execute complex clinical and administrative workflows**, offloading significant cognitive and procedural work from clinicians.

**Capabilities and Functionality.** These agents are goal-driven and can autonomously plan and execute multi-step tasks, with humans providing strategic oversight and handling exceptions. Their capabilities are transformative:

- **End-to-End Administrative Process Automation:** Managing entire administrative workflows. For prior authorizations, an agent could identify that a procedure requires one, gather the relevant clinical notes and lab results from the EHR, populate the payer-specific form, submit it via an API, and monitor for a response, only alerting a human if the request is denied or requires additional clinical input.
- **Clinical Pathway Orchestration:** Managing complex care pathways over time. For discharge planning, an agent could coordinate among nursing, pharmacy, and physical therapy; schedule the necessary follow-up appointments and transmit the discharge summary to the primary care provider; and schedule automated check-in calls with the patient post-discharge.
- **Active Documentation Scribe:** Moving beyond summarization to become an active scribe that listens to the physician-patient encounter, intelligently synthesizes information from the EHR in real-time (e.g., pulling up relevant past labs), and generates a structured, billable note for the clinician to review, edit, and sign.

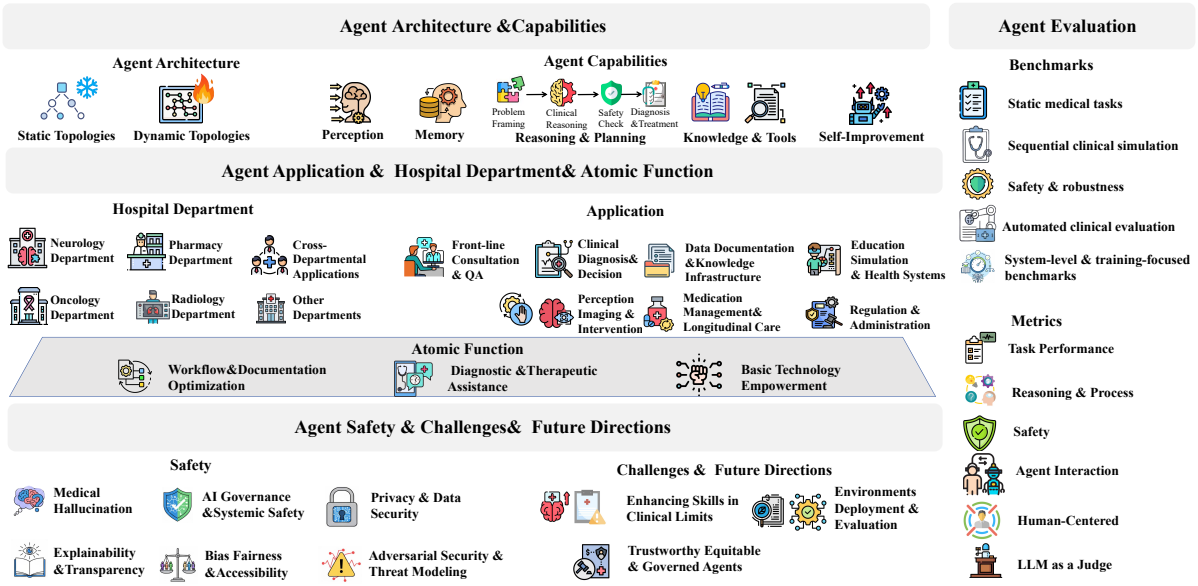


Figure 1: A conceptual framework for medical agents.The framework illustrates the flow from Agent Architecture and Capabilities, forming the basis for Atomic Functions, which underpin Agent Applications in Hospital Departments, while addressing Safety, Challenges, Future Directions, and Evaluation.These aspects have constructed a continuously evolving ecosystem for medical agents in the field of healthcare.

This requires a sophisticated architecture of chained tool use, dynamic planning and replanning, and sandboxed function-calling capabilities that securely interact with hospital systems while enforcing safety constraints.

**Integration and Governance.** These agents form an enterprise-wide **orchestration layer**, connecting disparate systems and coordinating actions across departments. Governance at this stage must be **process-centric and procedural**. This requires mature platform engineering with comprehensive, immutable audit trails to ensure every agent action is logged and attributable. A robust human-in-the-loop (HITL) framework is essential, with clearly defined rules for which actions can be fully automated (e.g., scheduling a routine follow-up) versus those that require explicit human approval (e.g., ordering a medication). The system must include clear "off-switches" and escalation pathways. This is the stage where organizational policy, clinical leadership, and regulatory engagement must be in full alignment to ensure that this powerful autonomy is deployed safely, ethically, and effectively.

### 1.4 Navigation

To better understand how medical agents function and the avenues for effective navigation within healthcare systems, we introduce a conceptual framework for medical agents, illustrated in Figure 1. This framework outlines the process of medical agents in the medical ecosystem, covering aspects from technology to application and evaluation, as well as challenges and future prospects. The rest of this survey can be read in four passes. First, Section 2 characterizes Medical Agents from a *model capability* perspective, decomposing agentic systems into planning, tool use, memory, self-improvement, reasoning, perception, and long-horizon control. Second, Section 3 examines *system designs*, while Sections 4 to 6 organise Medical Agents by *tasks, hospital departments, and applications* along the care pathway, from triage and consultation to longitudinal management, documentation, and back-office operations. Third, Section 7 focuses on *safety and governance*, and Section 8 develops a complementary framework for *evaluation*, covering static tasks, sequential clinical simulations, safety and robustness, automated clinical assessment, and system-level performance. Finally, Section 9 outlines open challenges and research directions.The overall structure is illustrated in Figure 2.

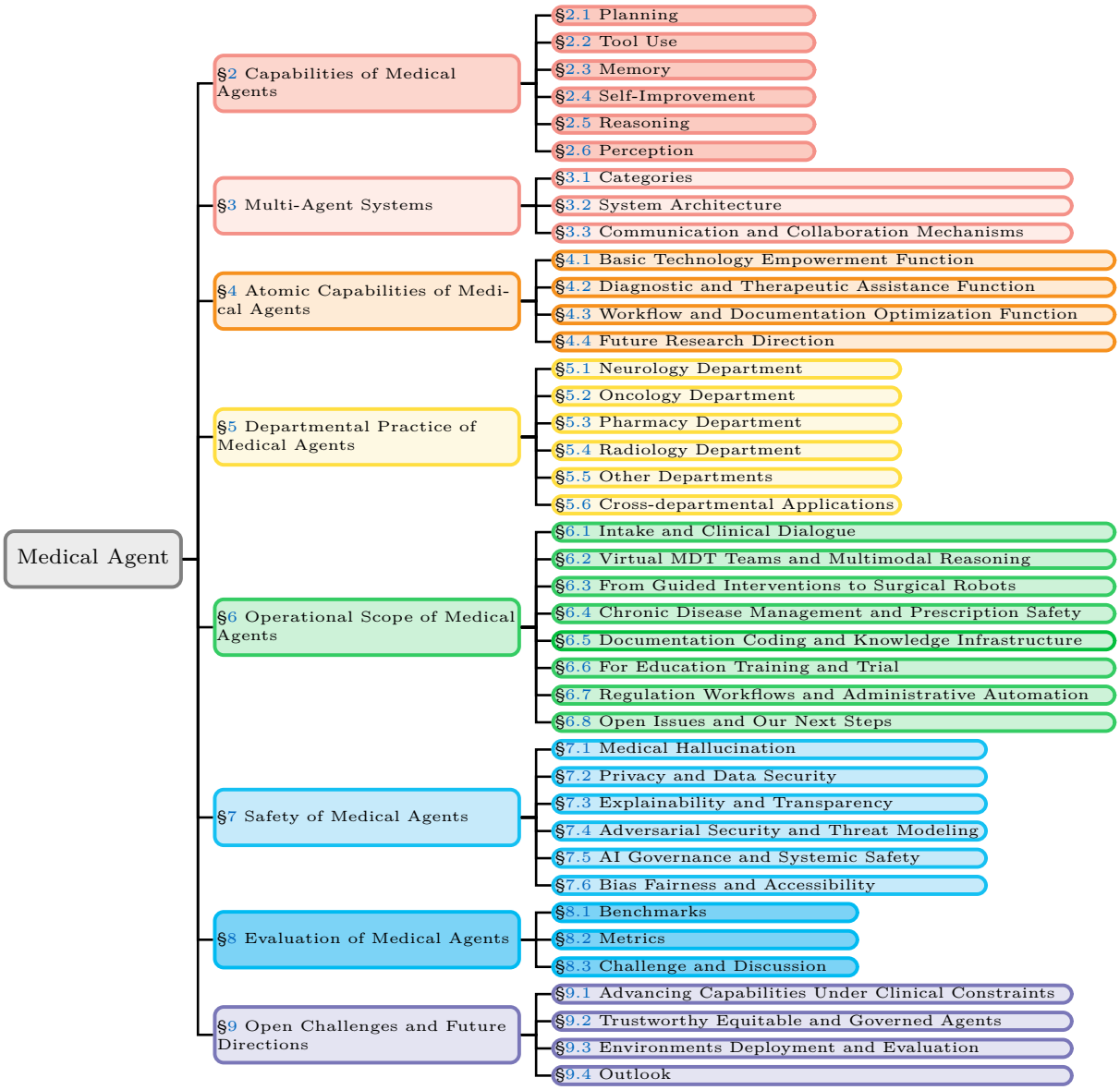


Figure 2: A comprehensive hierarchical categorisation of medical agent,illustrated with selected representative works.

## 2 From Workflows to Agentic Medical LLMs: Planning, Tool Use, Memory, Self-improvement, Reasoning and Perception

Building on the maturity levels outlined in Section 1, this section moves further to discuss how medical agents evolve in terms of capabilities. Large language models are reshaping clinical AI by shifting from fixed workflows to interactive systems that can plan, act, and learn within medical environments. Rather than a single chatbot, recent work treats the model as an agent that coordinates tools, memory, and feedback under regulatory and organizational constraints. This survey outlines this shift and argues that hybrid architectures, where agentic decision making is embedded in explicit workflows and safety protocols, are especially suitable for healthcare.

We group systems into **Three Paradigms**: *Workflow* based designs implement deterministic pipelines mirroring processes such as radiology reporting or guideline selection; the language model plays a local role,

| Dimension | Workflow-based | Agentic | Hybrid |
|---|---|---|---|
| **Planning** | Fixed pipeline [105; 357]<br>Hand-crafted steps [117; 271]<br>Protocol as state machine [186] | Dynamic sub-tasks [92; 146]<br>Multi-agent coordination [166]<br>Sense–Plan–Act–Reflect [225; 227] | Protocol skeleton [294; 125]<br>Mandatory checkpoints [193]<br>Agents fill in local details [270] |
| **Tool Use** | Fixed tool order [97; 286; 260]<br>RAG pre-wired [331; 236]<br>Rule-based API calls [305] | Tools as actions [92; 146; 61]<br>ReAct-style loops [276]<br>Learned when/what to call [139 | Safety tools hard-coded [233]<br>Other tools free [193; 49]<br>Judge inspects traces [49] |
| **Memory** | Short context [312; 2]<br>Slot-filling state [125; 132]<br>Read-only KB/ontology [164; 233] | Explicit read/write memory [232]<br>Episodic + patient history [193]<br>Experience library [243; 260; 292] | Shared structured store [271]<br>Workflow writes core facts [143]<br>Agents summarize [150; 140; 193] |
| **Self-improvement** | Offline rule updates [47]<br>Batch retraining [295; 55] | Reflection and peer review [244]<br>RL in simulators [232; 139; 57]<br>Self-training from traces [35] | Multi-level loop: language, policy, and workflow structure updated from feedback [17; 64] |
| **Reasoning** | Template CoT [33]<br>Staged checklist [18; 363; 260]<br>Scripted multi-agent flow [274] | Multi-round reasoning [92; 314]<br>Tool and memory-grounded [198]<br>Fast/slow modes [139; 204; 64] | Workflow enforces structure[268]<br>Agents optimize local decisions and explanations [88] |
| **Perception** | Separate perception models [235<br>Outputs converted to text [353]<br>LLM passive on raw signals [20] | Perception as tools [351]<br>Active viewing/zooming [301]<br>RL for "where to look" [184] | Protocol defines key models [270]<br>Results go to shared memory [150] |
| **Others (continual learning, uncertainty)** | Manual KB or guideline updates[268]<br>Simple risk flags [47; 295; 116] | Incremental KB updates [198]<br>Uncertainty triggers reflection Or handoff to humans [90; 112] | Versioned fact commons [271]<br>Uncertainty drives slowdown Or escalation to humans [193; 135] |

Table 1: Comparison of workflow-based, agentic, and hybrid agent paradigms across key design dimensions.

for example summarizing notes, while human designed logic governs overall control and safety checks. *Agentic* designs give one or more models central control: agents decompose tasks, select tools and memories, and iterate through cycles of planning, acting, and reflecting. *Hybrid* designs retain explicit workflows and rule based safety constraints but delegate local planning and decisions to agents, for example letting an agent choose question order or dispatch predefined sub workflows.

We compare these systems along **Six Capabilities**: *planning, tool use, memory, self improvement, reasoning, and perception*, which together describe how tasks are decomposed and ordered, how external tools are integrated, how short and long term knowledge is stored and retrieved, how behavior is updated over time, how structured reasoning limits hallucinations, and how multimodal signals such as images and waveforms are interpreted and grounded. In the following subsections we analyze each capability separately, focusing on the specific design choices and evaluation patterns that arise in medical settings.

## 2.1 Planning: From Static Pipelines to Protocol-Aware Agents

Planning determines how an agent moves from an initial clinical query to a complete diagnostic or management plan. The medical literature exhibits three broad patterns: static workflows, fully agentic planning, and hybrid protocol-aware planning.

### 2.1.1 Workflow-based planning

Workflow-based systems hard-code task decompositions that align with specific clinical pathways. A typical chest X-ray system follows a pipeline such as "parse free-text report → map findings to controlled concepts

→ fill a structured template → generate a standardized report". Protocols such as the ABCDEF schema for chest radiograph interpretation are compiled into fixed sections, and the LLM is used to extract and normalize concepts or generate section text.

**Strengths and limitations of static workflows.** Pipelines are easy to audit: each stage has a clear specification and can be validated independently against guidelines or expert annotations. Alignment with existing departmental workflows (e.g., standardized radiology reporting, OSCE checklists) eases integration into hospital systems, and hard-coded pipelines can block certain unsafe behaviors, such as free-form orders for high-risk medications. However, static workflows are brittle. Changes in upstream data or target tasks (new report templates, updated guidelines, different specialties) require engineering redesign. Pipelines rarely include mechanisms to detect and correct upstream errors, so mis-extracted facts propagate silently, and non-linear clinical reasoning which involves considering alternative hypotheses or revisiting assumptions in light of new evidence is difficult to express in purely forward pipelines.

**Workflow-style systems and evaluations.** Recent medical agent systems instantiate such workflow-based planning across a wide range of tasks. Remote monitoring and OS-style orchestrators such as REMONI and MedicalOS encode tool APIs and patient-flow graphs that structure data acquisition, summarization, and alerting around predefined care pathways [105; 357]. Triage, reception, and outpatient guidance agents like PIORS similarly script information gathering, eligibility checking, and routing before handoff to clinicians [19]. Many multi-agent decision-support frameworks, including MDAgents, MedAgents, MedAide, Adaptive Reasoning and Acting in Medical Language Agents, EHRAgent, AGENTiGraph, clinically-inspired multi-agent transformers for disease-trajectory forecasting, and SmartState, decompose clinical tasks into stages (e.g., history taking, hypothesis generation, evidence retrieval, consensus formation) and bind each stage to specialized agents or tools [133; 132; 260; 292; 57; 243; 346; 196; 13]. Pharmacy- and documentation-oriented platforms such as RxLens, Rx Strategist, and AI scribe systems (e.g., Sporo AI Scribe) implement multi-stage pipelines for OCR, entity extraction, safety checks, and structured note generation [117; 271; 142]. Evaluation pipelines mirror this workflow-centric view: agent-based uncertainty-aware radiology labeling, reinforcement-learning-based validation of ophthalmic VQA, and benchmark datasets like MedMCQA are organized as fixed sequences of parsing, reasoning, and scoring modules [20; 74; 205]. Framework papers on evaluating LLM agents in the clinic and designing agentic workflows for patient-friendly report generation emphasize aligning task graphs with clinical pathways and safety requirements [186; 254], while conversational health agents for well-being and chronic-disease self-management embed scripted flows for triage, education, and escalation even when using free-form dialogue models [312; 2]. Generic multi-agent search and spoken virtual-patient systems similarly hard-code dialogue phases and information-search steps, reflecting workflow-style planning under the hood despite using LLM components [60; 258].

### 2.1.2 Agentic planning: dynamic task decomposition and coordination

Agentic planning abandons fixed decompositions in favor of dynamic planning within a learned policy. Given a complex task (e.g., long-term follow-up management, multimodal workup of a rare disease), an agent generates sub-tasks and an execution order, chooses tools per sub-task, and may revise the plan based on intermediate results. Recent medical agent frameworks instantiate this pattern across visual reasoning, multimodal perception, retrieval-augmented question answering, and treatment planning [92; 276; 198; 91; 146; 54; 224; 316; 318; 55; 293; 28]. Several recurring motifs appear in medical agent planners.

**Dynamic task graphs.** Rather than a linear pipeline, agentic systems construct task graphs on the fly. A planner might generate nodes corresponding to "clarify chief complaint", "screen red-flag symptoms", or "review prior imaging", with edges encoding dependencies. Execution agents then expand nodes into specific questions, tool calls, or summary steps. This resembles hierarchical task networks but is instantiated via LLM prompting. Dynamic decompositions underlie planning-heavy agents for radiotherapy and focused ultrasound treatment [314; 344], autonomous navigation and image-quality control in interventional workflows [225; 179; 227], and cohort/feature extraction over electronic health records [143; 243; 287].

**Multi-role collaboration.** Many systems mimic multidisciplinary teams: a planner agent decomposes tasks, specialist agents reason within their domains, and a judge or moderator resolves disagreements. Separate agents may handle cardiology, oncology, and pharmacy; the planner decides when to consult each, and the judge aggregates their rationales into a final plan. This multi-agent structure allows planning to incorporate diverse domain expertise and to cross-check critical steps. Yuan et al. show that *mixed-vendor* multi-agent teams, where doctor agents are instantiated from different model families, consistently outperform single-vendor teams on rare and complex diagnoses by pooling complementary inductive biases that homogeneous teams collectively miss [329]. MedCollab further strengthens collaborative planning by adopting an IBIS-structured argumentation protocol and constructing Hierarchical Disease Causal Chains that transform flattened diagnostic predictions into structured models of pathological progression [334]. Multi-agent simulators and clinical benchmarks make these roles explicit in hospital- or clinic-scale environments, where doctor, patient, nurse, and tool-using agents interact under realistic constraints [63; 63; 150; 232; 166; 330]. Safety- and communication-oriented constellations further specialize agents for monitoring, coaching, and bias mitigation within the planning process [193; 112; 129; 175]. Knowledge-centric designs extend this idea toward adaptive collaborations and hierarchical teams, using knowledge graphs or role-specialized LLMs to structure the decision process [299; 224; 363]. Beyond virtual clinicians, other work instantiates distinct "doctor", "nurse", "regulator", "researcher", or educational copilots, including Doctor-R1, NurseLLM, Agentic-AI Healthcare, dual-agent nursing robots, Medco, OpenLens AI, and regulator-manufacturer simulators [139; 97; 237; 349; 291; 45; 94], often coupled with experiential learning mechanisms or longitudinal research workflows that further refine planning policies [341; 362; 257; 45].

**Protocol-aware planning.** Agentic planners in medical settings rarely operate in a vacuum. Instead, they embed protocols as soft or hard constraints. A planner may be required to ensure that all guideline-mandated diagnostics for a given cancer stage are considered before recommending therapy, or to obey a fixed high-level sequence (initial assessment, differential diagnosis, investigations, management) while filling in details adaptively. In some systems, an explicit protocol graph or checklist is maintained by a symbolic orchestrator, and the LLM planner only chooses local actions under that skeleton. Task-specific agents hard-code oncology guidelines or procedural pathways into their planning skeletons, such as NCCN-based breast cancer planning and radiotherapy pipelines [190; 314], while automatic-evaluation frameworks derive protocol graphs directly from standardized-patient pathways or causality-aware diagnostic progressions [168; 172; 164; 100]. Other agents bind planner actions to external calculators, tools, and institutional databases through nested tool calling and RAG-style interfaces [359; 143; 243; 225], and hospital-scale simulators embed triage and consultation protocols as symbolic or partially scripted flows for agent evaluation [63; 63; 232; 166; 150]. Emerging "next-generation" medical agents align planning heuristics with tool-augmented reasoning styles (e.g., o1-like deliberate planning) to improve stability and safety in complex scenarios [307].

**Benefits and risks.** Agentic planning offers significant flexibility: agents can adapt to novel case types, reorder questions to improve patient rapport, and tailor follow-up to longitudinal histories. Yet planning quality is hard to guarantee. Common issues include local optima (prematurely committing to an incomplete workup), omission of critical investigations, and unnecessary tool calls that increase cost and risk. Addressing these issues requires coupling planning with explicit uncertainty estimates and process-level auditing. Recent work moves in this direction via trust-verification and auditing layers around RAG and knowledge-graph agents [198; 224; 45], agent-based tools for preference-aligned data collection and simplification [55; 175], and large-scale evaluations of conversational safety and clinical competence across multi-agent constellations [193; 112; 129; 63; 168]. Taken together, these systems suggest that dynamic planning is not only a mechanism for solving single cases, but also a substrate on which safer, more auditable medical agents can be trained from simulated and real clinical experience [139; 362; 257; 316; 318].

#### 2.1.3 Hybrid planning: protocol skeletons with agentic detail

Hybrid planning uses predefined workflows to anchor safety and compliance, and agentic components to adapt within those boundaries. Designers typically encode a guideline or care pathway as a high-level sequence of stages and required checkpoints, use deterministic code to ensure that all mandatory stages are visited and

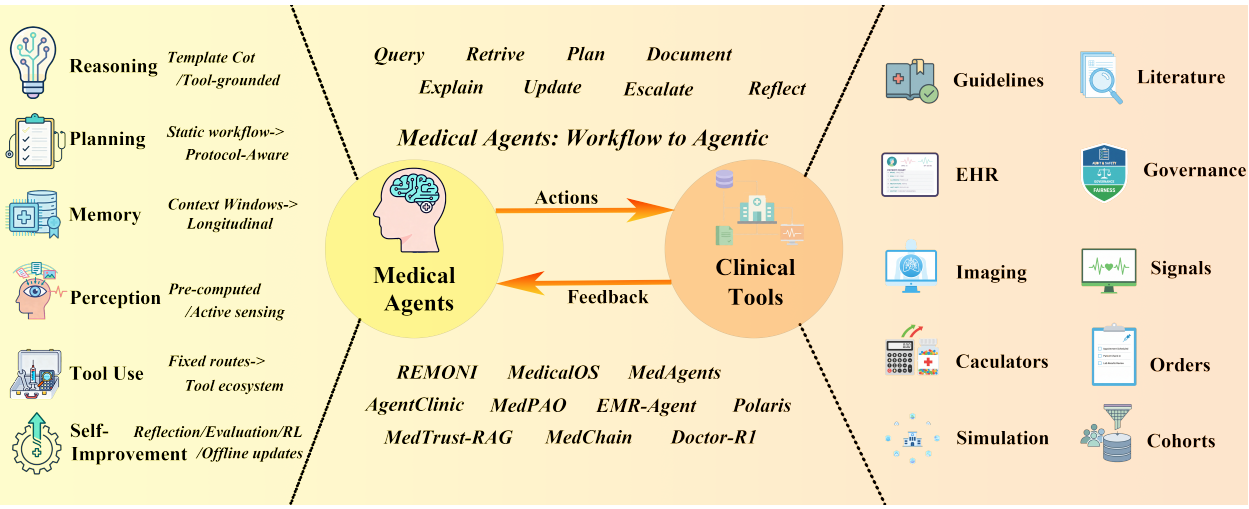


Figure 3: Medical agents with capabilities in reasoning, planning, memory, perception, tool use, and self-improvement interact with clinical tools (e.g., EHR, imaging, calculators, orders) via iterative action–feedback loops, enabling tool-calling agents that carry out and refine complex medical tasks.

that safety checks (e.g., allergy review before prescribing) are enforced, and then delegate within each stage to an agentic planner to select questions, choose tools, and tailor actions to the patient.

**Applications of hybrid planning.** This pattern appears in systems that simulate chronic care management, surgical co-pilots, multi-agent multidisciplinary consultations, standardized cancer patients, and professional medical Q&A services, where schemas or care pathways define the outer loop and agents handle the inner loop [294; 125; 156]. In surgical settings, for example, the environment can embody the protocol (allowed next actions given the current operative step), while an agentic co-pilot reasons about which knowledge to retrieve and what advice to surface [294]. Report-structuring and conversational agents increasingly adopt the same design. MedPAO encodes chest X-ray interpretation protocols as a Plan–Act–Observe loop, enforcing coverage of required views while delegating concept extraction and tool selection to an agentic controller [270]. Polaris organizes a primary nurse-like agent and multiple specialist safety agents around explicit checklists and escalation rules, so that protocol conformance is handled by the orchestration layer while LLMs supply localized reasoning [193]. Xu et al. show that diagnostic frameworks such as CoD, MedAgents, and AgentClinic also follow a staged pipeline (symptom gathering, candidate generation, consensus) where the workflow is fixed but o1-based agents reason within each stage and decide when to retrieve knowledge or call tools [307].

**Control plane vs. reasoning plane.** Hybrid planning is particularly appealing in healthcare because it mirrors how clinicians operate: high-level pathways are standardized, yet individual cases require local adaptation. Technically, hybrid designs force an explicit separation of the control plane, which enforces protocols, and the reasoning plane, in which LLM-driven decisions occur. This separation simplifies verification and opens the door to learning local policies (e.g., via reinforcement learning) inside fixed protocol graphs.

### 2.2 Tool Use: From Scripted Workflows to Tool Ecosystems

Tool use is central to medical agents because clinically relevant information and capabilities are exposed through external systems: literature databases, guidelines, knowledge graphs, EHRs, imaging backends, calculators, and institution-specific APIs. We observe a progression from tools wired into fixed pipelines, to tools surfaced as first-class actions for agents, to hybrid schemes where some tools are always invoked and others are under agentic control.

### 2.2.1 Workflow-based tool use

In traditional pipelines, tools are invoked at predefined locations. A common pattern for question answering is "retrieve from a literature database or guideline repository → summarize with an LLM → output structured answer". Retrieval-augmented generation (RAG) is used behind the scenes: retrieval results are concatenated into the prompt, but the LLM has no control over *when* or *whether* to retrieve. Similar patterns appear for knowledge graphs and domain APIs: given recognized entities, a rule-based process queries medication databases, imaging analysis services, or ontologies, and the LLM paraphrases or aggregates results. Tool usage is implicitly determined by pipeline design rather than by the model.

This design eases integration with existing hospital systems: each tool interface can be validated independently, and only a small set of tool combinations is possible. However, rigid tool sequences limit coverage on complex cases: the system cannot decide that additional imaging is needed, that a different knowledge base is more relevant, or that a calculator should be called again after new information arrives. Supporting new tools or tasks often requires redesigning pipelines and re-validating all tool call paths.

Many systems labeled as "agentic" still follow this workflow-style tool routing. For decision support and diagnosis, NurseLLM, AMANDA, MoMA, KERAP, AgentMD, MDAgents and MedAgents [97; 286; 80; 305; 123; 132; 260] orchestrate fixed stages of retrieval, calculator use, and EHR access; agents mostly fill in text around a predetermined sequence of tool calls. Workflow-style control in causality-aware diagnosis agents, triage tools, and protocol-based CDSSs [164; 93; 131; 13; 269] similarly prescribes which tools are invoked at each stage.

Imaging and documentation pipelines, including interactive radiology systems, labeling agents, and scribes [257; 20; 331; 236; 142; 112; 281; 223], usually treat perception models, OCR engines, and style checkers as fixed subroutines: the LLM does not choose which perception tool to run or when to rerun it. Safety- and evaluation-focused workflows such as Polaris, tiered agentic oversight, GuidelineGuard, AgentClinic, AI Hospital, and spoken virtual patients [193; 135; 233; 232; 63; 258; 168; 55] also hard-code tool use by specifying which guideline checkers, retrieval modules, and scoring tools are invoked in each evaluation step.

### 2.2.2 Agentic tool use: tools as first-class actions

Agentic tool use exposes tools as callable APIs that the LLM can select, sequence, and combine. In ReAct-style frameworks, the model alternates between natural language reasoning and tool calls, deciding at each step whether to act and which tool to invoke; medical agents adopt the same abstraction across diagnosis, treatment planning, education, administration, and knowledge curation [92; 276; 146; 61; 71; 314].

In this setting, tools are typically organized into four broad categories:

- **Knowledge tools**, which provide vector search over guideline or literature corpora, role-specific knowledge bases, fact-checking and uncertainty estimation, or agentic knowledge graphs [198; 224; 166; 318; 49; 363].
- **Clinical calculators**, which implement dosing calculators, risk scores, and severity indices, sometimes accessed via specialized calculator bridges such as MeNTi/Menti and related frameworks [359; 341].
- **Perception tools**, which wrap imaging classifiers and segmenters, ECG and waveform analyzers, and lab normalization services, allowing multimodal agents to call radiology, pathology, and neuroimaging models as tools [92; 146; 61; 71; 109; 294].
- **System tools**, which expose EHR query APIs, scheduling systems, order-entry interfaces and audit logs, as seen in EMR and EHR agents, pharmacy platforms, reporting workflows, and privacy-preserving deployment layers [143; 243; 117; 271; 83; 254; 237; 270].

Concrete frameworks illustrate how agents learn to navigate these tool ecosystems. Visual and multimodal agents such as Med-VRAgent, Proactive MLLMs, MMedAgent, MedRAX, radiotherapy planners, oncology decision agents, and emerging lightweight offline agents like Meissa [92; 276; 146; 61; 314; 71; 44] select among perception tools, guideline retrieval and calculators to answer complex queries and propose treatment

options. Newer systems push beyond static tool sets: MACRO autonomously discovers composite multi-step tool sequences from verified execution trajectories and registers them as reusable high-level primitives that continuously expand the agent's behavioural repertoire [62]; TEA-CXA learns the practical trustworthiness of each tool through multimodal agentic reinforcement learning, resolving conflicting tool outputs by empirically accepting or rejecting responses per query type [111]; and MedScope performs coarse-to-fine evidence seeking over long-form clinical videos by interleaving intermediate reasoning with targeted tool calls, grounding predictions in temporally localised visual evidence [155]. For 3D volumetric analysis, 3DMedAgent coordinates heterogeneous visual and textual tools through a flexible MLLM agent, progressively decomposing complex 3D CT analysis into tractable subtasks without 3D-specific fine-tuning [290]. In pathology, LAMMI-Pathology adopts a tool-centric, bottom-up architecture in which customised domain-adaptive tools are clustered by domain style into component agents coordinated through a hierarchical planner, avoiding excessively long contexts that could induce task drift [253]. MedSAM-Agent reformulates interactive segmentation as a multi-step decision-making process, using a two-stage training pipeline that integrates multi-turn outcome verification with clinical-fidelity process rewards to promote interaction parsimony [170]. CARE decomposes multimodal medical reasoning into coordinated sub-modules, where a compact VLM proposes medical entities, an expert segmentation model produces pixel-level ROI evidence, and a grounded VLM reasons over the augmented image, with VLMs optimised via reinforcement learning with verifiable rewards [75]. Knowledge-centric systems like MedTrust-RAG, agentic medical knowledge graphs, MedChain, AgentClinic, tree-based RAG test recommenders, KG4Diagnosis and correction pipelines such as IryoNLP [198; 224; 166; 232; 318; 363; 49] focus on which knowledge tools to invoke for retrieval, verification and trust calibration. System-oriented agents including EMR-AGENT, EHRAgent, RxLens, Rx Strategist, MedPAO, patient-facing report generators and MCP-based deployments [143; 243; 117; 271; 270; 254; 237] exercise EHR, cohort-building, prescription and documentation tools as actions in their control loops.

A recurring question is *when* tools should be called and *which* tools should be preferred. Agentic systems therefore add meta-control over tool use. Confidence-aware heuristics and learned controllers can, for example, query a guideline only when internal confidence drops below a threshold, or always consult a calculator for dosing questions. RL-style approaches treat tool calls as actions in an environment, with rewards reflecting answer correctness and tool cost. Experiential RL agents such as Doctor-R1, architecture-search frameworks like Learning to Be a Doctor, and mathematical–medical agents such as Haibu's system [139; 362; 341] explicitly optimize tool-selection policies and tool-routing topologies, while MeNTi/Menti and ReflecTool [359; 162] add nested calculator calls and reflection loops that can revise earlier tool decisions. Earlier non-LLM diagnosis agents based on RL and interpretable inquiry policies [100; 172] can themselves be wrapped as callable tools, further enriching the action space.

Long clinical trajectories make tool-credit assignment non-trivial: an early retrieval or calculator choice may determine ultimate outcomes. New RL algorithms for long-horizon tool use (e.g., stepwise advantage estimation in GiGPO and SpaRL) target this setting, but medical applications remain limited. Simulation environments such as Agent Hospital, ClinicalLab, MedChain, AgentClinic, SurgBox and dual-agent nursing-robot frameworks [150; 311; 166; 232; 294; 349] provide rich tool-using environments (orders, consults, procedures) on which such controllers can be trained. Specialized pipelines, including ClinicalAgent for clinical trials, MedCoAct for doctor–pharmacist collaboration, MedAgentAudit for failure analysis, disease- and specialty-specific agents, and radiotherapy planners [330; 351; 90; 250; 109; 314], treat each interaction with registries, trial databases, or planning software as an explicit tool action. Educational and omni-assistant systems such as MedCo, MedAide and schema-guided standardized patients [291; 292; 125] similarly integrate calculators, question banks and simulated patients as tools that can be called in different orders depending on the learning goals.

Agentic tool use greatly increases flexibility and reuse: tool APIs can be shared across tasks, and agents can discover non-obvious sequences of tool calls. At the same time, it introduces failure modes such as tool misuse, error chains across tools, or unbounded exploration of unsafe tool combinations, motivating explicit constraints on which tools can be called and how their outputs are checked. Importantly, tool-using pipelines also create new pathways for demographic bias: DUCX introduces a stage-wise fairness decomposition that separates end-to-end bias in chest X-ray agents from three agent-specific sources—tool exposure bias, tool

transition bias, and model reasoning bias—showing that intermediate tool-use behaviours exhibit subgroup disparities not predictable from end-to-end evaluation alone [309].

#### 2.2.3 Hybrid tool use: safety-critical scaffolding

Hybrid frameworks combine fixed guarantees around certain tools with agentic freedom elsewhere. A common pattern is to *mandate* calls to safety-critical tools while leaving other tools at the agent's discretion. Medication recommendation systems, for example, may require querying drug–drug interaction checkers and allergy databases for every candidate regimen, regardless of the agent's internal confidence; the agent can choose additional knowledge tools as needed but cannot bypass these checks. Imaging co-pilots may be required to invoke specific segmentation or staging models before issuing structured reports.

Evaluation and oversight agents themselves constitute a second class of "meta-tools". A separate LLM (or set of LLMs) inspects tool-call traces, intermediate states, and proposed actions to assess correctness, completeness, and adherence to predefined tool-use policies. If the trace violates criteria (e.g., omitting a mandatory safety tool or exceeding a cost budget), the evaluation agent can trigger localized retries or discard outputs. This LLM-as-a-judge pattern, widely studied in code generation and mathematical reasoning, is increasingly applied to tool auditing in medical systems.

Finally, deployed systems often mix static and agentic tool pathways. Low-risk informational tasks are routed through simple RAG pipelines with a fixed retrieval tool, while higher-stakes or more complex requests are delegated to agents with broader tool access. Triage heuristics or meta-models choose which tool regime to use. Safety-focused constellations such as Polaris, debiasing and auditing frameworks, pharmacy and prescription-verification agents, and early multi-agent information-search systems [193; 129; 90; 49; 117; 271; 60] exemplify this pattern: calls to critical checkers and knowledge bases are hard-wired, while more exploratory tool use (e.g., additional retrieval or explanatory tools) is left under agentic control.

### 2.3 Memory: From Context Windows to Longitudinal Clinical Memory

Memory in medical agents spans multiple timescales and forms: immediate conversation context, long-term patient histories, structured medical knowledge, and experience repositories of past reasoning episodes. We observe a shift from implicit, fixed memory in workflow systems to explicit, controllable memory modules in agentic and hybrid designs.

#### 2.3.1 Workflow-based memory

Pipeline systems typically handle memory in two ways. **Short-term context.** Dialogue systems and report generators pass recent turns and extracted entities to subsequent modules by concatenating them into prompts or structured feature vectors, such as fillable slots for symptoms and risk factors. This design suffices for short encounters but fails for longitudinal care, where relevant information spans many visits and modalities. **Read-only knowledge.** Guidelines, ontologies, and static knowledge graphs act as immutable memory. Workflow components query them via symbolic rules or retrieval modules, but the knowledge itself does not evolve online; updates require manual curation and system redeployment. These designs lack mechanisms for retaining and exploiting past interactions or adapting memory structures to new tasks, and errors in read-only knowledge (e.g., outdated guidelines) cannot be corrected by agents themselves.

#### 2.3.2 Agentic memory: explicit read and write mechanisms

Agentic systems treat memory as a first-class resource. Agents decide when to write to and read from memory, and memory is often organized in multiple layers. Recent medical multi-agent systems, from hospital-scale simulators and patient generators to consultation copilots, explicitly expose such memory modules as part of their architectures [150; 140; 232; 291; 222; 193].

**Episodic memory within encounters.** Within a single encounter, agents maintain a structured state: history of questions and answers, hypotheses and their probabilities, tool results, and safety flags. Rather than passing raw conversation text, systems store normalized representations (e.g., JSON objects for symptom

status or lab values) that multiple agents can share. This enables cross-agent coordination and supports auditing, as seen in simulated clinics and virtual standardized patients where each case is represented by a rich interaction schema [232; 125; 291]. Multi-agent consultation frameworks further maintain shared episode state across collaborating specialists, allowing agents to exchange partial assessments and intermediate plans within a single case [260; 292; 42; 222].

**Long-term patient memory.** In longitudinal settings, agents maintain per-patient memory of prior visits, diagnoses, medication changes, and patient preferences. Entries may be indexed by time, condition, or clinical episode. When a patient revisits, the agent retrieves and summarizes relevant history before planning new actions. Patient simulators such as Agent Hospital and Patient-Zero build persistent trajectories for synthetic patients across many encounters, enabling agents or human learners to observe disease progression and treatment response over time [150; 140]. Rare-disease copilots and deployed consultation agents similarly track longitudinal histories to support continuity of care and specialist hand-offs [42; 222; 193].

**Experience repositories.** Several medical agents cache successful reasoning traces, tool-call sequences, or code snippets as reusable experiences. When facing a new case, the agent retrieves similar past traces and adapts them. This case-based memory parallels classical case-based reasoning and recent "program memory" approaches in EHR query agents. EHRAgent, for example, maintains a library of successful multi-table query programs and replays or adapts them for new information needs [243]. AgentEHR challenges agents to execute complex decision-making tasks directly within raw, high-noise EHR databases and introduces RetroSum, a retrospective summarisation mechanism that dynamically re-evaluates interaction history to prevent long-context information loss and ensure unbroken logical coherence, coupled with an evolving experience strategy that bridges domain gaps [163]. TARSE separates retrievable resources into *skills* (reusable clinical procedures such as guidelines and pharmacologic mechanisms) and *experience* (verified reasoning trajectories indexed by step-level transitions), performing lightweight test-time adaptation to align intermediate reasoning with clinically valid logic [275]. Collaborative consultation frameworks such as MedAgents, MedAide, MEDCO, and RareAgents also treat past multi-agent discussions and expert reports as exemplars that can be revisited for similar cases [260; 292; 291; 42].

**Memory-space adaptation and structured clinical state.** Recent work further demonstrates that adaptation can be shifted entirely from weight space to memory space. MemSeg-Agent conditions a fixed backbone with lightweight static, few-shot, and test-time working memories dynamically composed by an agentic controller, enabling few-shot learning, federated supervised learning, and test-time adaptation within a single architecture while substantially reducing communication overhead in federated settings [32]. TRACE introduces a dual-memory architecture for temporal reasoning over streaming EHRs: a static Global Protocol encoding institutional clinical rules and a dynamic Individual Protocol tracking patient-specific state, with four agentic components (Router, Reasoner, Auditor, Steward) coordinating over this structured memory to maintain bounded inference cost and auditable reasoning traces [220]. SELSM distils simulated clinical trajectories into entity-agnostic operational rules within an abstract skill space, using a Query-Anchored Two-Stage Retrieval mechanism to dynamically fetch logical priors that guide step-by-step reasoning for locally deployable agents on FHIR-based EHR tasks [317].

**Knowledge memory and writable knowledge bases.** Structured knowledge graphs and vectorized literature corpora are increasingly treated as memory whose organization can evolve. Agents may add edges to a knowledge graph when discovering new associations (subject to validation), or update relevance scores in vector stores based on usage and outcomes. KG4Diagnosis explicitly couples a hierarchical team of diagnostic agents with a dynamically constructed medical knowledge graph, while earlier multi-agent search systems for medical information maintain shared indices and caches at the retrieval layer [363; 60]. Such designs blur the boundary between "knowledge base" and "working memory", as agents continuously rewrite both.

Crucially, agentic memory is *writable*. Agents can store failures and incorrect reasoning paths for later reflection, and multi-agent systems can maintain blackboard-style shared memories that record partial conclusions from different specialists. Safety-focused constellations such as Polaris and healthcare consultation agents increasingly pair these memories with governance mechanisms, including access control, audit trails, and

safety filters, to regulate which events are remembered and how they influence future behavior [193; 222]. This raises new questions: which events should be remembered, at what granularity, and under what governance?

#### 2.3.3 Hybrid memory and cross-module coupling

Hybrid systems treat memory as a shared infrastructure across modules rather than a set of isolated buffers. **Shared clinical state.** The workflow layer ensures that key clinical facts (diagnoses, procedures, allergies) and structured entities are always written to memory in standardized formats (e.g., FHIR resources). Agentic components then reorganize and compress this memory, constructing abstractions tailored to tasks such as phenotype summaries for diagnostic workups or risk trajectories for chronic disease management. Learning components mine memory for experience: RL agents can sample trajectories for policy updates, reflection agents analyze past failures, and uncertainty calibrators estimate reliability as a function of case features.

**Systems perspective.** In this view, designing memory itself becomes a central systems problem: schemas, access policies, and update rules determine how effectively downstream modules can reuse information. Clinical agent frameworks such as EMR-AGENT, which coordinates multi-agent SQL workflows over EMR databases to construct reusable, standardized cohorts and feature tables, and Rx Strategist, which couples a multi-stage prescription-verification pipeline with a shared knowledge graph and active-ingredient database, instantiate this pattern by treating cohorts, feature tables and drug knowledge as explicit memories that multiple components can query and, where appropriate, update rather than as static repositories [143; 271].

### 2.4 Self-Improvement: Reflection, Evaluation, and Reinforcement Learning

Modern medical agents are not only expected to perform well at deployment time, but to improve as they encounter new patients, institutions, and guidelines. Turning this aspiration into reality requires explicit mechanisms for using feedback from data, simulated or real environments, and human or automated evaluators without violating the stringent safety, governance, and auditability constraints of clinical practice. We structure this landscape in terms of three nested self-improvement loops operating over data, policy, and architecture.

#### 2.4.1 Workflow-based self-improvement

**Offline revisions and data annotation.** In pipeline systems, self-improvement is largely offline. Engineers collect error cases, refine rules, update prompts, or expand training datasets. Multi-agent pipelines may incorporate fixed mutual checking (e.g., a second LLM revises reports), but the structure remains static and self-improvement manifests as new system versions rather than online learning. Recent multi-agent systems for pharmacovigilance, report generation, and risk assessment, such as MALADE for ADE extraction, agentic workflows for patient-friendly imaging reports, Medical AI Consensus for radiology reporting, and Med-TAMARA for dialogue risk assessment, still generally follow this pattern, with models updated between deployments rather than continuously in situ [47; 254; 58; 295].

Some works explore automated data annotation and preference alignment. For medical dialogue, LLMs can assign preference labels or quality scores to candidate responses under multi-criteria (accuracy, safety, empathy), and RL from AI feedback (RLAIF) can train dialogue policies without exhaustive human labeling. Dou et al. systematically study LLM-based data annotation strategies for medical dialogue preference alignment, building multi-agent annotators that generate conversations and preference labels as training signals [55]. However, these methods are typically applied in batch and not yet tightly coupled to the live deployment workflow.

**Toward workflow-coupled adaptation.** Recent medical-agent systems begin to push self-improvement closer to the workflow itself. Doctor-R1 applies experiential agentic reinforcement learning in a multi-agent consultation environment, combining a two-tier reward and an experience repository so that interaction trajectories directly drive policy updates [139]. Dutta and Hsiao introduce automatic correction for doctor agents interacting with simulated patients, allowing reasoning and action policies to adapt after failed diagnoses [57]. At the system level, Zhuang et al. treat medical agents as graphs of workflow nodes and search over this architecture space, so that the diagnostic workflow itself evolves according to feedback [362].

MedChain couples an interactive, sequential benchmark of clinical cases with a MedChain-Agent that reuses prior cases via RAG and feedback signals, gradually adapting its decision-making over episodes [166]. Taken together, these lines of work suggest a continuum from static pipelines with occasional offline revisions, through agentic workflows that embed reflection, consensus and risk assessment, toward architectures where data collection, evaluation, and workflow evolution form a single closed loop of self-improvement across medical tasks [47; 254; 58; 295; 55; 139; 57; 362; 166].

### 2.4.2 Agentic self-improvement

Agentic systems push these ideas further by making self-improvement an explicit part of the agent's ongoing interaction with its environment. Rather than relying solely on offline updates, the agent participates in generating critiques, collecting feedback, and adapting its own strategies. The resulting mechanisms span reflection and critique, multi-agent peer review, reinforcement learning in simulated environments, and iterative self-training.

**Reflection and critique.** Reflection methods [244; 79] let agents inspect their own reasoning chains, compare outcomes against expectations, and revise. In medical agents, reflection may be triggered when uncertainty is high or when an evaluator flags potential guideline violations. Agents generate critiques, identify missing evidence, and replan; confidence-aware variants condition reflection on estimated reliability to avoid unnecessary cost. Recent empirical audits such as MedAgentAudit explicitly analyze collaborative failure modes in medical multi-agent systems and argue for auditable reflection over opaque deliberation, while protocol-driven agents like MedPAO and code-based agents such as EHRAgent embed these critique loops into plan–act–observe workflows grounded in clinical protocols and executable tools [90; 270; 243].

**Multi-agent peer review.** Multiple agents with different roles evaluate one another's outputs. A quality-review agent checks completeness and coherence; a safety agent checks for harmful recommendations; a coding agent verifies billing consistency. Feedback is either used on the fly (to revise outputs) or stored for offline training. This pattern mirrors peer review in clinical practice and can surface subtle errors. Frameworks such as MDAgents, MedAide, Medco, and AMANDA instantiate this pattern with moderators, specialist consultants, educator agents, and dedicated perceptual or evaluation agents, showing that structured group deliberation with external knowledge can improve diagnostic accuracy, multimodal reasoning, and educational feedback [132; 292; 291; 286].

**Reinforcement learning in simulated environments.** Several medical agents treat clinical interaction as a sequential decision problem and apply RL. Environments include simulated patients, EHR sandboxes, and full hospital simulations. Rewards combine diagnostic accuracy, resource cost, patient satisfaction, and safety penalties. For example, agents in simulated clinics learn when to ask additional questions versus when to decide, and how to trade off test ordering cost against diagnostic certainty. GRPO-style objectives and process-level rewards such as rewarding good inquiry trajectories rather than only final answers improve planning behavior. AgentClinic provides a multimodal benchmark for such training, and adaptive doctor agents built on it refine their reasoning and action policies via automatic correction after incorrect diagnoses, illustrating how RL-style updates can be layered on top of language-model agents [232; 57]. ATPO formulates multi-turn medical dialogue as a Hierarchical Markov Decision Process and proposes an uncertainty-aware Adaptive Tree Policy Optimization algorithm that adaptively allocates rollout budgets to high-uncertainty states, enabling a Qwen3-8B model to surpass GPT-4o on medical dialogue benchmarks [28]. MIND unifies inquiry and diagnosis within a single RL framework for psychiatric consultation, building a Criteria-Grounded Psychiatric Reasoning Bank and enforcing explicit clinical reasoning with rubric-based process rewards to jointly improve information acquisition and diagnostic decision-making [149]. MedSAM-Agent reformulates interactive medical image segmentation as a multi-step autonomous decision-making process, introducing a two-stage pipeline that integrates multi-turn outcome verification with clinical-fidelity process rewards [170]. Zhang et al. introduce an agentic framework for scaling medical reasoning verification, training verifiers to iteratively query external medical corpora during evaluation via tool-augmented RL, achieving substantial gains over prior reward-model baselines with an 8× reduction in sampling budget [338].

**Iterative self-training.** Agents may generate tasks, attempt them, evaluate performance, and add successful (or curated) trajectories to training sets. This self-play pattern echoes AlphaZero-style approaches and recent LLM self-training methods [17; 64], and can be adapted to medical coding games, note writing, or treatment planning, provided careful safety filtering. Self-evolving consultation frameworks and multi-agent education copilots already move in this direction by replaying multi-agent consultations or didactic interactions, curating successful trajectories, and updating specialist roles over time [35; 291]. EvoClinician introduces a "Diagnose–Grade–Evolve" loop: an Actor agent attempts a diagnosis, a Process Grader evaluates each action for clinical yield and resource efficiency, and an Evolver updates the Actor's strategy by evolving its prompt and memory at test time [103]. ExperienceWeaver shifts the focus from data retrieval to experience learning: instead of recalling past examples, it distils noisy, multi-dimensional feedback into structured Tips and Strategies that are injected into an agentic pipeline, teaching the model "how to revise" rather than "what to revise" for clinical text improvement [304]. Differential Reasoning Learning (DRL) extracts reasoning graphs from reference rationales and the agent's chain-of-thought, diagnoses graph-level discrepancies via clinically weighted graph edit distance, and stores the resulting natural-language instructions in a knowledge base that patches likely logic gaps at inference time via RAG [167].

### 2.4.3 Hybrid self-improvement

**Multi-layer feedback loops.** Hybrid systems distribute self-improvement across pipeline and agent layers. Knowledge bases, rules, and templates are updated as new evidence or guidelines emerge; agentic components refine their strategies via reflection and RL; and shared memories capture the long-term effects of those updates. For instance, a system may update its vector store with new guidelines, retrain its retrieval controller via RL to make better use of this memory, and revise its multi-agent orchestration based on error audits, as seen in recent medical frameworks that couple environment benchmarks (AgentClinic), collaboration audits (MedAgentAudit), and adaptive collaboration structures (MDAgents, MedAide, Medco) [232; 90; 132; 292; 291]. Specialized components such as AMANDA for visual question answering and EHRAgent for tabular EHR reasoning similarly evolve retrieval, tool-use, and controller policies while keeping core language models frozen [286; 243].

**Designing rewards and signals.** An important theme is multi-layer feedback. Environment-level metrics, such as end-to-end diagnostic yield and time-to-treatment, guide architecture adjustments; process-level metrics (guideline adherence, tool coverage) guide agent strategies; and per-utterance metrics (safety scores, empathy ratings) guide language style. Self-evolving consultation frameworks make this explicit by updating both agent policies and coordination mechanisms from longitudinal performance signals across levels [35]. Designing these reward signals and their interactions is an open challenge, especially when human feedback is scarce.

## 2.5 Reasoning: preventing hallucination propagation

Safe medical reasoning requires not only powerful models but also disciplined structure and an explicit treatment of uncertainty. Rather than relying on a single monolithic chain of thought, recent work decomposes reasoning into complementary modes that can be orchestrated and audited. We highlight two such modes, which underpin most agentic designs and shape how hallucinations are generated, detected, or suppressed.

### 2.5.1 Workflow-based reasoning

Many systems use template-based chains of thought: "list symptoms → generate differential diagnoses → propose tests → choose final diagnosis". LLMs fill in each step, and rule-based modules check consistency (e.g., whether the chosen diagnosis explains observed findings). CoD formalises this pattern by explicitly representing a chain of diagnoses with intermediate hypotheses and calibrated confidence distributions, yielding an interpretable medical agent [33].

**Template-based and multi-agent workflows.** Multi-agent diagnostic frameworks generalise the same workflow by distributing stages such as guideline retrieval, knowledge-graph reasoning, multimodal evidence aggregation and specialist consultation across dedicated agents, for example, MAGDA, KG4Diagnosis,

MedAgents, Zodiac, CARE-AD, MedRAX, FEAT and tiered agentic-oversight systems for safety-critical settings [18; 363; 260; 354; 152; 61; 238; 135]. Vision–language pipelines such as Proof-of-TBI, graph-augmented VLMs and agent-based uncertainty-aware radiology labelers implement similar staged workflows for imaging-heavy tasks, typically moving from perception to structured feature extraction, longitudinal context fusion and final explanation [88; 147; 20]. Template-like multi-agent conversations are also used to mitigate cognitive bias and improve reliability: frameworks for multi-specialist consultation, multi-agent debate and active-inference prompting orchestrate roles like generalist, specialist, devil's advocate and supervisor before committing to a final diagnosis [274; 129; 246].

**Interpretability, coverage, and error propagation.** Outside core diagnosis, similar agentic workflows appear in prescription verification and pharmacy operations [271; 117], multi-agent search and integration of heterogeneous medical data [60; 212; 173], RL-based controllers for interactive image analysis and timely interventions [234; 184; 74; 240], and infrastructure-focused systems such as SmartState, federated lifelong imaging, spoken virtual patients, well-being companions, fraud analysis and data-annotation agents [13; 350; 258; 312; 116; 55]. These designs improve interpretability and make auditing easier, but they still inherit upstream errors: if retrieval, perception or data-quality modules are biased or incomplete, multi-agent workflows may converge quickly on an incorrect yet well-structured explanation.

### 2.5.2 Agentic reasoning

Agentic systems employ multi-round reasoning with self-correction, and have now been instantiated across diagnosis, treatment planning, report structuring, EMR mining, guideline interpretation, simulation, and medical education [92; 314; 270; 139; 143; 225; 140; 188; 250; 97; 238; 197; 78; 18; 132; 260; 291; 292; 330; 152; 71; 289]. Agents revise hypotheses as new evidence arrives, and role-specific agents handle different inference tasks (e.g., collaborative neurological or forensic reasoning, nursing and pharmacy support, or longitudinal disease management), often via multi-agent decompositions that coordinate guideline-driven diagnosticians, data-fusion specialists, and task-specific tools [18; 363; 287; 243; 117; 349; 7; 106; 326; 33; 40].

**Tool-grounded inference and oversight.** Reasoning in these systems is tightly coupled to external tools and memory: agents actively retrieve relevant guidelines and case histories, integrate structured knowledge graphs, and interact with calculators, prediction models, and RAG-style retrieval modules to ground their inferences and improve medical QA [198; 299; 224; 363; 55; 359; 54; 318; 316; 279; 212; 173; 272; 112]. MA-RAG iteratively evolves both external evidence and internal reasoning history within an agentic refinement loop, transforming semantic conflict among candidate responses into actionable retrieval queries that drive a boosting-like convergence toward stable medical consensus [298]. MedCoG introduces meta-cognitive regulation of reasoning, where self-assessments of task complexity, familiarity, and knowledge density dynamically gate the utilisation of procedural, episodic, and factual knowledge, yielding a 5.5× improvement in inference density on hard medical benchmarks [348]. MedCoRAG constructs patient-specific evidence packages by jointly retrieving and pruning UMLS knowledge graph paths and clinical guidelines, then performs multi-agent collaborative reasoning where a Router Agent dynamically dispatches Specialist Agents based on case complexity for iterative deliberation and targeted re-retrieval [161]. CARE decomposes multimodal medical reasoning into coordinated sub-modules—entity proposal, expert segmentation for pixel-level ROI evidence, and grounded VLM reasoning—with reinforcement learning aligning answers to supporting evidence and an agentic coordinator providing tool-invocation planning and evidence–answer consistency verification [75]. Some systems explicitly aim to prevent hallucination propagation or miscalibration by adding specialist critic, auditor, or oversight agents, or by modeling uncertainty and trust alignment in multi-agent discussions [90; 135; 193; 283; 95; 268; 209; 88; 156; 13; 271; 354].

**Balancing fast and slow modes.** A key open problem is blending fast and slow reasoning. Fast, shallow reasoning is cheap but error-prone; slow, structured reasoning (long chains of thought, extensive tool use, multi-agent debate) is safer but costly. Beyond generic controllers, medical agent work explores reinforcement-learning and active-inference style controllers for deciding when to escalate reasoning depth, ask more questions, or trigger multi-agent consultations, as well as for optimizing downstream policies such as treatment planning, triage, and workflow automation [139; 246; 234; 100; 184; 74; 227; 225; 13; 350]. RL-based meta-controllers

that learn when to switch from fast to slow modes based on estimated difficulty and risk are an active research topic [204; 64]. Medical agents require such controllers to avoid both over-thinking benign questions and under-thinking high-stakes ones, especially in multi-modal and multi-agent settings such as radiology, ophthalmic surgery planning, oncology, and chronic-disease management [92; 61; 188; 88; 270; 71; 361; 152; 7].

#### 2.5.3 Hybrid reasoning

Beyond purely template-based or purely agentic designs, many recent medical agents adopt *hybrid* reasoning that couples guideline workflows, multi-agent collaboration and multimodal perception. In oncology, Cancer-GUIDE uses internal disagreement between multiple guideline-trajectory predictors as a confidence signal, only surfacing treatment plans that are both guideline-concordant and self-consistent [268]. In ophthalmology, CataractSurg-80K benchmarks cataract-surgery planning with explicitly annotated intermediate reasoning steps, encouraging models to expose structured assessments rather than only final recommendations [188].

Similar patterns appear in multi-modal and conversational agents: multi-agent diagnostic frameworks [40], collaborative VLMs for lung X-ray analysis [361], multimodal progression predictors such as OAAgent [7], and argumentation-based systems like ArgMed-Agents [106] all combine heterogeneous reasoning modes and make disagreements observable. MedCollab further strengthens this paradigm by adopting an Issue-Based Information System (IBIS) argumentation protocol that requires agents to provide "Positions" backed by traceable evidence, constructing Hierarchical Disease Causal Chains through explicit logical operators to transform flattened diagnostic predictions into structured models of pathological progression, and iteratively filtering low-quality reasoning through logic auditing and weighted voting [334]. This hybridisation turns inconsistency into a trigger for re-checking or escalation, reducing silent propagation of early hallucinations through the pipeline.

### 2.6 Perception: multimodal sensing and environment interaction

Perception pushes medical agents beyond text-only interfaces into direct contact with imaging, physiologic signals, and other structured data streams. What began as static feature-extraction front-ends feeding a language model is increasingly becoming an interactive sensing layer that can be queried, steered, and audited. We contrast workflow-based perception and fully agentic perception because they determine how well systems can ground their reasoning, avoid perceptual blind spots, and explain downstream decisions.

#### 2.6.1 Workflow-based perception

Multimodal pipelines often pre-process non-text data with specialized models: CNNs for imaging, transformers for ECGs, random forests for lab panels. Outputs (e.g., lesion labels, risk scores) are converted to text or structured fields and passed to the LLM, which remains oblivious to the raw modalities and cannot choose which images to inspect or which analyses to run. Benchmarks and vision–language systems for medical report interpretation and outcome prediction, such as MedRepBench and CT pulmonary angiogram VLMs, typically follow this pattern by evaluating fixed visual encoders whose structured outputs are consumed by downstream language models [235; 353].

#### 2.6.2 Agentic perception

Agentic systems treat perception models as tools. The agent can decide when to view images, which region or modality to analyze, and what quantitative outputs to request. Some systems support active perception: the agent issues bounding box queries, requests zoomed views, or asks for alternative reconstructions. Others use RL-trained multimodal models whose reward reflects task performance (e.g., lesion detection mean average precision).

**Interactive perception as a tool.** In clinical communication and VQA, systems such as AMANDA and ChatMyopia instantiate this pattern by orchestrating LLM agents around image encoders and classifiers, deciding when to trigger visual inspection and knowledge retrieval to answer patient-specific queries [286; 301]. MedScope extends this paradigm to long-form clinical videos, performing coarse-to-fine evidence seeking

by interleaving intermediate reasoning with targeted tool calls and verification on retrieved observations, producing predictions explicitly grounded in temporally localised visual evidence and optimised via Grounding-Aware GRPO [155]. MedXIAOHE incorporates diverse medical reasoning patterns via reinforcement learning and tool-augmented agentic training, enabling multi-step diagnostic reasoning with verifiable decision traces and achieving state-of-the-art performance across medical benchmarks [241]. 3DMedAgent enables 2D MLLMs to perform general 3D CT analysis by coordinating heterogeneous visual and textual tools through a flexible agent, maintaining a long-term structured memory that aggregates intermediate tool outputs and supports query-adaptive, evidence-driven multi-step reasoning across over 40 tasks [290]. In dentistry, OPGAgent coordinates specialised perception modules with a consensus mechanism through hierarchical evidence gathering (global, quadrant, and tooth-level phases), a toolbox encapsulating spatial, detection, utility, and expert models, and a consensus subagent that resolves conflicts through anatomical constraints [328]. Multi-agent clinical and control frameworks likewise tightly couple perception and action: MedCoAct coordinates specialist agents that request and cross-check perceptual evidence for complete clinical decisions, while adaptive multi-agent DRL and Q-learning policies operate over telemetry and endoluminal visual streams to trigger interventions and navigate instruments [351; 234; 184]. Even ostensibly text-only agents such as the term-aware T-Agent expose structured domain knowledge as a tool-like perceptual channel, a design that naturally extends to structured outputs from imaging and reporting systems [110].

**Opportunities and challenges.** Agentic perception closes the loop toward embodied clinical agents that can interrogate imaging archives, telemetry streams, and device outputs as needed to support decisions. At the same time, it raises new challenges in credit assignment (which perception actions were useful), safety (avoiding over-reliance on a single perception model or modality), and explanation (communicating how perception informed reasoning and recommendations).

### 2.6.3 Hybrid perception

Between rigid pipelines and fully agentic perception lies a practically important class of hybrid systems. Here, some perceptual modules remain fixed and always-on (e.g., mandatory safety checks or legacy FDA-cleared CAD), while others are exposed as optional tools that the agent can invoke when uncertainty, novelty, or task demands justify extra scrutiny. A small backbone of standardized summaries is computed for every case, preserving stability, auditability, and compatibility with existing infrastructure, while higher-cost queries (fine-grained segmentations, temporal trend analyses, cross-modality retrieval) are triggered selectively under an attention or compute budget. This design both contains the risk of more flexible tool use and offers a natural locus for governance: institutions can specify which perception tools are mandatory, which are optional, and when disagreements require human escalation, then audit whether deployed agents respect this "perception contract."

## 3 Multi-Agent Medical Systems: Taxonomy, System Architectures, and Collaboration Protocols

The previous section analyzed capabilities at the level of a single medical agent. In this section, we turn to discuss how agents work together in real healthcare settings to cover the entire spectrum of clinical and organizational responsibilities safely and efficiently. The goal of Collective Intelligence is to solve complex problems better as a group than as an individual. This concept, often inspired by nature and human teams, relies on multiple entities working together. This teamwork brings better, stronger, and more flexible results. Modern healthcare is a very complex field. It involves huge amounts of different data, constantly changing patient conditions, and the need for expert decisions from many teams. To handle this scale, we need a computing structure that uses distributed intelligence. Multi-Agent Systems (MAS) are the perfect technology for this. MAS use specialized, autonomous agents to handle specific jobs, like checking diagnoses [40], tracking drug interactions [115], or planning personalized treatments [81]. By distributing the work, MAS reduce the burden on any single system and allow real-time use of expert knowledge. This coordinated group effort improves clinical accuracy, manages hospital resources better, and starts a new era of highly effective patient care.

### 3.1 Categories of Medical Multi-Agent Systems

Modern healthcare demands intelligent systems that can continuously adapt and coordinate effectively to manage its inherent complexity. This subsection classifies medical multi-agent systems (MAS) into three categories, distinguished by their specific combinations of collaboration goals and operational norms. While not exhaustive, these categories encompass a broad range of MAS designs and clearly illustrate how system objectives influence agent interactions and resulting outcomes [216].

**Strategic Learning in Healthcare.** Strategic Learning in Healthcare focuses on embedding agents within a game theoretic context, where agents pursue individual, partially conflicting, or fully cooperative goals. The interactions which can be competitive, cooperative, or mixed are explicitly governed by predefined game rules and interaction norms. This setting often aligns with non cooperative (strategic) and cooperative concepts in traditional game theory, allowing the system to optimize high stakes resource allocation and competitive decision making processes. Agents utilize techniques like Reinforcement Learning (RL) to learn optimal strategies under dynamic constraints [302; 336].

Clinic Resource Management: This strategic environment is crucial for optimizing critical resource allocation [12], such as dynamic management of ICU beds or real time emergency room triage protocols [25], where agent goals (*e.g.,* minimizing patient wait time vs. maximizing unit capacity utilization) may be partially conflicting.

Medical Insurance Negotiations and Drug Pricing: Agents can model the competitive dynamics between pharmaceutical companies, insurance providers, and healthcare systems [209]. These agents learn optimal bidding, coverage, and pricing strategies (e.g., Nash Equilibria) under regulatory constraints and budget limits to predict market outcomes and inform cost effectiveness analysis.

Large Scale Public Health Policy Design: MAS facilitates the simulation of long term impacts of interventions, such as vaccination mandates, resource investment in underserved regions, or quarantine policies [94]. By modeling diverse societal responses and policy maker objectives, agents can iteratively refine public health policy to achieve desired population health outcomes with minimal societal disruption.

**Healthcare System Modeling and Simulation.** MAS excels at creating high fidelity virtual environments, essential for risk free system testing and optimization. Multiple agents collaborate to build a "Virtual Hospital" or a broader "Virtual Social Healthcare System." Li et al. [150] introduced Agent Hospital, where LLM-powered doctors and patients interact frequently , helping doctors continuously improve diagnostic and treatment decisions. This virtual environment supports large-scale evaluation and training of autonomous clinical agents, and can also serve as a controllable testing platform for exploring counterfactual "hypothetical scenarios". Operational Optimization: The virtual environment is used to simulate patient flow dynamics, predict bottlenecks in departments like radiology or labs, and analyze capacity during emergency peaks. Workflow Validation: MAS supports the rigorous evaluation of new clinical processes, such as new outpatient pathways [58], standardized pre operative preparation protocols, and the impact of technology integration. Crisis Preparedness: It facilitates large scale emergency strategy simulation for major public health events, such as epidemic outbreaks. Training and Validation: Systems like the Agent Hospital and virtual medical ecosystem platforms [294; 157] support real world training, doctor patient communication practice, and non disruptive validation of new clinical decision support rules and complex treatment plans.

**Collaborative Clinical Task Solving.** Collaborative task solving brings multiple agents together to pursue a well-defined objective through a structured and tightly coordinated workflow. In contrast to strategic learning, which often involves competition or negotiation, or to modeling and simulation, where agents operate autonomously, collaborative systems organize agents into cohesive teams that progress through sequential or parallel stages of problem-solving. A systematic empirical review by Vatsal et al. across 49 studies confirms that multi-agent design is the dominant architectural pattern ($\sim$82% fully implemented) in current healthcare agents, although orchestration layers remain mostly partial and action-oriented capabilities such as treatment planning still show substantial gaps [273].

The MAS architecture naturally facilitates Collaborative Task Solving by mirroring the stringent, workflow driven division of labor found in real world clinical teams, ensuring both accountability and patient safety. Specialized Role Fulfillment: The system employs a hierarchy of specialized agents to execute complex

clinical workflows, analogous to human expert roles [85]: Diagnostic Agent: Establishes the initial diagnostic hypotheses and determines the necessary investigative framework. Data Analysis Agents (e.g., Imaging [68], Genomics [195]): Perform deep, focused analysis on specific, complex datasets to confirm or refine the diagnosis. Treatment Planning Agent: Synthesizes inputs from all specialists to propose a comprehensive therapeutic plan [300]. Safety Agent (QA): Acts as a critical checkpoint, verifying drug compatibility, dosage against patient biometrics, and correcting potential misdiagnoses before implementation [271]. Documentation Agent: Automatically organizes patient data, generates required reports, and ensures all actions are recorded as structured data. [331; 323] Integrated and Traceable Medicine: This structure enforces a strict workflow driven collaboration model where agents sequentially execute and validate steps to complete treatment plans holistically. This mandated coordination ensures complete traceability of every decision and guarantees high clinical safety by integrating mandatory specialized checks, thus moving beyond fragmented, siloed AI systems.

### 3.2 System Architecture

Section 3.1 focused on what medical MAS are designed to achieve at different scales. To realize these goals in practice, agents must be embedded in concrete architectures that specify who can talk to whom and under what control. In this section, we therefore move from what MAS are trying to achieve to how they are structurally organized, analysing common system architectures used in medical settings.

#### 3.2.1 Static Topologies

Static topologies rely on predetermined patterns of connectivity that remain fixed throughout runtime. In medical multi-agent systems, such as those supporting clinical decision pathways or hospital workflow coordination, the relationships and communication channels among agents or between agents and a central orchestrator are typically defined through explicit clinical guidelines, operational protocols, or rule-based heuristics. This rigidity ensures predictable information flow, which is essential for safety-critical environments and regulated care processes. Common static configurations include centralized, decentralized, and hierarchical structures.

**Hierarchical Structure.** Hierarchical structures arrange agents in layered roles, with higher-level agents directing and integrating the work of lower-level ones. In medical MAS, this mirrors clinical team organization: a coordinating agent assigns specialised diagnostic or analytic tasks to subordinate agents in areas such as cardiology or pathology and then synthesizes their outputs into a coherent clinical conclusion [24]. This structure also supports safety and guideline adherence. Many systems use tiered oversight in which junior agents propose actions, mid-tier agents check for harm or compliance, and a senior agent or human clinician provides final approval [260]. NutriOrion exemplifies a hierarchical multi-agent framework with a parallel-then-sequential reasoning topology: specialised domain agents with isolated contexts first decompose nutrition planning for multimorbid patients, followed by a conditional refinement stage with a multi-objective prioritisation algorithm to resolve conflicting dietary requirements and inject pharmacological contraindications as hard constraints [296]. Vivaldi adopts a role-structured multi-agent system for interpreting multivariate physiological time series in emergency care, revealing that agentic pipelines substantially benefit non-thinking and medically fine-tuned models while explicit tool-based computation is decisive for codifiable clinical metrics [77]. Hierarchical designs enhance traceability and maintain consistent clinical reasoning, though they can create bottlenecks when upper-level agents become overloaded or when rapid adaptation is required in dynamic care settings.

**Centralised Structure.** In centralized structures, a single coordinator agent gathers information and directs the activities of peripheral agents. In medical multi-agent systems, this design is often reflected in architectures where a central clinical decision engine or hospital command system aggregates inputs from diagnostic agents, monitoring agents, and workflow assistants to maintain a unified view of patient status or operational demands. Such centralization supports comprehensive resource management and enables consistent, protocol-aligned decision-making, similar to the global coordination mechanisms described in cultural simulations and collaborative settings [345]. However, centralized approaches frequently face scalability challenges as more clinical or operational agents are added, resulting in heavier communication loads

and an elevated risk of single-point failures. Although centralized structures deliver consistency and strong oversight, their structural rigidity constrains adaptive responsiveness in dynamic healthcare environments.

**Decentralised Structure** Decentralized structures rely on peer-to-peer interactions among agents without a central controller, forming flexible networks that enhance fault tolerance and resilience. In medical MAS, such architectures allow specialist agents to collaborate on complex clinical reasoning, independently reviewing evidence, debating differential diagnoses, or integrating multimodal data such as radiology images and laboratory results [36]. This structure also supports research workflows by enabling continuous knowledge exchange and deliberate dissent to prevent bias. While decentralized systems promote diversity of perspectives and robustness, they require sophisticated consensus, synchronization, and conflict-resolution mechanisms to ensure coherence, safety, and compliance with clinical guidelines.

**Limitations.** Overall, static topologies provide predictability, ease of implementation, and straightforward maintenance due to their clearly defined communication structures and predefined roles. In medical MAS, such designs are well suited for stable workflows with fixed clinical protocols, routine diagnostic pipelines, or structured patient monitoring tasks. Their rigidity ensures reliable information flow and consistent adherence to guidelines, which is critical in safety-sensitive healthcare environments. However, the main limitation of static topologies is their inability to adapt dynamically to unforeseen circumstances, such as sudden agent failures, unexpected patient conditions, evolving clinical requirements, or shifting system objectives. This inflexibility can reduce effectiveness in complex or rapidly changing medical scenarios, highlighting the need for more adaptive and resilient topological solutions.

#### 3.2.2 Dynamic Topologies

Static topologies offer stability and predictability, making them suitable for structured medical workflows such as routine patient monitoring or standardized diagnostic pipelines. However, their rigidity limits adaptability in dynamic clinical scenarios, where patient conditions, resource availability, and treatment goals can change rapidly. Dynamic topologies address these challenges by continuously adjusting inter-agent connections based on real-time feedback, environmental cues, or evolving objectives. Systems like AI Hospital [63] and Agent Hospital [150] demonstrate iterative restructuring of communication networks to support adaptive coordination and maintain reliable, guideline-compliant decision-making. Despite significant advances, dynamic and adaptive multi-agent topologies continue to face critical research challenges.

**Generalizability**: Most systems are optimized for single-task domains, and extending adaptability across multiple clinical scenarios requires lifelong learning capabilities.

**Resource Efficiency**: Dynamic MAS incur high computational and training costs, necessitating more efficient optimization strategies.

**Inference Efficiency**: Complex multi-agent setups can struggle with task adaptability and resource allocation during inference, highlighting the need for scalable adaptive mechanisms.

Addressing these challenges will substantially improve the practical utility, scalability, and robustness of dynamic multi-agent systems in healthcare, enabling more adaptive, efficient, and reliable deployment in real-world clinical environments.

#### 3.2.3 Scalability Exploration

Scalability remains a critical challenge in medical multi-agent systems, particularly as the number of clinical or diagnostic agents grows. In fully connected networks, communication pathways increase rapidly, leading to higher computational costs and potential synchronization bottlenecks. Centralized or hierarchical architectures provide coordination but may become overloaded under high message volumes, while decentralized networks improve fault tolerance yet require complex consensus protocols to maintain coherence.

In healthcare contexts, scalable MAS are essential for tasks such as hospital-wide patient monitoring, distributed diagnostic reasoning, or large-scale clinical simulations. Approaches such as hybrid architectures combining supervisory agents for global coordination with local agents operating semi-independently, can mitigate communication bottlenecks and adapt team sizes to task complexity. Techniques like graph-based

task allocation, adaptive message routing, and asynchronous communication further support scalable, efficient coordination [286; 271]. Achieving scalability enables robust, real-time decision-making across hundreds or thousands of medical agents, while maintaining reliability and adherence to clinical protocols.

#### 3.2.4 Latency Problems

Although MAS have superior reasoning depth, their inherent latency has become a key obstacle in their clinical applications. The latency is not just a matter of computational cost, but also a problem of trust and usability. In fast-paced environments such as emergency triage or monitoring in intensive care units, the lack of response capability can easily disrupt the doctor's workflow. The latency in medical MAS has three main causes: (i) sequential reasoning: multi-round reasoning processes (such as MedAgents [260], MDTeamGPT [34], MedMMV [165]) require repeated calls to large foundational models, and as the debate length increases, the latency accumulates linearly. (ii) communication overhead: scalable architectures (such as Agent Hospital [150]) involve a large amount of message routing and serialization among agents, and as the number of agents increases, the cost becomes very high. (iii) tool integration: external API calls for electronic health record retrieval or image analysis (such as MedOrch [102], AgentMD [123]) introduce unpredictable I/O latency.

To alleviate the problem of latency, at the framework level, it can be classified based on the complexity of the problem, routing urgent or simple issues to fast and streamlined paths (such as MedAgents [132] ), or adaptive activation: using routing mechanisms (such as MoMA [80], MMedAgent-RL [302]) to only activate necessary agents and prune redundant steps. At the technical level, the retrieved context can be reused to avoid redundant "conversations" between agents. Future work must systematically benchmark latency and accuracy to ensure that multi-agent systems are feasible in time-sensitive medical environments.

### 3.3 Communication and Collaboration Mechanisms

The previous subsection examined MAS architectures as connection patterns. Architecture alone, however, does not determine behaviour. It must be combined with concrete communication and collaboration mechanisms that specify what information agents exchange and how they coordinate decisions. In this subsection, we analyze these mechanisms in terms of message types, interaction paradigms, and human–agent collaboration.

#### 3.3.1 Messages Types

The effectiveness of Medical Multi-Agent Systems (Medical MAS) is fundamentally determined by how agents exchange clinical information, coordinate diagnostic reasoning, and execute healthcare workflows. Drawing from advancements in general MAS, communication in medical agents has evolved to balance the rigorous demands of clinical precision with the nuances of patient interaction. We categorize these mechanisms into three domains: Structured Messages, Unstructured Messages, and Standardised Protocols.

**Structured Messages:Precision and Clinical Interoperability.** In healthcare, precision is paramount. Structured communication serves as the backbone for deterministic tasks where ambiguity entails high clinical risk. As observed in general MAS, structured formats such as JSON and XML are widely adopted to ensure high machine readability. In the medical domain, this aligns with established interoperability standards (e.g., HL7 FHIR resources represented in JSON). Agents utilize these formats to encode patient vitals, medication lists, and lab results into unambiguous parameters [152]. Beyond data exchange, structured messages often include executable code (e.g., SQL or Python) allowing agents to interact precisely with external medical databases or calculation tools [81]. The explicit structure ensures verifiability and ease of parsing, which is critical for Clinical Decision Support Systems. It facilitates system-level optimization and persistent memory logging, allowing for retrospective audits of an agent's reasoning path, a crucial requirement for accountability in healthcare.

**Unstructured Messages: Semantic Richness in Clinical Narrative.** While structured data handles metrics, medicine is deeply rooted in narrative. Unstructured communication, primarily through Natural Language (NL), allows agents to handle the "human" aspect of healthcare. Natural language preserves the rich

semantic details found in patient histories, subjective symptom descriptions, and medical literature. As noted in general agent research, this modality is essential for tasks requiring high expressiveness [206]. Unstructured messaging is the primary vehicle for agents performing consultation simulations (e.g., Doctor-Patient agents), drafting discharge summaries, or interpreting unstructured clinical notes (SOAP notes). It captures the tone, urgency, and subtle cues (e.g., a patient's hesitation) that structured formats often strip away. Despite the fluency of modern Large Language Models (LLMs), unstructured communication introduces challenges regarding ambiguity and potential misinterpretation. However, for open-ended tasks like medical education or empathetic patient support, the trade-off favors the expressiveness of natural language.

**Standardised Protocols.** To move beyond isolated interactions, Medical MAS requires standardised protocols to govern how agents collaborate within a complex healthcare ecosystem. These protocols ensure security, privacy, and effective delegation. Horizontal Communication (Peer-to-Peer): Inspired by protocols like A2A (Agent-to-Agent), medical agents employ peer-to-peer delegation models to simulate Multi-Disciplinary Teams (MDTs). For instance, a "General Practitioner Agent" may dynamically negotiate with a "Radiology Agent" and a "Pathology Agent" to solve a complex diagnostic case [41; 40]. Protocols like Agora can function as a meta-layer, allowing these specialist agents to switch between high-bandwidth natural language (for debating a diagnosis) and efficient structured routines (for exchanging patient IDs) [107; 148]. Vertical Communication (Client-Server): Similar to the Model Context Protocol (MCP), vertical protocols standardize the interface between agents and external medical tools (e.g., PubMed search tools, Drug Interaction Checkers [81]). This ensures that agents can reliably access and retrieve external knowledge through a unified client-server interface without hallucinating data sources. Advanced protocols (e.g., ANP) incorporate Decentralised Identity (DID). In medical contexts, this translates to strict access control, ensuring that only authorized agents can access sensitive patient data (PHI), thereby aligning agent communication with privacy regulations (e.g., HIPAA/GDPR).

### 3.3.2 Collaboration Paradigms

Building on the structural foundations of agent topologies and communication mechanisms, we now shift focus to the interaction dynamics that animate Medical Multi-Agent Systems (Medical MAS). While topology defines the architectural skeleton, collaboration paradigms determine how medical agents negotiate diagnostic uncertainty, evolve their clinical knowledge, and execute complex healthcare workflows. Drawing inspiration from sociological theories and clinical practice, we categorize these interactions into three distinct paradigms: Consensus-oriented, Collaborative Learning, and Task-oriented Interaction.

**Consensus-oriented Interaction: The Digital Multi-Disciplinary Team.** In medical contexts, diagnostic accuracy often supersedes speed. Consensus-oriented interaction focuses on aligning diverse specialist perspectives to resolve ambiguity and mitigate clinical risk. This paradigm effectively simulates human Multi-Disciplinary Teams (MDTs) or Tumor Boards. As demonstrated in systems like MedAgents [260] and AI Hospital [63], agents representing different specialties (e.g., a Radiologist Agent, a Pathologist Agent, and an Oncologist Agent) engage in multi-round structured dialogues. Unlike simple voting, these agents utilize debate and mutual critique to challenge hallucinations and refine reasoning steps. This interaction relies on mechanisms such as iterative refinement and confidence-weighted voting. For instance, if a General Practitioner agent proposes a rare diagnosis, a Specialist agent may challenge it based on contradictory lab evidence, forcing the system to "re-think" and converge on a high-fidelity conclusion grounded in collective deliberation. This paradigm is critical for complex problem-solving where no single agent possesses complete knowledge. It transforms the "Black Box" of individual LLMs into a transparent, self-correcting reasoning process, significantly enhancing reliability in high-stakes medical decisions [11].

**Collaborative Learning Interaction: Evolving Medical Intelligence.** While consensus aims for a single decision, collaborative learning focuses on the mutual enrichment of the agents themselves. This paradigm mirrors Continuing Medical Education (CME) and residency training, where agents share experiences to accelerate individual skill acquisition without necessarily seeking immediate agreement. Agents engage in post-hoc analysis of clinical cases. For example, in frameworks like MEDCO [291], student agents discuss diagnostic paths with mentor agents to refine their internal reasoning models. By explaining, critiquing, and revising their logic in a peer-to-peer setting, agents uncover diverse reasoning paths that a solitary model might miss. Agents operating in different "virtual hospitals" or processing different datasets can pool their

insights. Through observational learning, an agent that successfully identifies a rare pathology can share this "experience" (encoded as high-level rationales rather than raw data) with peers, thereby updating the collective knowledge base of the system. A key challenge in this paradigm is preventing the propagation of errors (hallucination amplification). Effective medical collaborative learning requires strict filtering mechanisms to ensure that only valid, high-quality clinical insights are integrated into the agent collective.

**Task-oriented Interaction: Clinical Workflow Automation.** In contrast to the deliberation of consensus or the exploration of learning, task-oriented interaction is pragmatic and execution-driven. It aligns agents along structured Clinical Pathways or administrative pipelines to maximize operational efficiency. This paradigm adopts a "production line" approach, where interactions are governed by strict modularity and temporal ordering. For instance, a Patient Intake Agent collects symptom data and passes it to a Triage Agent, which then triggers a Lab Ordering Agent [217]. Unlike the conversational depth of consensus interactions, task-oriented agents rely on clear hand-offs and predefined deliverables (e.g., a standard FHIR object). Collaboration is achieved through the successful transfer of intermediate results (e.g., a lab report) rather than mutual persuasion. This paradigm is ideal for deterministic, high-volume tasks such as hospital scheduling, automated discharge planning, or prior authorization processing. It minimizes the cognitive overhead of dialogue, ensuring that healthcare workflows proceed with logical consistency and speed.

#### 3.3.3 Human-Agent Collaboration

While previous content details the internal topology and collaboration among agents, a complete Medical MAS ecosystem is defined by its interface with human clinicians. Human-agent collaboration serves as the critical bridge grounded in clinical reality, enabling healthcare professionals to harness the computational power of MAS while maintaining the "Human-in-the-Loop" for safety and accountability. We categorize these interactions into three evolutionary paradigms: Single-turn Task Delegation, Multi-turn Interactive Instruction, and Immersive Human-Agent Collaboration.

**Single-turn Task Delegation.** In its simplest form, human clinicians delegate discrete, well-defined tasks to agents via one-off commands. This paradigm focuses on efficiency and administrative offloading. Examples include a physician asking a retrieval agent to "List the contraindications for Metformin" or a scribe agent being tasked to "Generate a discharge summary based on the provided lab results." The MAS processes the request and returns a complete response without requiring further dialogue. As noted in general agent literature [215; 260], this mode currently dominates the medical landscape. It is highly effective for low-risk, deterministic tasks such as medical coding, appointment scheduling, and information retrieval from Electronic Health Records (EHR), acting essentially as an intelligent command-line interface for healthcare.

**Multi-turn Interactive Instruction.** Moving beyond static delegation, this paradigm introduces iterative feedback loops, positioning the human clinician as a supervisor and the agent as a reasoning assistant [138]. This is crucial for complex clinical decision-making where ambiguity exists. In diagnostic scenarios, a doctor does not simply accept an agent's initial differential diagnosis. Instead, they engage in a multi-round dialogue: correcting the agent's misinterpretations of symptoms, adding new context ("Patient traveled to the tropics recently"), or steering the agent to focus on specific organ systems. This interaction style mimics the attending-resident relationship [265]. The human expert guides the agent toward a satisfactory outcome, ensuring that the agent's reasoning aligns with clinical guidelines and patient values. This is prevalent in Clinical Decision Support Systems (CDSS), where agents help draft complex treatment plans or write research papers. The human provides "guardrails" to prevent hallucinations, making this the primary mode for ensuring safety in generative medical AI.

**Immersive Human-Agent Collaboration.** In the most integrated paradigm, agents evolve from subordinates to peer collaborators. Immersive collaboration is characterized by symmetry, continuous interaction, and shared situational awareness. Unlike interactive instruction where the human initiates, immersive agents can proactively propose actions or flag risks in real-time. For instance, during a surgery, a robotic agent might not just follow commands but actively anticipate the surgeon's next move, adjusting the camera angle or holding tissue automatically [315]. In a "Smart ICU" setting, agents and humans work as a cohesive team. The agent monitors continuous vital signs and alerts the team to subtle trends (e.g., early sepsis signs) that humans might miss due to cognitive overload [136]. The interaction is fluid and bidirectional—akin to a

seasoned nurse whispering a suggestion to a doctor. This paradigm represents the frontier of Medical MAS. It requires agents to possess high-level "Theory of Mind" to understand the clinician's intent and the patient's emotional state, fostering a seamless partnership that transcends sequential command chains.

| **Clinical Function** | **Task** | **Sub-Types** | **Paper** |
|---|---|---|---|
| Basic Technology Empowerment Function | Interactive Medical Image Segmentation | Click-Driven | [239; 176; 114] |
| | | Prompt-Driven | [240] |
| | | Audio-interaction | [181] |
| | Medical Image Classification | Xray-based | [18] |
| | | WSI-based | [85; 221] |
| | Medical Structural Information Processing | Clinical notes-based | [142] |
| | | EHR based | [243; 143] |
| | Medical Knowledge Graph Construction | - | [335; 59] |
| Core Diagnostic and Therapeutic Assistance Function | Medical Question-Answer | QA | [198; 224; 52] |
| | | VQA | [262; 333] |
| | Multi-turn Doctor-Patient Dialogue | Diagnostic Simulation and Training | [69; 11; 56] |
| | | Doctor Communication Assistance | [197; 158] |
| | | Quality Evaluation and Optimization | [327] |
| | Clinical Surgery | Surgical Navigation | [225; 184; 297] |
| | | Surgical Instrument Operation | [51] |
| | | Surgical Planning and Scene Understanding | [174; 344] |
| Workflow and Documentation Optimization Function | Clinical Service Optimization | Emergence Department Workflow Optimization | [93] |
| | | Reception Workflow Optimization | [19] |
| | | Examination Pathway Optimization | [318] |
| | Report Generation | Report Precision | [323; 278; 58] |
| | | Report Standardization | [270] |
| | | Patient-friendly Report | [10; 254] |
| | Medical Text Optimization | Medical Error Correction | [49] |
| | | Medical Concept Standardization | [8] |
| | | Medical Text Simplification | [175] |
| | | Prescription Verification | [271] |

Table 2: Classification of Medical Agent Tasks.

# 4 Atomic Capabilities of Medical Agents: A Functional Task-Level Perspective

Having outlined how capabilities are assembled into multi-agent systems, we next zoom in on the tasks that medical agents are actually designed to solve. This section analyzes the technical details of agents in different medical tasks from a task-level problem definition perspective. It focuses on the core chain of the clinics and classifies tasks into three clear categories based on clinical value: Basic Technology Empowerment Function, Core Diagnostic and Therapeutic Assistance Function, and Workflow and Documentation Optimization Function. Each category addresses distinct clinical needs while contributing to the overall conversion of AI capabilities into practical medical value. The overall structure is displayed in Table 2.

## 4.1 Basic Technology Empowerment Function

As the underlying technical foundation of clinical AI applications, basic technology empowerment provides core support for upper-level diagnosis and treatment processes including image analysis, information extraction and knowledge structuring. It serves as an essential prerequisite for turning AI potential into practical clinical results.

### 4.1.1 Interactive Medical Image Segmentation

Studies in this paragraph focus on agent-driven medical image segmentation (MIS), with each agent's key design and core interaction tailored for human-AI collaboration. In [239], each voxel is an independent agent that adjusts segmentation probabilities via discrete actions from a shared policy, paired with an action-based confidence network to assess action-ground truth alignment, interacting via user clicks. In [176], the proposed method extends this multi agent reinforcement learning (MARL) framework, where voxel agents use a boundary-aware reward, i.e., global cross-entropy + edge weighting, and interact via supervoxel-level clicks for noise robustness. In [114], Zhang's agent is a RL-trained "clicking agent" that samples clicks from a probability map, receives binary "better/worse" rewards, and interacts via clicks + preference feedback. In [240], the authors design the agent as a temporal prompt optimizer based on markov decision process and deep Q-network, which selects optimal prompt types, i.e., points and bounding boxes, interacting through prompt selection to reduce SAM's medical prompt sensitivity. In [181], the propsoed SCOPE is a speech-guided agent empowered by LLM and vision foundation models, which processes verbal commands, tracks instrument tips and interacts through speech and a depth-aware virtual cursor. Across works, agents focus on action/prompt decision-making and feedback integration, with intuitive interactions, i.e., clicks, prompts and speech, to cut human effort.

### 4.1.2 Medical Image Classification

Researches in this part focus on agent-driven medical image classification, categorized by data type, i.e., 2D radiographs vs. 3D whole slide images, (WSIs). For 2D chest X-rays diagnosis, [18] deploys three agents—Screening Agent (contrastive VLM for guideline-aligned findings), Diagnosis Agent (chain-of-thought reasoning for prediction), Refinement Agent (adjusts for disease interdependencies), which all operate with only guidelines and images. For 3D WSIs diagnosis, Ghezloo et al.'s PathFinder [85] integrates four agents—Triage (benign/risky classification), Navigation (text-conditioned patch sampling), Description (histopathological patch texts), Diagnosis (synthesizes descriptions); Quang et al.'s GMAT [221] deploys four domain agents—Planning (description structure), Generate (class texts), Verify (medical accuracy), Finalize (output structuring), to extract textbook knowledge for WSI subtype classification.

### 4.1.3 Medical Structural Information Processing

Beyond image-centric technologies, structured processing of clinical text and EHR data is another key basic technology empowerment component. For clinical text structured extraction, Lee et al.'s Sporo AI Scribe [142] uses a multi-agent system with fine-tuned medical LLMs to extract key info from doctor-patient dialogues, generates hallucination-suppressed structured SOAP notes, and converts unstructured text into EHR-compatible docs aligned with PDQI-9 [252]. For EHR structured preprocessing, Lee et al.'s EMR-AGENT [143] uses two agents: Cohort and Feature Selection Agent (which enables iterative SQL interaction for

cohort/feature extraction with error feedback) and Code Mapping Agent (which handles feature localization and candidate matching for code standardization), and it adapts to heterogeneous schemas (MIMIC-III, eICU) without manual rules; Shi et al.'s EHRAgent [243] employs an agent with Python code generation and execution capabilities, integrates medical knowledge and to handle multi-table EHR queries via natural language-to-code, and interacts through iterative debugging.

#### 4.1.4 Medical Knowledge Graph Construction

High-quality medical knowledge graphs (KGs) organize scattered medical information to support knowledge-intensive clinical tasks such as multi-turn doctor-patient dialogue [327], interactive chatbots [347] and medical question answering (MQA) [224]. They are a key part of basic technology empowerment that enhances the knowledge capabilities of clinical AI. Sone works also introduce the Agent-based frameworks for constructing more effective medical knowledge graphs [335; 59]. For examples, MedKGent [335] mainly contains two agents, where the Extractor agent identifies relational triples with sampling-based confidence scores from decades of PubMed abstracts, while the Constructor agent incrementally integrates these triples into a time-evolving KG, resolving conflicts and reinforcing recurring knowledge.

### 4.2 Core Diagnostic and Therapeutic Assistance Function

This function integrates directly into the core process of clinical diagnosis and treatment, addressing key issues such as diagnostic decision-making, treatment implementation and doctor-patient communication. It serves as the primary link for delivering clinical value.

#### 4.2.1 Medical Question Answering (MQA)

These studies focus on agent-driven Medical Question Answering (MQA) and are categorized into two non-overlapping types: pure QA and Visual Question Answering (VQA). For text-only QA, MedTrust-RAG [198] uses an agent to conduct iterative retrieval-verification (assessing evidence adequacy via Medical Gap Analysis and refining queries) and reduce hallucinations with Negative Knowledge Assertions. AMG-RAG [224] employs an agent to dynamically build confidence-scored Medical Knowledge Graphs (MKGs) and enable multi-hop reasoning for accurate answer grounding. SOLVE-Med [52] integrates a Router Agent to classify queries to select relevant specialty agents, and an Orchestrator Agent to synthesize outputs from 10 domain-specific small language models, for cross-specialty text-based consultations. For multi-modal VQA, Thakrar et al.'s system [262] uses agents to mimic clinical collaboration (peer consultation) and reference-checking, processing dermatological images to align visual features with text questions. Zhan et al.'s conditional reasoning framework [333] uses an agent that adjusts reasoning strategies by question type to enhance cross-modal alignment between medical images and answers.

#### 4.2.2 Multi-turn Doctor-Patient (DP) Dialogue

Multi-turn doctor-patient (DP) dialogue goes beyond static MQA to simulate real clinical interactions. It emphasizes dynamic two-way communication that aligns with core diagnostic workflows such as inquiry, information accumulation and diagnosis, and prioritizes adaptability to clinical processes. Existing agent-based works could be categorized into three parts: diagnostic simulation and training, doctor communication assistance, and dialogue quality evaluation and optimization. For diagnostic simulation and training, DoctorAgent-RL [69] employs a MARL framework where a doctor agent optimizes questioning strategies via interactions with a patient agent, guided by a multidimensional evaluator. MedAgentSim [11] integrates doctor, patient, and measurement agents with self-improvement mechanisms, i.e., memory replay and chain-of-thought reasoning, for realistic clinical interactions. EvoPatient [56] proposes a coevolution framework where patient and doctor agents refine responses and questions through simulated dialogues, supported by attention and trajectories libraries. For doctor communication assistance, Li et al. [158] propose a proactive dialogue generator, which uses a two-stage recommendation structure, i.e., query generation and candidate ranking, to enable doctor agents to proactively collect diagnostic information. Dr.Copilot [197] consists of three DSPy [130]-optimized agents, including Scorer, Recommender and Reconciliation that provide targeted improvement suggestions for Romanian telemedicine. For dialogue quality evaluation and optimization,

MedKGEval [327] leverages a knowledge graph-driven multi-agent system, containing Doctor, Patient, Judge and Director agents, to simulate dynamic dialogues and conduct real-time turn-level evaluation of clinical appropriateness and factual accuracy.

#### 4.2.3 Clinical Surgery Task

Clinical surgery demands high precision and adaptability, and agent-based technologies are increasingly supporting key surgical processes as part of core therapeutic assistance. Studies in this paragraph focus on agent-driven surgical AI applications, categorized into surgical navigation, surgical instrument operation, and surgical planning and scene understanding. For surgical navigation, Robertshaw et al. [225] propose a agent based on TD-MPC2 [96] for autonomous endovascular navigation in mechanical thrombectomy, outperforming previous SOTAs across ten patient vasculatures. Medina et al. [184] develop a Q-learning agent to navigate bronchial endoluminal channels, leveraging state-action Q-values and energy gradient adjustment for trajectory optimization. Wu et al. [297] design a magnetically actuated milli-spinner agent with helical fins and slits, achieving rapid vascular navigation and overcoming pulsatile flows. For surgical instrument operation, Deng et al. [51] present a deep imitation learning agent trained on simulated data to automate drop-in gamma probe manipulation, using visual input and robot state for end-to-end control. For surgical planning and scene understanding, SurgRAW [174] employs a CoT-driven multi-agent framework with RAG and panel discussions to enhance surgical scene interpretability. FUAS-Agents [344] integrates segmentation, dose prediction and strategy-generation agents to produce personalized focused ultrasound ablation treatment plans.

### 4.3 Workflow and Documentation Optimization Function

This function focuses on the standardization and automation of clinical workflows and medical documentation. It reduces redundant operations and human error, lowers the consumption of medical resources, improves efficiency in core diagnostic and treatment processes and indirectly ensures the effective implementation of clinical value.

#### 4.3.1 Healthcare Service Optimization

Optimizing healthcare services is key to reducing inefficiencies and improving patient care, and agent-based systems are helping streamline hospital workflows. The studies of healthcare service optimization can be categorized into reception workflow optimization, emergency department (ED) workflow optimization, and medical examination pathway optimization. For reception services optimization, PIORS [19] uses two core agents of PIORS-Nurse and HospInfo-Assistant to boost outpatient reception efficiency and personalization. PIORS-Nurse agent is builted by fine-tuing VLM for department guiding and pre-diagnosis info gathering, and HospInfo-Assistant is constructed by a LLM agent interacting with hospital information system for patient archives and admin info retrieval. For ED workflow optimization, Han et al propose a multi-agent [93] system integrates four specialized agents: Triage Nurse, Pharmacist, plus Emergency Physician and ED Coordinator, all collaborating via CrewAI for diagnosis, treatment planning, and resource allocation. For medical examination pathway optimization, HiRMed [318] adopts a tree-structured RAG-agent framework, i.e., Root/Department/Item layers. Each agent implements RAG-enhanced reasoning with a dual-layer knowledge base, uses memory-augmented logic, and leverages a fine-tuned LLM agent to weight test recommendations, ensuring high coverage and low miss rates.

#### 4.3.2 Report Generation

Medical report generation is a critical documentation task, and multi-agent architectures centered on LLMs or VLMs are transforming how these reports are created to meet clinical needs [278; 58; 270; 254; 10; 323]. These studies focus on agent-driven radiology report optimization, categorized into three goals: improving report accuracy, standardizing reports, and enhancing readability/patient-friendliness: For improving report accuracy. Yi et al. [323] utilize multimodal multi-agent system, which employs Retrieval, Vision, and Synthesis agents to align visual evidence with textual context, reducing hallucinations; MRGAgents [278] fine-tunes 13 disease-specific agents to balance normal/abnormal findings, enhancing diagnostic coverage; Elboardy et al.

[58] present a ten-agent framework includes Quantitative Segmentation and Diagnostic Classifier agents to supplement precise measurements and validate findings. For standardizing reports, MedPAO [270] uses a Plan-Act-Observe (PAO) loop agent with protocol-driven (ABCDEF) concept categorization and ontology mapping to structure unstructured text; For enhancing readability and patient-friendliness, Alam et al. propsoe a [10] multi-agent RAG framework integrates Concept Bottleneck Model (CBM)-based interpretable concept vectors and disease-specific ReAct agents to retrieve clinical documents and generate clinically consistent reports. Reflexion agent [254] iteratively optimizes radiology reports by verifying International Classification of Diseases (ICD)-10 codes and adjusting Flesch-Kincaid grade levels; their agent balances medical accuracy and readability to meet AMA's 6th-grade level recommendation.

#### 4.3.3 Medical Text Correction and Simplification

Beyond report generation, agent-based frameworks are also improving other medical text processing tasks that support workflow standardization and accessibility. One study [49] proposes a system employing four specialized agents, i.e., MedReAct, MedEval, MedReFlex and MedFinalParser, which collaborate in a fixed structured workflow. This workflow includes observation-action cycles, multi-criterion evaluation, reflective optimization and result formatting for medical text correction. Another study [8] introduces the Agentic Model Context Protocol framework, which equips LLM agents with access to external vocabulary tools and two-step reasoning such as keyword inference and concept selection for medical concept standardization. Medical Simplifiers [175] contains five role-specific agents, i.e., Layperson, Simplifier, Medical Expert, Language Clarifier, Redundancy Checker, organized into iterative interaction loops, enabling collaborative refinement for accurate medical text simplification. Rx-Strategist [271] introduces a multi-agent system that combines ICD Finder, ICD Matcher, Dosage Retriever and Checker, to enable precise prescription verification.

### 4.4 Future Research Direction

By reviewing Table 2, we observe that agents have relatively limited applications in basic technology empowerment, primarily as basic tasks like medical image segmentation and classification boast mature end-to-end solutions via deep learning models such as UNET. These tasks require no intermediate process decomposition, making it hard to leverage agents' core strength in complex task breakdown. Most existing agent studies focus on tasks that alleviate direct medical burdens by enhancing individual doctors' efficiency, such as MQA, report generation and clinical surgery. However, workflow optimization has been significantly overlooked in current agent research. In fact, such tasks not only reduce hospitals' operational burden globally but also align closely with agents' capabilities in task decomposition and cross-subject collaboration. Thus, future research should focus on unlocking the potential of agent frameworks for overall clinical workflow optimization, yet two key challenges remain. First is benchmark construction: cross-departmental processes involve multiple participants and scattered data across systems, posing significant challenges for data collection and standardization. Second is weak robustness. From my point of view, existing static pipeline agents [19; 93; 318], even with task decomposition based on prior knowledge, would struggle to dynamic clinical workflows, particularly across different countries. Methodological innovations are therefore needed, such as the proactive, self-optimizing agents that sense changes, adjust strategies on the fly, and refine collaboration via real-time clinical feedback.

## 5 Medical Agents: The Hospital Department-level Practice Perspective

Sections 2- 4 took a technological view of medical agents: we described their capabilities, how they form multi-agent systems, and the task formulations they address. In this section, we will explore how these components are reflected in specific specialty areas from the perspective of clinical departments. The rapid development of artificial intelligence in healthcare has created unprecedented opportunities for the deployment of medical agents across hospital departments. Beyond improving the efficiency of routine workflows, these systems are progressively reshaping how clinical resources are organized and coordinated. While department-specific medical AI systems leverage specialized domain knowledge and established clinical norms to optimize local decision-making, the growing complexity and continuity of patient care have motivated a transition toward cross-departmental collaborative systems, which can provide more integrated and holistic support

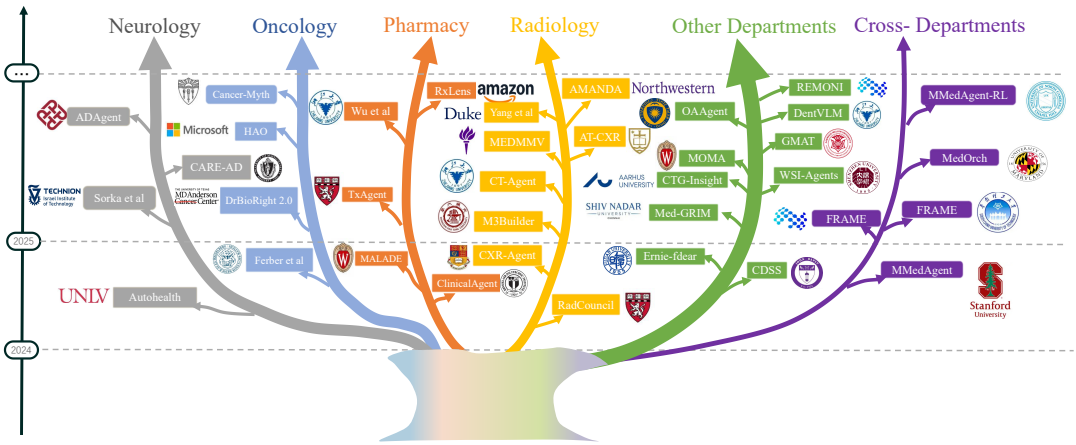


Figure 4: The Evolution Landscape of Medical Agents in Hospital Department.

than any single department in isolation. Focusing on high-volume specialties with rich digitised data and well-defined clinical workflows, in this section, we review the departmental-level practices of medical agents in six key areas: **Neurology Department** (Section 5.1),**Oncology Department** (Section 5.2),**Pharmacy Department** (Section 5.3),**Radiology Department** (Section 5.4) and **other departments of the hospital** (Section 5.5).These department-oriented systems together form a continuously evolving ecosystem of medical agents within hospitals, as illustrated in Figure 4.

### 5.1 Neurology Department

The neurology department is fundamentally oriented toward the non-surgical management of complex disorders of the nervous system and therefore represents a key proving ground for medical agents in routine clinical practice.

Existing work on medical agents mainly focuses on disease-specific decision support and complex clinical reasoning. ADAgent [109] is the first single LLM-driven agent specifically developed for analyzing Alzheimer's disease to address user queries and support decision-making. To better predict the onset of Alzheimer's disease, CARE-AD [152] employs a multi-agent LLM framework to simulate the collaborative diagnostic process of clinicians. AutoHealth [29] interacts with Parkinson's disease patients through an Internet of Medical Things (IoMT) system for early detection and health management.Beyond disease-specific applications, Sorka et al. [250] demonstrated that a structured multi-agent paradigm, which emulates specialized cognitive processes, can substantially enhance complex clinical reasoning in neurology.

**Challenges.** A core challenge is integrating multimodal or longitudinal information into transparent diagnostic reasoning. Frameworks such as CARE-AD [152] alleviate this problem by assigning dedicated LLM agents to extract key information and conduct domain-specific evaluations to improve early risk prediction and interpretability. Current medical agents in neurology department still lack clinically grounded, workflow-aware evaluation. Recent works begin to alleviate this by validating agents on board-style exams, real patient cohorts, and comparative case studies against human experts or incumbent tools, but large-scale, prospective and multi-site assessments remain scarce.

### 5.2 Oncology Department

The oncology department is responsible for the diagnosis, treatment, and comprehensive management of benign and malignant tumors, with a particular emphasis on malignant neoplasms. As a data and guideline-intensive professional field, building medical agents based on large language models (LLMs) holds great promise in oncology.

In terms of patient information summary generation, Blondeel et al. proposed the multi-agent framework HAO [24] to address the problems of time-consuming and inconsistent information quality that arise when summarizing patient information using traditional manual methods. To address the challenges of integration of multimodal data and multidomain expertise in oncology clinical decision-making, Ferber et al. proposed the RAG-Agents system in 2024 [71]. Subsequently, they established a two-stage system [72] to support personalized oncology clinical decision-making.They found that combining language models with precise oncology and search tools could significantly enhance clinical accuracy.Manually developing treatment plans that comply with the guidelines of the National Comprehensive Cancer Network (NCCN) for breast cancer patients is both time-consuming and labor-intensive.To enhance efficiency and accuracy, Mohammed et al. proposed two AI-driven methods: Agentic-RAG and Graph-RAG [190]. Converting the information of non-small cell lung cancer patients into treatment recommendations that comply with the NCCN guidelines requires a significant amount of time and expertise. Therefore, Unell et al. use a hybrid Agent and meta-classifier framework [268], which can strike a balance between accuracy, interpretability, and regulatory requirements, while reducing annotation costs. Zhu et al. [358] found that LLMs often fail to identify the common false assumptions (patients' incorrect hypotheses and beliefs) in real-world patient situations, posing potential medical safety risks. Therefore, they introduced "Cancer-Myth" (a adversarial dataset) to assess this risk.

**Challenges.** In oncology department, medical agents face intertwined challenges in guideline adherence, patient communication, multimodal reasoning, and process interpretability. Generating personalized yet strictly guideline-compliant treatment plans remains difficult, motivating enhanced retrieval pipelines, guideline-aware reasoning, and disagreement detection mechanisms inspired by systems such as Agentic-RAG, Graph-RAG [190], and CancerGUIDE [268]. LLMs often fail to recognize and correct patients' false presuppositions, as shown in Cancer-Myth [358], underscoring the need for oncology-specific safety layers and benchmarks that explicitly target misconception detection rather than generic factual QA. Oncology decisions rely on integrating heterogeneous data such as imaging, pathology, genomics, and clinical records and high demands for traceability. While tool-augmented agents like HAO [24] can orchestrate specialized models and databases, they also increase system complexity and error propagation risks.

### 5.3 Pharmacy Department

Pharmacy department is central to ensuring the safe and rational use of medications, covering prescription review, dispensing, and therapeutic monitoring. In this subsection, we review how such agents are being used to support prescription validation, medication management, and medication safety.

In response to the 40% prescription error rate in hospitals and the shortage of pharmacists, Van et al. proposed a multi-stage LLM agent system [271], which utilizes knowledge graphs and various search strategies to significantly reduce prescription errors and enhance patient outcomes. In the field of medication management,due to the difficulty in interpreting handwritten prescriptions in emerging markets, customers urgently need an automatic, fast, accurate and scalable prescription digitization system to achieve seamless online pharmacy ordering. Jagatap et al. designed the LLM-driven multi-agent framework RxLens [117] used for automatically building the pharmacy shopping cart. In terms of pharmacovigilance, Choi et al. proposed the LLM-driven agent MALADE [47] for extracting Adverse Drug Events(ADEs) from drug label data. There are issues such as drug interactions, redundant prescriptions, and conflicting clinical goals in the treatment recommendations for patients with multiple diseases. In terms of medication safety,Wu et al. [300] proposed a dynamic multidisciplinary team (MDT) framework led by general practitioner (GP) to generate a treatment plan for patients with multiple diseases. LLM has limited capabilities in obtaining and integrating external knowledge resources (such as DrugBank [284]) in clinical pharmacy trials. Yue et al. proposed ClinicalAgent [330], which is a well-structured multi-agent system used for predicting results, explaining the causes of failures, and estimating the duration of the trial. In order to generate personalized treatment suggestions for precise treatment, Gao et al. proposed TxAgent [81] to analyze drug interactions, contraindications and treatment strategies for specific patients. In drug combination extraction, RexDrug adopts a two-stage training strategy that first uses a multi-agent collaborative mechanism to automatically generate high-quality reasoning traces for supervised fine-tuning, then applies reinforcement learning with a multi-dimensional reward function specifically tailored for n-ary drug combination extraction, consistently

outperforming baselines on both n-ary and binary drug interaction tasks [285]. NutriOrion addresses personalised nutrition intervention for multimorbid patients through a hierarchical multi-agent framework with a parallel-then-sequential reasoning topology, decomposing planning into specialised domain agents with isolated contexts and injecting pharmacological contraindications as hard negative constraints during synthesis [296].

**Challenges.** Current systems still struggle with incomplete or shallow recommendation generation for multimorbidity patients, often missing valid therapeutic options, under-detecting drug conflicts, and providing limited substitution strategies, as highlighted by Wu et al [300]. Agents remain highly dependent on static or manually curated knowledge and lack robust, autonomous knowledge retrieval and updating mechanisms. It motivates tighter integration of structured knowledge graphs and region-specific retrieval pipelines, as in Rx Strategist [271]. Medical agents are still evaluated on narrow technical metrics and small-scale datasets. There is a need for standardized, clinically grounded benchmarks to assess real-world safety and effectiveness in diverse healthcare settings.

### 5.4 Radiology Department

Radiology Department is a department that uses various imaging techniques to conduct examinations, diagnoses and some interventional treatments on the human body.With the increasing volume and complexity of imaging studies, radiology department has become a key application area for medical agents that integrate visual perception with clinical reasoning. In this subsection, we review the applications of medical agents in radiology department , focusing on their roles in assisting image analysis and reasoning,radiotherapy plan generation, and processing radiology reports.

- **Image analysis and reasoning** In response to the shortcomings of medical multimodal large language models in integrating image details with professional knowledge, a series of LLM-driven agent systems have emerged recently. In response to the problems that existing multimodal medical models neglect image details and lack professional knowledge, Wang et al. proposed AMANDA [286]. To enhance the reliability of image analysis, AT-CXR [159] and MedMMV [165] respectively reduce hallucinations through uncertainty awareness and reasoning path sampling. MedRAX [61] and AURA [65] focus on improving the interpretability and adaptability of chest X-ray interpretation. Additionally, to optimize the diagnostic process, MedAgent-Pro [288] introduces step-by-step evidence-based reasoning, while MAM [355] and PASS [70] achieve more flexible, efficient, and adaptable medical image diagnosis workflows through modular multi-agent collaboration and probability-driven sampling, respectively. TEA-CXA addresses the critical challenge of conflicting tool outputs by enabling an agent to empirically learn the practical trustworthiness of each tool across different query types via multimodal agentic reinforcement learning [111]. CARE decomposes multimodal medical reasoning into coordinated sub-modules with pixel-level ROI evidence grounding and reinforcement-learning-based answer–evidence alignment, advancing clinical accountability beyond end-to-end black-box systems [75]. For volumetric imaging, 3DMedAgent enables 2D MLLMs to perform general 3D CT analysis by progressively decomposing complex analyses into tractable subtasks and maintaining a long-term structured memory for evidence-driven multi-step reasoning [290]. MedSAM-Agent reformulates interactive medical image segmentation as a multi-step autonomous decision-making process with a two-stage training pipeline integrating multi-turn outcome verification and clinical-fidelity process rewards [170], while MemSeg-Agent shifts adaptation from weight space to memory space, conditioning a fixed backbone with dynamically composed lightweight memories for few-shot, federated, and test-time adaptation [32].

- **Radiotherapy plan generation** To address the complexity and time-consuming issues in radiotherapy planning, Nusrat et al. proposed DOLA [201], which is fully deployed locally to strictly protect patient privacy. Yang et al. [314] adopted a zero-sample LLM agent and achieved fully automatic IMRT planning in the commercial Eclipse system without any training data or professional fine-tuning. Both of the two works adopted the approach of decomposing the planning optimization process into multiple steps, and utilized explicit chain reasoning and natural language explanations to enhance the comprehensibility.

- **Processing radiology reports** Automated medical report generation aims to alleviate the burden on radiologists, but it faces challenges such as hallucinations, unstructured data, and lack of clinical details. To improve the accuracy and consistency of report generation, RadCouncil [331] and Yi et al. [323] introduced multi-agent collaboration and feedback mechanisms. CXR-Agent [236] and MRGAgents [278] respectively solved the problems of uncertainty quantification and incomplete clinical information coverage through modular design. In specific application scenarios, Sudarshan et al. [254] transformed the reports into language understandable to patients using a self-reflection framework, while Ben-Atya et al. [20] solved the problem of structured extraction of non-English reports. Additionally, in response to the limitations of existing evaluation metrics, GEMA-Score [342] proposed a fine-grained interpretable multi-agent scoring system to more comprehensively evaluate clinical value.

**Challenges.** In radiology department, medical agents still face substantial challenges in delivering reliable, end-to-end support across heterogeneous imaging tasks and modalities. In thoracic radiography triage, systems such as AT-CXR [159] must make upgrade or deferral decisions under real-world resource constraints and uncertain findings, yet principled, uncertainty-aware governance for when to stop, escalate, or postpone remains underdeveloped. In medical visual question answering and image-centric reasoning, current multimodal agents often lack sufficiently fine-grained visual understanding and dynamic knowledge integration, leading to hallucinations. This motivates multi-step problem decomposition and tightly coupled, continuously updated medical knowledge retrieval, such as AMANDA [286]. In medical report generation and structuring, single-step LLM approaches remain prone to factual inconsistency, hallucinations, and weak structuralization, motivating multi-step, retrieval-enhanced and protocol-driven agents like MedPAO [270] that can support verifiable, traceable and clinically actionable imaging reports.

### 5.5 Other Departments

The applications of medical agents are not limited to the departments discussed above, but are also emerging in a variety of other hospital units with distinct workflows and information needs. In this subsection, we briefly review department-level practices in additional settings including emergency department, dermatology department, ophthalmology department and the rare disease diagnosis and treatment center.

**Emergency Department.** The emergency department is a specialized department in the hospital that is responsible for treating critically ill patients. To effectively utilize the rich and complementary patient information from multimodal electronic health record (EHR) data, Gao et al. developed the LLM-driven multimodal agent mixture (MoMA) architecture [80], which employs specialized LLM agents to convert non-textual data into multimodal summaries and generate clinical predictions. Emergency medical dispatch often faces challenges such as unclear information, caller distress, and high cognitive load for dispatchers, which undermine the accuracy of decision-making and operational efficiency. Li et al. [157] developed a multi-agent system based on fact sharing and powered by LLM, which is used to simulate real emergency call scenarios. Traditional LLM medical agents face issues such as poor real-time adaptability and insufficient multi-step reasoning capabilities in high-risk environments like ICUs. Xu et al. [307]replaced GPT-4 with o1 as the main LLM for the three medical agent frameworks, namely CoD [33], MedAgents [260], and AgentClinic [232], to study the impact of o1 on agent reasoning, tool usage adaptability, and real-time information retrieval in different clinical scenarios. They found that the diagnostic accuracy was significantly improved, but o1 performed worse than GPT-4 in simple tasks. The triage and treatment planning in the emergency department require comprehensive integration of multimodal patient data and adherence to the KTAS standard. However, manual assessment is inefficient. Han et al. proposed an LLM-driven clinical decision support systems(CDSS) [93], aiming to assist emergency room doctors and nurses in patient triage, treatment planning, and overall emergency management.

**Dentistry.** Agentic systems are beginning to extend into dental imaging, a domain where vision–language models underperform task-specific models on individual tasks. OPGAgent proposes a multi-tool agentic system for auditable orthopantomogram (OPG) interpretation, coordinating specialised perception modules through a Hierarchical Evidence Gathering module that decomposes analysis into global, quadrant, and tooth-level phases with dynamically invoked tools, a specialised toolbox encapsulating spatial, detection, utility,

and expert models, and a Consensus Subagent that resolves conflicts through anatomical constraints [328]. The accompanying OPG-Bench protocol, based on (Location, Field, Value) triples derived from real clinical reports, enables comprehensive review of findings and hallucinations beyond conventional VQA indicators.

**Pathology.** Computational pathology is rapidly evolving with the integration of foundation models and agentic AI. Da et al. provide an international expert assessment of how these systems can be responsibly integrated into medical practice, connecting deployable clinical relevance with technical maturity, operational readiness, and economic and regulatory context [50]. LAMMI-Pathology adopts a tool-centric, bottom-up LVLM-agent architecture in which customised domain-adaptive tools are clustered by domain style to form component agents, coordinated through a top-level planner that avoids excessively long context lengths inducing task drift, and introduces Atomic Execution Nodes for trajectory construction and trajectory-aware fine-tuning [253].

**Dermatology.** Dermatology is a discipline dedicated to the diagnosis, treatment and prevention of diseases related to the skin, hair, nails and related mucous membranes. In the context of dermatology, existing visual language models (VLMS) often lack explicit modeling of dermatological medical guidelines during diagnosis. Vashisht et al. [272] used GPT-4V with Naive Chain-of-Thought for retrieval and employed LLM-driven multi-agent conversations for comment-based diagnosis, which can accurately diagnose dermatological conditions early. The manually designed medical agent workflow lacks flexibility and scalability. To address this issue, Zhuang et al. [362] were inspired by automatic machine learning (AutoML) and designed the first LLM-driven fully automatic design framework for medical agents, significantly improving the accuracy of dermatological diagnosis. Med-GRIM [178] is a model specifically designed for medical VQA tasks. It utilizes graph-based retrieval and prompt engineering to integrate domain-specific knowledge, enabling both low computational requirements and maintaining the accuracy and robustness of responses. Thakrar et al. proposed a two-layer LLM architecture [262] that directly simulates the clinical workflow and is used for the multimodal medical visual question-answering task in remote dermatology diagnosis.

**Ophthalmology.** Ophthalmology is a specialized field that studies diseases of the visual system, mainly involving disorders of the eyeball and its associated tissues. LLM lacks domain-specific knowledge in ophthalmology and has insufficient ability to interpret medical images, with a low degree of task integration. Wu et al. proposed the LLM-driven multi-tool agent ChatMyopia [301], This system provides personalized, accurate, and secure responses for tasks such as myopia macular lesion grading and individual question consultation. MLLMs are prone to generating illusions in ophthalmic diagnosis, lacking transparency and traceability. Pan et al. proposed EH-bench [208] to assess the illusion problem of MLLMs and designed a LLM-driven agent-centered three-stage framework, which significantly alleviates both visual understanding and logical synthesis types of illusions. To enhance the ability of LLM to interpret heterogeneous ophthalmic reports, Meng et al. proposed a knowledge-driven multi-agent system (MAS) [188], which decomposes ophthalmic surgery planning into collaborative expert agents. To address the issues of hallucinations, limited interpretability, and insufficient specific domain medical knowledge faced by LLM in glaucoma diagnosis, Liu et al. proposed a multi-agent diagnostic framework named MedChat [169], which combines professional visual models with multiple LLM agents of specific roles, all coordinated by a chief agent. To address the issue of the lack of deep annotations, high-quality, multimodal visual instruction data for intelligent ophthalmic diagnosis, Li et al. proposed Eyecare Kit and designed the EyecareBench benchmark [153]. They also developed EyecareGPT, which adopts an adaptive resolution mechanism and hierarchical dense connectors to enhance visual understanding. Shi et al. proposed EyeAgent [242], an LLM-driven agentic framework that orchestrates 53 validated ophthalmic tools across 23 imaging modalities with textbook-grounded RAG to provide interpretable clinical decision support.

**The Rare Disease Diagnosis and Treatment Center** The Rare Disease Diagnosis and Treatment Center is a comprehensive platform dedicated to the diagnosis, treatment, follow-up and research of rare diseases. It is a specialized center that involves multi-disciplinary collaboration (MDT) rather than a single department. The current medical agent framework is difficult to adapt to the complex needs of rare diseases. To address this issue, Chen et al. proposed RareAgents [41], which is the first LLM-driven multi-disciplinary team decision support tool. Chen et al. developed a multi-agent dialogue (MAC) framework for disease diagnosis, simulating the multidisciplinary team (MDT) discussions of a real medical team [40]. This framework significantly enhances the diagnostic capabilities of LLMs. Zhao et al. proposed DeepRare [345], which is

the first rare disease diagnosis agent system driven by LLM. It can handle heterogeneous clinical inputs and generate graded diagnostic hypotheses for rare diseases. The accuracy of LLMs in prioritizing genes for rare genetic diseases is unknown. Although they perform well in medical examinations, their performance declines when dealing with large gene sets, and there are positional deviations and literature deviations. Neeley et al. [195] used multiple LLMs to generate evaluations of gene-phenotype associations, identified case-specificity through HPO classification, and adopted a divide-and-conquer strategy to have LLMs process 5-gene groups instead of the entire gene set, avoiding the positional sensitivity of LLMs.

**Challenges.** Medical Agents still face several shared limitations. In ophthalmology, multi-agent VLM systems and graph- or knowledge-driven retrieval such as EyecareGPT [153] improve visual reasoning, yet fine-grained cross-modal alignment and traceable evidence chains across heterogeneous images and reports remain difficult. In emergency and triage settings, agent mixtures and protocol-aware clinical decision support systems (e.g., MoMA [80], KTAS-based systems [93], dispatch MAS [157]) enhance multimodal fusion and reasoning, but real-time adaptation to noisy inputs and uncertainty-aware escalation or deferral are still underexplored. In dermatology, LLM-driven multi-agent dialogues, AutoML-style workflow search, and knowledge-driven MAS such as DermPrompt [272] and Zhuang et al. [362] reduce hallucinations and improve accuracy, yet workflows remain brittle and hard to generalise across lesions, procedures, and care settings. In rare disease diagnosis and treatment center , agentic systems such as RareAgents [41] and DeepRare [345], together with LLM-based gene prioritisation, struggle with extreme data sparsity, position and literature biases, and the need for robust MDT-style collaboration on ultra-rare, diagnostically uncertain cases.

### 5.6 Cross-departmental Applications

The cross-departmental Agent is an intelligent agent system that can span multiple medical specialties and fulfill unified diagnosis tasks for patients from different departments, addressing their diverse diagnostic needs. It needs to be implemented within a unified framework, through the integration of multimodal and multi-source information and the orchestration of tools, to support interactive clinical reasoning across multiple specialties and tasks, while maintaining an interpretable and traceable decision-making process.

- **Surgical navigation and multimodal collaborative analysis** In response to the complexity of surgical and imaging analysis, SCOPE [181] innovatively combined voice commands with visual models to achieve real-time segmentation and tracking of surgical instruments and anatomical structures. Chen et al. [36] proposed an intermediary guidance multi-agent framework and TissueLab [154], which respectively utilized multi-expert intelligent agents' collaboration and human-machine co-evolution mechanisms to effectively solve cross-modal medical decision-making problems covering fields such as pathology, radiology, and ophthalmology.

- **Clinical reasoning practice and automation of research processes** To replicate the real clinical thinking process, MEDDxAgent [226] simulated the doctor's differential diagnosis (DDx) process through a multi-agent iterative update mechanism; MedAgentSim [11] constructed a fully realistic simulation environment with doctor-patient roles to dynamically evaluate diagnostic performance; FRAME [325] extended this approach to the field of research, using a feedback-driven iterative framework to achieve the automation and quality improvement of medical paper generation.

- **Tool Arrangement and Adaptive General Framework** To overcome the limitations of a single model in integrating knowledge and tools, MedOrch [102] and MMedAgent [146] adopted a modular tool arrangement architecture, enabling flexible scheduling and transparent reasoning for various external tools such as medical image analysis and web search; on this basis, MMedAgent-RL [302] introduced a reinforcement learning mechanism, allowing the intelligent agent to conduct dynamic collaboration optimization across multiple specialties such as radiology and neurosurgery, significantly enhancing the generalization ability of the system. MACRO shifts from static tool composition to experience-driven tool discovery: from verified execution trajectories, it autonomously identifies recurring effective multi-step tool sequences, synthesises them into reusable composite tools, and registers these as new high-level primitives that continuously expand the agent's behavioural repertoire across diverse medical imaging datasets [62]. Meissa brings full agentic capability offline

as a lightweight 4B-parameter medical MM-LLM, learning both when to engage external interaction (strategy selection) and how to execute multi-step interaction (strategy execution) by distilling structured trajectories from frontier agent systems, matching or exceeding proprietary frontier agents in 10 of 16 evaluation settings across 13 medical benchmarks [44]. MedScope extends tool-using clinical video reasoning to long-form procedures by interleaving intermediate reasoning with targeted tool calls and verification on retrieved observations, grounding predictions in temporally localised visual evidence via Grounding-Aware GRPO [155].

**Challenges.** A common challenge is the complexity of the tool and model ecosystem, which makes it difficult to expand and maintain. Different departments require completely different tool chains. The traditional approach is to build a separate set of models for each department, which incurs extremely high maintenance costs and makes it difficult to migrate to new tasks. Unified scheduling of multiple imaging tools, through tool-calling, for cross-modal and cross-department tasks such as MMedAgent [146] can effectively alleviate this problem. Another challenge is the hallucination problem caused by multi-modal high-dimensional data. By decomposing the tasks and having each part completed by a dedicated model, and integrating the results by LLM, this can explicitly reduce the risk of hallucination.

## 6 Medical Agents: Applications Across Clinical Workflows and System Operations

The departmental perspectives in Section 5 showed how medical agents are being prototyped and deployed within individual specialties. In this section, we reorganize the landscape along the patient's journey through the healthcare system,

from the moment when patients arrive at the "front door" of the system to the backstage infrastructures that sustain day-to-day operations. We begin with agents that sit closest to patients in triage, consultation, and clinical decision support, then move to systems that assist high-stakes procedures and multimodal diagnostic reasoning. Building on these frontline capabilities, we next consider agents for longitudinal health management, patient communication, and professional training, where short encounters are extended into ongoing guidance and education. Finally, we turn to the back-office layer, where agents manage documentation and knowledge assets, support research workflows, and coordinate administrative, financial, and regulatory processes that shape how clinical work is actually delivered.

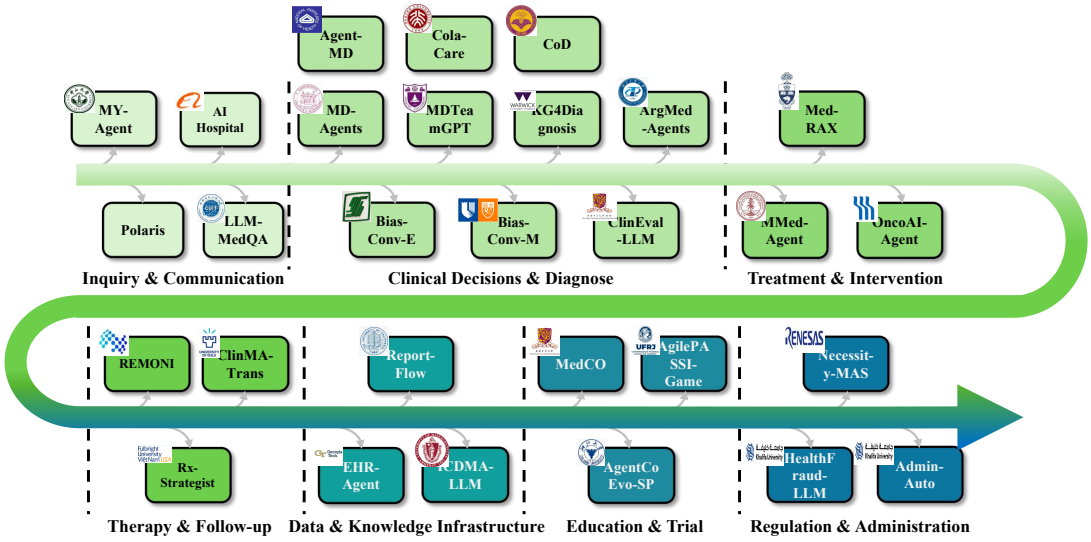


Figure 5: The application landscape: from frontline encounters to back-office operations.

### 6.1 Early Stage of the Care Pathway: Intake and Clinical Dialogue

At the start of the patient journey, medical agents interact with patients before any imaging or procedures, turning vague complaints into structured information, guiding patients through reception and triage, answering questions, and ensuring conversations are safe and trustworthy. This work focuses on five themes: diagnostic questioning, full outpatient flow simulation, disease-specific counselling and role-aware dialogue.

#### 6.1.1 Optimizing diagnostic questioning

Several studies treat the sequence of questions as a learnable policy. A two-stage prompting strategy broadens differential diagnoses and then narrows down to high-yield cues, cutting redundant turns while covering plausible conditions. *Doctor-R1* models interactive diagnosis as a reinforcement learning problem, rewarding strategic and empathic questioning to improve communication and performance on HealthBench and MAQuE [158; 139]. *MedClarify* takes an information-theoretic approach, computing a list of candidate diagnoses analogous to a differential diagnosis and proactively generating follow-up questions aimed at reducing diagnostic uncertainty by selecting the question with the highest expected information gain, reducing diagnostic errors by ∼27 percentage points over single-shot baselines [293]. *MIND* unifies inquiry and diagnosis within a single reinforcement learning framework for psychiatric consultation, building a Criteria-Grounded Psychiatric Reasoning Bank that retrieves semantically similar reference consultations and distils reusable criteria-grounded supports, while rubric-based process rewards provide fine-grained supervision over intermediate decision steps [149]. *ATPO* formulates multi-turn medical dialogue as a Hierarchical Markov Decision Process and proposes an uncertainty-aware Adaptive Tree Policy Optimization algorithm that adaptively allocates rollout budgets to high-uncertainty states, enabling smaller models to surpass GPT-4o on dialogue benchmarks [28]. *EvoClinician* introduces a self-evolving agent that learns efficient diagnostic strategies at test time via a "Diagnose–Grade–Evolve" loop, where a Process Grader evaluates each action for clinical yield and resource efficiency and an Evolver updates the Actor's strategy [103]. Other work focuses on transparency and clinical plausibility: an interpretable system links each follow-up question to its purpose, while a causality-aware framework ensures the agent follows clinically meaningful paths, not superficial text patterns [172; 164]. Large-scale actor–critic architectures adapt inquiry policies to traffic and noisy inputs, enabling online agents to respond to distribution shifts rather than fixed scripts [100]. These systems aim to optimize "what to ask next" as a key design goal.

#### 6.1.2 Outpatient workflows and multi-role consultation

Another line of work reconstructs the entire outpatient encounter as a chain of coordinated roles. A reception-to-physician pipeline simulates how a patient moves from front-desk intake through triage to specialist handoff, maintaining context so downstream agents can reuse earlier information instead of re-asking questions [19]. For rare diseases, a multi-disciplinary consultation framework assigns specialist agents to different organ systems and allows them to merge hypotheses into a single conversation, mimicking real-world MDT practice [42]. Large-scale hospital simulators like *AI Hospital* populate an entire ward with LLM-driven doctors, nurses, and patients, using the sandbox to stress-test diagnostic conversations under safety, coverage, and failure constraints before deployment [63]. Closer to point-of-care use, a lightweight *Healthcare agent* splits consultations into planning, memory, and reporting stages, while self-evolving systems like *SeM-Agents* let doctor and helper roles update each other, making the virtual clinic more structured, rational, and cautious over repeated cases [222; 35]. At the dialogue-management level, *DiagGPT* tracks topics, goals, and action items over turns, keeping long diagnostic conversations coherent while switching between data gathering, hypothesis revision, and recommendation [27]. These designs move from single-turn QA to distributed responsibilities across conversations.

#### 6.1.3 Patient education and condition-specific counselling

A third group of systems supports patients and caregivers who need explanations rather than diagnoses. *NoteAid-Chatbot* turns discharge summaries into structured discussions that resemble human teaching, balancing clarity, informativeness, and patient satisfaction with reinforcement learning [118]. In primary eye care, *ChatMyopia* acts as both a pre-consultation agent for myopia and a caregiver agent for newborn

auricular deformities, providing accessible explanations grounded in specialist knowledge [301; 281; 356]. The *Multi-OphthaLingua* benchmark targets multilingual ophthalmology QA in low- and middle-income settings, using RAG-enhanced dialogue to reduce cross-lingual performance gaps [223]. These agents translate expertise into patient-friendly dialogue.

#### 6.1.4 Safety, role, and affect-aware dialogue

Several works focus on the outer layer of conversation—identity, ethics, safety, and emotion. A multilingual, privacy-focused framework ensures inquiry agents operate across languages and jurisdictions while respecting data-protection rules, using configurable policies to control sensitive information [237; 156]. For emotional well-being, grounded conversational agents with embodied avatars recreate aspects of in-person presence, improving comfort and adherence when patients talk to empathetic agents [312]. *Polaris* organizes a constellation of LLM agents around long, safety-critical patient conversations, matching human nurses on safety, readiness, and bedside manner [193]. *T-Agent* models when and how medical terms should be introduced and rephrased, ensuring technical vocabulary supports understanding [110]. These systems prioritize safety and affect as key design goals.

### 6.2 Decision-Making at the Clinical Core: Virtual MDT Teams and Multimodal Reasoning

After triage and consultation, agentic systems at the clinical core process collected information into diagnostic assessments and care plans. In this subsection, we focus on diagnostic decision-making agents that operate as virtual multidisciplinary teams, learn from clinical experience, and perform radiologist-like multimodal reasoning and reporting. These systems sit between front-door dialogue agents and downstream treatment and follow-up agents: they consolidate signals from earlier stages, ground decisions in tools and guidelines, and make their reasoning more inspectable for clinicians.

#### 6.2.1 Virtual MDT-style teams for complex diagnosis and management

Medical agents are treated as virtual MDT members who share responsibility for diagnosis and care planning. *MDAgents* routes cases to solo agents or multi-agent teams based on task complexity, mirroring MDT practice and improving diagnosis benchmarks while minimizing collaboration costs [133; 132]. Yuan et al. demonstrate that *mixed-vendor* multi-agent teams, where doctor agents are instantiated from different model families (e.g., o4-mini, Gemini-2.5-Pro, Claude-4.5-Sonnet), consistently outperform single-vendor configurations by pooling complementary inductive biases that surface correct diagnoses individual models collectively miss [329]. *MedCollab* emulates hierarchical consultation workflows with a dynamic specialist recruitment mechanism, an IBIS-structured argumentation protocol requiring traceable evidence, and Hierarchical Disease Causal Chains that structure diagnostic predictions through explicit logical operators [334]. *MedCoRAG* combines UMLS knowledge graph path retrieval with clinical guideline pruning, then performs multi-agent collaborative reasoning where a Router Agent dispatches Specialist Agents based on case complexity for iterative deliberation on hepatic disease cases from MIMIC-IV [161]. *Vivaldi*, a role-structured multi-agent system for interpreting multivariate physiological time series in emergency care, reveals that agentic pipelines substantially benefit non-thinking models while explicit tool-based computation is decisive for codifiable clinical metrics [77]. *MedCoAct* connects doctor and pharmacist agents in a confidence-aware loop, enabling them to revise diagnostic hypotheses and medication plans together, especially useful in telemedicine with incomplete information [351]. Multi-role frameworks like *MDTeamGPT* and *MAM* split work between general practitioners, specialists, and coordinators, pooling multimodal inputs while ensuring each role's scope remains interpretable and audit-friendly [34; 355]. Knowledge-centric designs like *KAMAC* and *KG4Diagnosis* expand collaboration by recruiting specialists or traversing knowledge graphs as needed, aiding in broad differential diagnosis and tasks like cancer prognosis [299; 363]. Logic-driven systems such as *MedLA* and *ArgMed-Agents* require explicit logic trees or argument graphs, resolving inconsistencies before making final decisions, turning MDT discussion into structured reasoning chains [177; 106]. Multi-agent systems for ICU and ward management use similar modular designs to predict mortality and length-of-stay, outperforming traditional baselines while remaining transparent and controllable [40].

#### 6.2.2 Interactive diagnostic agents that learn from clinical experience

Rather than relying on one-shot question answering, a focus on interactive consultation and learning from experience is emerging. In the *AgentClinic* environment, doctor agents adapt to incorrect decisions by updating hypotheses and tool use based on feedback, studying how different reasoning policies affect efficiency [232; 57]. Experience-driven systems like *PPME* train agents using millions of electronic medical records, improving diagnostic performance and narrowing the gap between interactive and full-information diagnoses [257]. Workflow-centric systems such as *MedChain-Agent* extend this approach to sequential care, reusing case-derived feedback to improve performance on benchmarks [166]. Lightweight agents like *IMAS* adapt consultation planning, safety checks, and triage to rural and resource-limited settings, assisting semi-trained workers with decision-making on when to escalate care [78]. Scenario-specific systems, like those for guiding clinicians during intracerebral hemorrhage transfers, embed interactive designs into critical workflows [131]. Architecture-search systems like *Learning to Be a Doctor* optimize diagnostic workflows over time based on feedback [362].

#### 6.2.3 Radiologist-like multimodal reasoning and reporting

In one group of studies, agents act as collaborative radiologists, analyzing images, linking them to guidelines, and generating structured reports. In chest X-ray analysis, systems like MAGDA and MedRAX/ChestAgentBench decompose the process into retrieval, image reading, guideline-based reasoning, and explanation, with each step grounded in evidence [18; 61]. CXRAgent adds an uncertainty-aware layer, modeling when evidence is weak or conflicting, incorporating follow-up recommendations [236]. Zhuang et al. and Zhong et al. fuse CTPA detection, report generation, and outcome prediction into single pipelines, making chest imaging agents capable of covering diagnosis, documentation, and prognosis in one workflow [361; 353].

Beyond single-task chest pipelines, several systems extend radiology reasoning across tasks and modalities. *Med-VRAgent* combines a visual backbone with an LLM planner for radiology VQA, optimizing tool calls for efficiency [92]. *MMedAgent* integrates segmentation, measurement, and retrieval tools, enabling an LLM to orchestrate free-form multimodal queries [146]. Inquire–Interact–Integrate takes a similar approach, interacting with external tools and fusing multi-step evidence before providing a diagnosis [91]. Other systems incorporate clinical context: Peng et al. integrate images, EHR, and dialogue history to guide decision-making [212], while *EyecareGPT* combines fundus images and text for triage, diagnosis, and counseling [153].

A third line of imaging agents targets complex cases like rare diseases and volumetric scans. *DeepRare* integrates phenotypic features and imaging with literature retrieval, supporting rare-disease hypotheses [345]. *Med-GRIM* encodes images and reports into a graph structure, improving consistency and interpretability on radiology VQA tasks [178]. *CT-Agent* focuses on 3D CT scans, combining segmentation, measurement, and report-generation tools into a single loop [182]. *Proof-of-TBI* ensembles vision–language models to assess traumatic brain injury from CT scans, checking internal consistency before making triage recommendations [88]. These systems expand imaging agents from classifiers to programmable colleagues capable of reasoning and justifying decisions across images, text, and external knowledge.

### 6.3 Treatment Procedures: From Guided Interventions to Surgical Robots

After diagnostic decision-making has produced a provisional care plan, agents increasingly move from recommending options to supporting concrete therapeutic procedures and surgeries. In this subsection we focus on agents that help execute treatment: tool- and guideline-grounded planners that structure medication and procedure workflows, systems that use imaging and measurements to guide interventions, and surgical robots and navigation agents that assist with operative decisions and control. These applications sit between the diagnostic agents of the previous subsection and the longitudinal therapy and follow-up agents discussed later, helping to transfer clinical plans into consistent actions in practice.

#### 6.3.1 Tool- and guideline-grounded planners for calculators and treatment workflows

Some systems integrate agents with clinical calculators and knowledge bases to ensure decisions are numerically sound and policy-compliant. *AgentMD* automates the execution of clinical calculators, selecting appropriate

scores and calculating risks more reliably than vanilla GPT-4, illustrating the value of tool learning for bedside risk management [123]. *MeNTi* generalizes this by enabling meta-tool selection and nested tool calls, evaluated on the CalcQA benchmark where agents must select the right calculators, fill parameters, and interpret health status [359].

Guideline-based planners are prevalent in oncology and internal medicine. *CancerGUIDE* estimates internal disagreement to decide when guideline-based NSCLC recommendations are reliable and when extra review is needed [268]. The broader "natural language programming" approach uses retrieval-augmented agents to execute evidence-based workflows from free-text instructions, mapping guideline logic into executable plans instead of treating queries as isolated QA problems [269]. At the infrastructure level, *MedicalOS* converts clinician instructions into sequences of EHR queries, order entries, reporting, and treatment-planning commands, ensuring compliance with guidelines inside hospital systems [357]. Multi-agent CDSS systems like *ColaCare* combine EHR expert models with DoctorAgents, MetaAgents, and MSD-guideline RAG to improve interpretability in mortality and readmission predictions [287]. Adaptive systems, such as the one integrating BioBERT-based retrieval with web resources, support decisions while modeling uncertainty in MedQA and MedBullets-style decision-making [202]. Knowledge-fusion designs like *MedGen* support clinical goal setting, data collection, argument linking, and plan selection, ensuring recommendations for multimorbidity (e.g., breast cancer and depression) remain traceable and justifiable in multidisciplinary settings [173].

#### 6.3.2 Imaging-guided intervention and treatment planning

After interpreting images, agents assist in planning treatments and guiding interventions. In oncology and radiotherapy, *AutonomousAI* links GPT-4, vision transformers, and OncoKB to parse images and retrieve genomic and guideline knowledge, synthesizing personalized treatment plans for tumor boards [72]. Zero-shot IMRT agents generate treatment plans from contouring and prescription information, reducing trial-and-error in radiotherapy workflows [314]. Other systems focus on procedural guidance: a multimodal agent for focused ultrasound adapts sonication strategies based on pre-operative imaging and device parameters, while *CataractSurg-80K* predicts intraocular lens power and post-operative risk from multimodal data [344; 188]. These agents not only describe what they see but also propose, score, and sometimes automatically instantiate treatment options.

#### 6.3.3 Surgical robots and navigation agents

At the far end of the perception–action spectrum, agents directly interact with surgical robots and navigation systems. *SurgBox* provides a platform where agents collaborate as surgical copilots: one plans steps, while others monitor sensors or translate commands into robot trajectories [294]. This platform supports training and evaluation before deployment in real operating rooms, orchestrating language, control, and simulation components within one framework. For vascular interventions, a navigation agent learns a latent model of catheter motion from historical data, using it to plan safe paths under fluoroscopy [225]. By embedding agents into robotic and navigation loops, these systems extend the agentic paradigm beyond reasoning into physically embodied assistance, linking perception directly to intervention along the patient journey.

### 6.4 From Therapy to Follow-up: Chronic Disease Management and Prescription Safety

Once diagnoses are made and treatments planned, care shifts to long-term therapy, monitoring, and safe medication use. Agent-based systems in this stage connect physiological signals, prognosis models, and prescription workflows to detect deterioration early, control drug risks, and manage patients both at home and in the clinic. Recent work includes real-time and remote monitoring, disease-trajectory prediction, digitized prescription pipelines, and conversational safety checks to monitor risk in everyday interactions.

#### 6.4.1 Real-time monitoring and remote health management

For unstable patients, agents monitor vital signs continuously. Shaik et al. propose a multi-agent reinforcement learning framework where agents track heart rate, blood pressure, and other metrics, adapt policies, and raise alarms as needed. This approach aims to improve sensitivity while reducing false alarms and personalizing alerts [234].Remote monitoring extends this concept to home care. REMONI connects wearable devices to

multimodal LLMs, allowing real-time feedback to patients and trend summaries for clinicians [105]. This shows how continuous monitoring can offer personalized supervision outside of clinical settings.

#### 6.4.2 Disease trajectory prediction and long-term prognosis

Some agents predict how diseases evolve over time. CLIMATv2 uses a multi-agent transformer where one module mimics a radiologist analyzing imaging data, while another integrates non-imaging clinical data [196]. This model provides forecasts for conditions like knee osteoarthritis and Alzheimer's disease, reflecting how long-term prognosis is managed in practice.

#### 6.4.3 Digitizing prescriptions and pre-dispensing safety checks

Medication workflows are becoming agent-driven. *RxLens* digitizes handwritten or scanned prescriptions by converting them into structured entries through OCR, entity extraction, and retrieval-augmented prompting to handle noise and missing information [117]. *Rx Strategist* treats prescription verification as a series of safety checks. Agents evaluate the indication, dose, schedule, and potential drug interactions, supported by a knowledge graph for evidence collection [271]. These systems transform prescriptions from unstructured notes to transparent decision trails, ensuring each step and its evidence are inspectable.

#### 6.4.4 Risk monitoring and safety gates in conversational care

Risk also arises in everyday conversations when patients inquire about treatments or adjust doses. *MedTAMARA* embeds safety checks into dialogue by framing risk evaluation as a multi-agent review process, where agents critique each other's responses before a final, trust-weighted recommendation is made [295]. This system adds a safety layer, complementing monitoring and prescription agents, ensuring long-term care aligns with clinical standards even in informal conversations.

### 6.5 Making Data Usable: Documentation, Coding, and Knowledge Infrastructure

Facing large volumes of text, audio, and structured records, systems are built to make this data usable. These agents clean and reshape clinical narratives, turn speech and free text into analyzable formats, manage codes and EHRs, and maintain knowledge bases for downstream use. They play a crucial role between care delivery (Sections A–D) and learning environments (Section F), stabilizing the information that supports daily care and long-term analysis. Recent work focuses on four themes: document quality, speech transcription, coding and EHR analytics, knowledge-graph infrastructure, and research operations for data collection and manuscript generation.

#### 6.5.1 Document quality, speech transcription, and patient-friendly summaries

Multi-agent workflows target clinical documentation quality and accessibility. *RadCouncil* coordinates drafting, reviewing, and consensus agents to produce readable radiology reports, reducing inter-institution variation [331; 270]. On the patient-facing side, a Reflexion-style agent rewrites expert reports into lay-friendly letters, and "Society of Medical Simplifiers" uses agents to improve readability while preserving key content [254; 175].

Other workflows align notes with guidelines. *MedReAct'N'MedReFlex* links retrieval-augmented access to corpora with critique and formatting steps to make error detection an auditable process [49; 233]. ASR-LLM Benchmark uses Whisper-based medical ASR to transcribe, structure, and evaluate clinical conversations [137; 142]. *ExperienceWeaver* addresses the small-sample setting by shifting from data retrieval to experience learning: instead of simply recalling past examples, it distils noisy, multi-dimensional feedback into structured, actionable knowledge (error-specific Tips and high-level Strategies) and injects this into an agentic pipeline, teaching the model "how to revise" clinical text rather than "what to revise," surpassing Gemini-3 Pro in small-sample settings [304]. These systems treat documentation as an iterative, multi-role process rather than a static output.

#### 6.5.2 Coding, claims, and agentic EHR analytics

Agentic systems support coding, billing, and analytics over EHR data. *Code Like Humans* frames ICD, CPT, and PCS assignments as a multi-agent task, improving interpretability and accuracy compared to black-box models [192]. For large EHRs, *EHRAgent* turns natural-language queries into multi-step SQL programs for cohort and temporal analysis [243]. *EMR-AGENT* automates cohort definition and feature extraction across multiple EMR schemas [143]. *Exploring LLM Multi-Agents for ICD Coding* stages roles over SOAP-structured notes, improving transparency and rare-code performance [151]. *AgentEHR* challenges agents to execute complex decision-making tasks such as diagnosis and treatment planning requiring long-range interactive reasoning directly within raw, high-noise EHR databases, and introduces RetroSum, a retrospective summarisation framework that dynamically re-evaluates interaction history to prevent information loss and ensure logical coherence, achieving performance gains of up to 29.16% over baselines [163]. *TRACE* enables temporal clinical reasoning with frozen LLMs over streaming EHRs by explicitly structuring context via a dual-memory architecture of static Global Protocol and dynamic Individual Protocol, with four agentic components coordinating to maintain bounded inference cost and auditable reasoning traces [220]. *SELSM* addresses the privacy-constrained deployment bottleneck by distilling simulated clinical trajectories into entity-agnostic operational rules and dynamically fetching these logical priors via a Query-Anchored Two-Stage Retrieval mechanism, achieving 100% task completion rates on FHIR-based EHR tasks [317]. These systems treat coding and EHR analysis as cooperative processes grounded in care semantics.

#### 6.5.3 Research operations: data collection, follow-up, and manuscript generation

Agent frameworks stabilize research workflows and longitudinal follow-up. *OpenLens AI* links literature, analysis, and writing agents to generate publication-ready manuscripts, improving transparency and reusability [45]. For longitudinal studies, *SmartState* assigns agents to guide protocol interactions, parse responses, and check adherence, improving data completeness and compliance [13]. These research agents connect clinical data, knowledge infrastructure, and the evidence base for future medical agents.

### 6.6 Simulation and Support Systems: For Education, Training and Trial

While the previous subsections focused on agents embedded directly in clinical care and hospital operations, an equally important line of work uses medical agents in simulated or indirect support roles. Instead of acting on real patients or live workflows, these systems create training grounds and decision aids for humans and models: virtual hospitals and patient simulators for education and skills training, agent-based environments for stress-testing clinical workflows, and multi-agent tools that support the design and monitoring of clinical trials.

#### 6.6.1 Medical education and training

Agent-based systems in medical education move beyond static textbooks and OSCE checklists, offering interactive environments where learners can practice reasoning, teamwork, and patient-centered dialogue. *MEDCO* organizes a teaching triad with a patient agent, clinician, and radiologist, allowing students to negotiate diagnostic reasoning and inter-specialty collaboration [291]. *ChatCoach* places a patient agent in realistic scenarios, while a coach agent analyzes the dialogue and provides feedback on empathy, clarity, and terminology. Schema-guided virtual patients like *SOPHIE* encode clinical pathways and emotional states, supporting reproducible, tailored OSCE-style encounters [112; 125]. A VR anatomy platform with a generative assistant allows learners to ask questions and study how avatar embodiment and task difficulty affect learning outcomes [46].

Simulated patient ecosystems treat both patients and clinicians as evolving agents in a virtual hospital. *Agent Hospital* creates a self-play environment where doctor, nurse, and patient agents interact, improving doctor policies through exposure to synthetic cases [150]. *EvoPatient* generates patient behaviors through co-evolution with doctor agents, aligning encounters with task requirements and human preferences without exhaustive scripting. *AIPatient* combines a clinical knowledge graph with retrieval-augmented generation to create simulated patients capable of handling history-taking and QA [56; 326]. *Patient-Zero* generates synthetic longitudinal records and patient agents from hierarchical knowledge, enabling MedQA-style training

without real charts [140]. LLM-driven VR avatars give virtual patients distinct personalities, handling ASR errors and label imbalances to improve intent classification in noisy environments [53; 258].

#### 6.6.2 Clinical research and trial support

In clinical research, multi-agent systems break down complex prediction tasks into domain-aligned parts. *ClinicalAgent* builds a trial-outcome prediction framework where agents focus on protocol understanding, efficacy, safety risks, and enrollment issues, instead of producing a single opaque forecast [330]. Each agent traces its reasoning from mixed trial data and textual reports, and their outputs are combined into a final prediction. This role-based structure improves accuracy and produces explanations that mirror how human trialists discuss benefit-risk balance, making decisions easier to critique and use in study design or evidence assessment.

### 6.7 Regulation, Payer Workflows, and Administrative Automation

Beyond bedside care, medical agents increasingly manage the broader processes in healthcare, such as regulation, payer workflows, fraud detection, and office administration. These systems model regulators, insurers, and hospital back offices as interacting agents, aiming to make complex processes more transparent and scalable.

On the regulatory side, Han et al. model regulators, manufacturers, and competitors as LLM-driven agents in a feedback-based simulation environment [94]. Their *Regulator–Manufacturer Agents* framework uses shared policy documents, market data, and compliance rules to help agents negotiate device approvals and adapt to regulatory changes, creating a sandbox for stress-testing rules without exposing real patients.

For payer workflows, *MAS-PA* transforms prior authorization review into a checklist-style process: agents match patient records to guidelines, filling in criterion slots and aggregating them into a final decision [209]. This structured approach makes it easier for human reviewers to understand the reasoning behind acceptance or denial and identify unclear or missing evidence.

A second group of systems focuses on financial integrity and billing. *LLM-HealthFraud* links blockchain-protected claim data with retrieval-augmented agents to detect fraudulent patterns like upcoding or unnecessary procedures [116]. The system surfaces high-risk cases with supporting evidence, complementing human auditors. Uddin et al. treat billing optimization as a sequential decision problem, where a reinforcement-learning agent learns to balance reimbursement, clinical appropriateness, and policy limits [267].

Administrative automation applies similar concepts to everyday operations. *Admin-Auto* coordinates multiple agents to handle tasks like registration, appointment scheduling, and documentation, breaking down high-level goals into tool calls across hospital systems [83]. These systems emphasize transparency, auditability, and consistency, alongside predictive performance.

### 6.8 Open Issues and Our Next Steps

This survey shows that medical agents now appear at almost every step of the patient journey, from first questions at the front desk to billing and research support in the back office. Yet most systems remain early prototypes. They are often tested on small datasets, in single hospitals, or in short simulations, so we still know little about how they work in busy, real clinics with diverse patients and staff.

Several limits are shared across layers. Many agents focus on a single narrow task, such as dialogue, imaging, or coding, and are rarely evaluated when workflows change or when data comes from new institutions. Safety and fairness controls are uneven: some systems add simple filters for hallucinations or risky advice, but few treat safety, bias, and traceability as central design goals. Most agents also depend on local tools and data silos, which makes reuse and common benchmarking difficult.

These gaps suggest clear next steps. In the near term, we need more end-to-end evaluations where agents are embedded in real or high-fidelity workflows and are measured on clinical impact, workload, and equity, not only on task scores. Future systems are likely to move from single agents to coordinated teams that share memories, tools, and safety rules across consultation, diagnosis, documentation, and management. It will be

| Application Layer | Sub-fields | Detail Types | Representative Papers |
|---|---|---|---|
| Front-line Consultation and Question Answering | Multi-turn dialogue | symptom querying, causal diagnosis | MyDoctor [172], CausalDiag [164] |
| | Patient education | discharge education, ophthalmology | Multi-OphthaLingua [223] |
| | Safety- and role-aware chatbots | safety-focused frameworks | Polaris [193] |
| | Evidence-grounded QA | retrieval, case-based, KG/RAG QA | TreeRec [318], LLM-MedQA [316], AMG-RAG [224] |
| Clinical Diagnosis and Decision Support | Virtual MDT teams | multi-specialist, KG-driven reasoning | MDAgents [132], MDTeamGPT [34], MAM [355], KG4Diagnosis [363], ArgMed [106] |
| | Experience-adaptive agents | simulated clinics, adaptive decision-making | AgentClinic [232], ARMA [57], Med-Chain [166] |
| | Tool- and guideline-grounded planning | clinical toolboxes, guideline workflows | AgentMD [123], MeNTi [359], CoLaCare [287] |
| | Evaluation, bias mitigation | diagnosis tracing, debiasing, structured exams | CoD/DiagnosisGPT [33], Debias-Agents [128; 129], ReflecTool [162], LCP-RAE [168], AI-SCE [185] |
| Perception, Imaging, and Intervention | Imaging reasoning and reporting | chest X-ray, radiology, multimodal QA | MedRAX/ChestAgent [61], MAGDA [18], Unc-CXR [236], MMedAgent [146], CtxDiagQA [212], EyecareGPT [153] |
| | Imaging-guided therapy and surgery | autonomous oncology, surgical copilots | Onco-Agent [72], SurgBox [294] |
| Medication Management Longitudinal Care | Remote monitoring | wearable monitoring, outcome forecasting | REMONI [105], CLIMATv2 [196] |
| | Prescription checks | indication, dosage, interaction checks | RxStrategist [271] |
| Data, Documentation, and Knowledge Infrastructure | Clinical documentation | drafting, reflexion-style letters, error detection | RadCouncil [331], PatientLetter [254], MedReAct/MEDREFLEX [49] |
| | Coding and query programming | ICD/PCS coding, natural-language-to-SQL | EHRAgent [243], Multi-role ICD [151] |
| | Knowledge graphs | ADE graph, KG querying | MALADE [47], AGENTiGraph [346] |
| | RAG for medical QA | KG + RAG for QA | AMG-RAG [224] |
| Education, Simulation, and Learning Health Systems | Case-based training | multi-agent, VR anatomy | MEDCO [291], VR-Anatomy [46] |
| | Synthetic patients | self-play, co-evolving behaviors | AgentHospital [150], EvoPatient [56], AIPatient [326] |
| | Trial prediction | trial outcome, study design | ClinicalAgent [330] |
| Regulation and Administration | Regulatory and payer workflows | device regulation, prior authorization | RegMan-Agents [94], MAS-PA [209] |
| | Administrative automation | registration, scheduling, documentation | Admin-Auto [83] |

Table 3: Overview of representative medical agent applications.

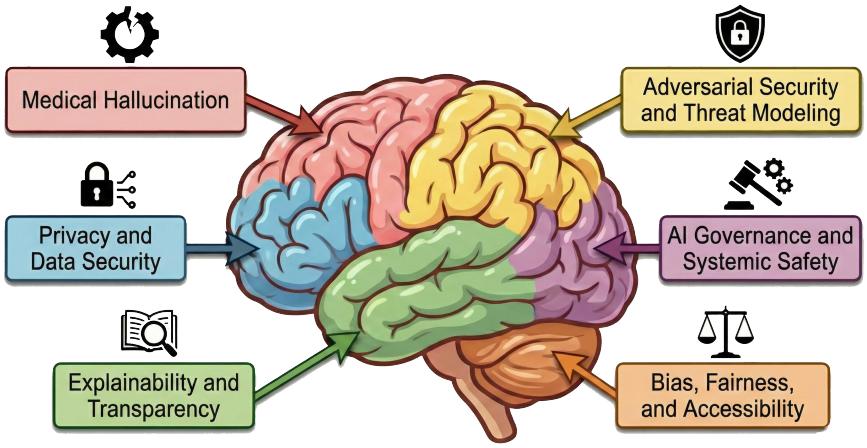


Figure 6: The Structure of Ensuring Safety in Medical AI.

important to encode clinical guidelines, hospital policies, and legal rules in machine-readable form, so that agents can check them automatically. As logs from real use and simulations grow, these applications could form learning health systems that improve step by step under human oversight.

## 7 Safety of Medical Agents: Are We Ready to Trust AI with Patient Lives?

Building on the capability, system, task, departmental, and workflow views in Sections 2–6, we now turn to the question of safety: how medical agents can fail, and what safeguards are needed to prevent patient harm and systemic risk. This section reviews recent work focusing on the safety of medical agents and analyzes safety challenges that span all previous settings, including Medical Hallucination, Privacy Security, Explainability of the system, Adversarial Security , AI Governance and Fairness among the patients.

### 7.1 Medical Hallucination: Reasoning Failure and Clinical Misjudgment

**Hallucinations as a Barrier to Safe Clinical Applications.** Hallucination in LLM-based medical agents refers to model-generated content that, although clinically plausible on the surface, is factually incorrect, internally inconsistent, or not supported by authoritative clinical evidence in ways that can distort diagnosis or treatment decisions [134]. The safety implications of such failures manifest hierarchically throughout the healthcare ecosystem. In the patient-facing intersection, hallucinations in LLM-based health-advice agents can mislead users about the severity of symptoms, screening options, and use of medications, prompting them to follow confident but incorrect recommendations that may delay appropriate care and thus cause preventable harm [5]. This risk is particularly acute in oncology, where models may fabricate unapproved combination therapies or overlook life-threatening drug interactions, directly endangering patient safety [199]. Beyond direct interaction, hallucinations can affect clinician decision-making and compromise the reliability of clinical documentation; automated documentation tools can introduce baseless diagnoses that corrupt the longitudinal veracity of medical records and propagate errors downstream [15]. Similar risks occurred in medical multi-agent architectures, where empirical audits show that shared model deficiencies can drive the group toward a flawed consensus while correct minority opinions are systematically suppressed [90]. Given these negative effects across all dimensions, the rigorous solution against hallucination has transitioned from a technical optimization to an urgent ethical and regulatory prerequisite [203].

**Integrative Strategies.** Current researchers pursue multiple strategies to reduce hallucinations in medical agent systems, which can be categorized by their point of intervention, from evidence-based grounding to structured reasoning and evaluated supervision. A first step in hallucination mitigation is to systematically ground model outputs in verifiable, preferably primary, clinical evidence. In practice, retrieval-augmented generation (RAG) pipelines realize this principle by constraining generation to curated biomedical and

public health knowledge sources. Empirical evaluations show that such RAG-based architectures achieve large reductions in hallucination rates in biomedical question-answering systems designed for public health education [306]. Beyond input grounding, recent work increasingly focuses on approaches that control stochasticity and enforce coherence in intermediate reasoning states. Multimodal agents like MedMMV [165] further suppress hallucinations by constraining model stochasticity, grounding reasoning in structured evidence graphs, and deploying dedicated hallucination detectors that actively supervise the validity of intermediate reasoning states across modalities. Other systems adopt self-refinement frameworks such as Re-KGR [200], which iteratively revise draft answers based on estimated hallucination risk, leading to more stable and coherent medical reasoning. In addition, many scholars argue that mitigating deployment risk requires sufficient transparency to enable meaningful human oversight; governance-oriented work therefore emphasizes the need for explicit, multi-dimensional reporting of model performance and safety-related failure modes in healthcare conversations [187]. Complementing this governance perspective, structured evaluation benchmarks such as MedHallu provide supervised datasets for detecting unsupported or hallucinated answers and show that introducing abstention options substantially improves precision and F1 scores by allowing models to abstain on low-confidence cases [210].

### 7.2 Privacy and Data Security

**Privacy Risks across the Application Lifecycle.** Privacy leakage in medical agent systems can occur at multiple stages of the development and deployment pipeline. During data collection and training, sensitive patient information contained in electronic health records (EHRs) can be memorized by large language models and later exposed in generated text or extracted via targeted data-extraction [30; 251]. In the deployment phase, cloud-based inference and API access further expose confidential data when raw clinical notes or imagings are transmitted to external model providers [201; 82]. At the interaction layer, empirical work shows that the increasing autonomy of LLM-powered agents creates substantially higher privacy risks than static prompting. PrivacyLens-Live [280] revealed markedly higher rates of contextual-integrity violations and sensitive-data leakage in realistic multi-step workflows, even when the underlying models score well on static privacy Q&A benchmarks. These vulnerabilities highlight that privacy assurance for LLM-based medical agents cannot only rely on policy compliance alone but also be embedded into the underlying system architecture through dedicated technical safeguards [343; 237].

**International Standards and Compliance Requirements.** Globally recognized frameworks such as the *General Data Protection Regulation* and the *Health Insurance Portability and Accountability Act* define strict rules for collecting, processing, and storing medical data, and are increasingly treated as the primary legal reference points for the governance of AI systems in clinical settings [310]. Additional regional acts, including *PIPEDA* and *PHIPA* in Canada, provide similar guarantees for accountability, auditability, and patient control over health information [31]. Collectively, these regulations support a privacy-by-design paradigm, in which adequate protection of health data in the digital age is seen as inconceivable without embedding transparency, traceability, and consent-aware data usage throughout the AI system lifecycle [26].

**Integrative Strategies.** To protect patient privacy in practical medical-agent deployments, researchers increasingly advocate for multi-layered technical and organizational safeguards that span data handling, model architecture, and clinical deployment. At the data level, the *Patient-Zero* framework [140] replaces real electronic health records with medically aligned synthetic records and realistic patient–agent interactions, enabling model development and evaluation without direct exposure of Protected Health Information (PHI). Complementary approaches such as *differential privacy* introduce calibrated noise to protect individual contributions while retaining statistical utility in medical learning systems [189]. Architecturally, the *Agentic-AI Healthcare* platform [237] demonstrates how a dedicated Privacy & Compliance Layer can be integrated into a multi-agent system in which symptom-checking, medication-support, and appointment agents are compartmentalized to reduce error propagation and minimize single points of failure. In terms of deployment, *local-first* and *edge-based* designs [201; 82] keep computation within hospital infrastructures and secure on-device environments, so that patient data remain within controlled boundaries.

### 7.3 Explainability and Transparency

**The role of Explainability in Safety-critical Medical AI.** Explainability and interpretability are fundamental to trustworthy use of medical AI systems, because clinicians must be able to understand how recommendations are produced before integrating them into diagnosis, treatment, or documentation workflows [108]. Empirical studies on medical LLM agents show that early diagnostic and decision-support agents often behave as black-box systems whose reasoning processes are misaligned with clinicians' cognitive workflows, leading to user distrust and limiting clinical adoption [33; 264]. For LLM-based medical agents more broadly, recent agent-centric frameworks emphasized that transparency, traceable rationales, and auditable decision paths are prerequisites for safe deployment, shared decision-making, and regulatory accountability in high-stakes clinical use [288; 84].

**Integrative Strategies.** To address these concerns, recent medical-agent frameworks promote explicit reasoning traces, such as tree-structured chain-of-thought and argumentation-based rationales, so that LLM-based diagnostic agents expose intermediate steps and the supporting clinical evidence in a form that clinicians can compare against established diagnostic and therapeutic pathways [213; 264]. Complementary retrieval-augmented medical agents ground their answers in curated guidelines, textbooks, and patient records, and use groundedness metrics to maintain a traceable link between each response and its supporting evidence [48]. At the multi-agent level, *MedAgentAudit* shows that even high-accuracy diagnostic frameworks can exhibit unsafe collaboration patterns, such as flawed consensus, suppression of correct minority views, and loss of key evidential units, when their internal discussions remain opaque, and argues for audit trails that log prompts, intermediate messages, and evidence flow as a prerequisite for trustworthy multi-agent medical systems [90]. Explainability is increasingly extended beyond internal model behaviour to the transparency of deployed medical-AI products. An empirical audit of CE-certified radiology AI tools in Europe finds that public documentation often lacks basic information on training data, validation cohorts, consent procedures, and deployment caveats, making it difficult for clinicians and institutions to assess safety and risk [67]. Based on these findings, governance and regulation work argues that transparency artifacts, such as model cards, subgroup performance reports, and accessible descriptions of reasoning and limitations, should be treated as core safety enablers rather than optional marketing material [187]. Taken together, these developments, from interpretable reasoning within single agents, to process-level auditing of multi-agent collaboration, to institutional transparency about product behaviour, support the view that explainability is not merely a desirable feature, but a foundational component of medical-agent safety and clinical trustworthiness.

### 7.4 Adversarial Security and Threat Modeling

**Types of Adversarial Attacks in Medical Agents.** Medical agents operating in clinical workflows face adversarial threats from both external malicious actors and internally compromised components. Externally, LLM-based agents are vulnerable to targeted jailbreaking, where adversarial prompts bypass safety alignment to force the generation of harmful treatment planning or contraindicated prescriptions [337; 319]. These risks are amplified by indirect prompt-injection attacks, where medical agents retrieve adversarially poisoned web content and thus hidden instructions can manipulate their tool usage, forcing them to email full conversation histories containing sensitive medical information or return malicious URLs that compromise the user's system [218]. Internally, the architecture of Multi-Agent Systems (MAS) can be attacked by communication manipulation. Research reveals that "agent-in-the-middle" attacks can compromise inter-agent dialogue, where a single infiltrated or faulty agent propagates subtle errors that contaminate the collective consensus, suppressing valid clinical signals from other agents [99; 127].

**Integrative Strategies.** To enhance system robustness against these threats, current researchers explore parallel defense strategies. For external prompt injection, pipeline-level cleaning of tool outputs and input filtering, such as schema validation, domain-restricted browsing, and policy-based rejection of unsafe instructions, can limit exposure to malicious content [218]. For internal MAS corruption, threat-aware multi-agent benchmarks propose resilience testing frameworks and topology-specific defense mechanisms, including redundancy, role separation, anomaly-based monitoring, and cross-agent verification [113]. Broader security analyses emphasize the integration of adversarial threat modeling and governance practices into the

deployment pipeline, including continuous auditing, provenance tracking, and policy-enforced least privilege [4]. Collectively, these approaches mark a shift toward security-centered medical agent design, aiming to ensure that clinical AI systems remain robust, trustworthy, and safe even under adversarial pressure.

### 7.5 AI Governance and Systemic Safety

Ensuring medical agents are trustworthy for clinical use requires governance structures that control how decisions are generated and monitored. Recent work emphasizes two complementary perspectives: architectural governance and autonomy-based governance. From the architectural side, medical multi-agent systems (MAS) such as Tiered Agentic Oversight demonstrate how a supervised hierarchical framework can be leveraged to guarantee clinical safety through the rigorous cross-verification of diagnostic results. By organizing agents into distinct tiers with explicit escalation protocols, this architecture ensures that diagnostic outputs are subjected to layered scrutiny, routing high-risk cases to human clinicians while maintaining continuous oversight over autonomous decisions [135]. From the autonomy perspective, AI-based clinical decision systems are classified by their level of autonomy, providing corresponding safety assurance from advisory decision support to systems that directly intervene on patients, and higher autonomy levels are associated with stronger safety-assurance obligations, including more demanding hazard analysis, run-time monitoring, and other structured evidence for safe use [73]. These recent insights treat governance as a system-level safety mechanism, aligning agent architecture and autonomy with the depth of oversight needed to protect patients and preserve clinical accountability.

### 7.6 Bias, Fairness, and Accessibility

Bias in medical-agent systems comes both from the data and models they use and from the way agents respond to patients: conversational and multi-agent clinical systems can inherit demographic and sampling biases from training data, and then express them as systematically different responses, plans, or resource allocations across race, gender, age, or socioeconomic groups [194; 160]. In practice, prompts, dialogue flows, retrieval components, and interaction patterns may encode assumptions centered on majority groups, leading agents to neglect the needs of marginalized populations [180]. Such bias needs to be addressed because fairness is a foundational requirement for medical-agent systems: algorithmic bias can exacerbate existing health disparities, misallocate scarce clinical resources, and erode trust in agent-assisted decision-making [282]. To make fairness operational, researchers adapt fairness metrics such as demographic parity, equalized odds, and counterfactual tests in healthcare prediction and evaluation frameworks and begin to extend them to agent outputs such as dialogues, triage decisions, or treatment recommendations [282; 1], while evaluation frameworks that vary patient attributes across subgroups show that medical LLM–based agents can shift their recommendations in ways that reflect underlying social biases, motivating explicit counterfactual-fairness evaluations for agentic systems [214]. Beyond measurement, agent-centric debiasing methods intervene within the agent pipeline, for example fairness-aware critic or bias-detector modules that monitor and re-rank retrieved evidence before reasoning in knowledge-augmented or multi-agent architectures [248], combined with inclusive data coverage and deployment choices such as multilingual agents that can be deployed in diverse care settings, including low-resource environments, so that historically disadvantaged populations see genuine improvements in access and outcomes rather than further marginalization [3; 180]. Importantly, tool-using agentic pipelines introduce novel pathways for demographic bias that extend beyond standalone model biases. DUCX introduces a stage-wise fairness decomposition for chest X-ray agents that separates end-to-end bias from three agent-specific sources—tool exposure bias (utility gaps conditioned on tool presence), tool transition bias (subgroup differences in tool-routing patterns), and model reasoning bias (subgroup differences in synthesis behaviours)—revealing that intermediate behaviours exhibit distinct subgroup disparities not predictable from end-to-end evaluation alone, with equalized odds gaps reaching up to 20.79% and conditioned subgroup utility gaps as high as 50% [309]. These findings underscore that fairness auditing for medical agents must move beyond aggregate performance to process-level analysis of tool use, routing, and reasoning across demographic subgroups.

# 8 How Should We Evaluate Medical Agents in Practice?

Safety considerations must be grounded in robust evaluation. In this section, we will discuss how we can reliably measure agents' performance and risks.

## 8.1 Benchmarks

We propose a comprehensive taxonomy that integrates both classic datasets and the latest research. To fully evaluate medical agents, we organize benchmarks into five distinct categories: Static Medical Tasks, Sequential Clinical Simulation, Safety and Robustness, Automated Clinical Evaluation, and System-Level Benchmarks. This structure provides a complete view of agent performance, ranging from basic knowledge to complex clinical interactions.

**Static medical tasks.** Static medical task benchmarks evaluate LLM or LLM-based agents on general medical knowledge in fixed input–output formats, without interaction with a simulated environment or the longitudinal state of the patient [261]. Representative knowledge-QA datasets include MedMCQA, PubMedQA, and the medical subsets of MMLU and MMLU-Pro [205; 122; 104; 352]. Static multimodal benchmarks such as VQA-RAD and PathVQA extend this setting to image–text pairs, assessing whether models can answer diagnosis- or finding-related questions from radiology or pathology images while still preserving a one-shot question–answer format [141; 101]. Med-CMR [87] moves beyond general medical VQA by specifically targeting complex diagnostic reasoning, decomposing multimodal QA into fine-grained visual understanding and multi-step, clinically grounded question types. In addition, some datasets focus on isolated medical tasks such as patient–doctor dialogue, MedDialog and MedDG, where multi-turn conversations are evaluated as self-contained encounters without explicit modelling of downstream orders and follow-up management decisions [38; 171]. Recent static benchmarks such as MedXpertQA push this regime towards expert-level difficulty by aggregating specialty board-style questions and multimodal clinical cases across 17 specialties and 11 body systems, while still evaluating models in a single-turn QA format without explicit environment interaction [364].

**Sequential clinical simulation.** Sequential clinical simulation benchmarks mirror real-world clinical workflows by embedding LLM-based agents in interactive environments where they iteratively gather information, update hypotheses, and make multi-stage diagnostic and treatment decisions [232; 63; 311; 261]. MedChain, for example, organizes 12,163 real clinical cases from 19 specialties into multi-stage trajectories spanning referral, history, examination, investigations, diagnosis, and management, requiring agents to choose actions at each step rather than answer a single static question [166]. AgentClinic converts medical QA problems into simulated encounters in a virtual clinic, supporting both multimodal analysis and dialogue-based scenarios in which agents must conduct consultations, order tests, and refine decisions over several turns [232]. Benchmarks such as AI Hospital's MVME environment, ClinicalLab's ClinicalBench, and related multi-agent hospital simulations further emphasize end-to-end pathways across departments, modelling interactions between doctor, patient, staff, and tools to evaluate whether medical agents can sustain coherent reasoning over time and adapt to evolving patient states [63; 311].

**Safety and robustness benchmarks.** Safety-oriented benchmarks test medical agents for hallucination, robustness, and adversarial behaviours in pratical settings. MedFact defines a fact-checking benchmark where model-generated clinical claims must be verified against curated biomedical evidence, explicitly quantifying factual consistency and hallucination rates [39]. AMQA builds adversarial medical QA scenarios by systematically perturbing questions and options, thereby probing whether models maintain correct diagnoses under distribution shifts and deceptive distractors [303]. Dynamic Automatic and Systematic (DAS) red-teaming proposes a multi-agent framework that automatically generates targeted adversarial prompts and safety tests for medical LLMs, enabling scalable discovery of unsafe behaviours across clinical tasks [207]. MedRepBench focuses on radiology reporting and introduces benchmarks where hallucinated findings and unsupported impressions are explicitly annotated and measured, linking generation quality to patient safety risks [235]. These datasets complement knowledge benchmarks by focusing evaluation on factual reliability, robustness to adversarial inputs, and risk-sensitive error modes that are critical for deploying medical agents in

practice. LiveMedBench addresses the critical limitations of data contamination and temporal misalignment by continuously harvesting real-world clinical cases from online medical communities on a weekly basis, ensuring strict temporal separation from model training data. It employs a Multi-Agent Clinical Curation Framework for data validation and an Automated Rubric-based Evaluation Framework that decomposes physician responses into granular, case-specific criteria, revealing that even the best-performing model achieves only 39.2% and that 84% of models exhibit performance degradation on post-cutoff cases [313].

Table 4: Representative benchmarks for LLM-based medical agents, grouped by evaluation focus.

| **Benchmark** | **Key focus** | **Resource link** |
|---|---|---|
| | *Static medical tasks* | |
| MedMCQA [205] | Large-scale multi-subject medical MCQ dataset | GitHub |
| PubMedQA [122] | QA over PubMed biomedical abstracts | GitHub |
| MMLU [104] | Multi-task exam benchmark with medical-related subdomains | GitHub |
| VQA-RAD [141] | Single-turn radiology image–text VQA | GitHub |
| Med-CMR [87] | Complex visual understanding and multi-step clinical reasoning spanning 12 imaging modalities | GitHub |
| | *Sequential clinical simulation.* | |
| MedChain [166] | Multi-stage interactive clinical cases across 19 specialties | GitHub |
| AgentClinic [232] | Multimodal virtual clinic with dialogue and tool interaction | GitHub |
| ClinicalLab [311] | Clinical prediction and diagnosis across 24 departments | GitHub |
| 3MDBench [259] | Multimodal multi-agent telemedicine consultations | GitHub |
| | *Safety and robustness benchmarks* | |
| MedFact [39] | Evidence-based fact-checking of LLM-generated medical claims | GitHub |
| AMQA [303] | Adversarial medical QA for bias and fairness assessment | GitHub |
| | *Automated clinical evaluation* | |
| MedQA-CS [322] | OSCE-style multi-station benchmark for clinical skills, using MedExamLLM as an LLM-as-judge for structured scoring of multi-section encounters | HuggingFace |
| 3MDBench [259] | Medical multimodal multi-agent dialogue benchmark with automated scoring of diagnosis and doctor–patient communication | GitHub |
| | *System-level and training-focused benchmarks* | |
| MedAgentBoard [360] | System-level comparison of multi-agent, single-LLM and conventional pipelines across QA, summarization, EHR prediction and workflow automation tasks | GitHub |
| MedAgentAudit [90] | Diagnostic framework and benchmark logs for analyzing collaborative failure modes inside medical multi-agent systems at the process level | GitHub |
| MedResearcher-R1 [324] | Agentic training and evaluation framework for deep medical research agents | GitHub |

**Automated clinical evaluation.** A growing line of work builds benchmarks that explicitly reduce dependence on clinician raters by automating large parts of the evaluation pipeline for medical agents [186]. The AI-SCE concept proposes Artificial Intelligence Structured Clinical Examinations, inspired by OSCEs in medical education, to assess agents on realistic clinical scenarios and structured checklists of clinical competencies [186]. 3MDBench incorporates a dedicated assessor agent that automatically scores dialogue quality and task completion in multi-agent telemedicine consultations, further illustrating how LLM-as-judge designs can scale evaluation to thousands of cases [259]. MedQA-CS instantiates an AI-SCE benchmark where LLMs are evaluated on multi-station clinical skills, using an examiner LLM (MedExamLLM) as an LLM-as-judge to grade multi-turn interactions with standardized vignettes [322].

MedRepBench similarly uses LLM-based judges calibrated against expert annotations to rate radiology reports along axes such as clinical correctness and completeness, substantially reducing manual grading effort [235]. These automated frameworks are particularly relevant for medical agents, where rich multi-turn trajectories make purely human evaluation expensive and difficult to reproduce at scale.

**System-level and training-focused benchmarks.** Finally, several benchmarks focus on system-level and training-oriented aspects of medical agents rather than a single downstream task. FedAgentBench studies federated training and evaluation of LLM-based agents across multiple clients, providing tasks, datasets, and metrics to analyze how agent performance, communication cost, and privacy constraints interact in decentralized clinical settings [228]. MedBrowseComp designs tasks where agents must conduct deep medical web research, thus benchmarking the entire retrieval–reasoning–reporting pipeline rather than isolated responses [37]. MedAgentBoard, beyond its task-level conclusions, also exemplifies a system-centric design by jointly reporting accuracy, efficiency, and workflow automation metrics for multi-agent, single-LLM, and conventional systems on a shared platform [360].
Collectively, these benchmarks highlight that evaluating medical agents requires not only task-level correctness but also careful measurement of training paradigms, resource usage, and integration with broader clinical information systems.

### 8.2 Metrics

Metrics are essential for translating the theoretical capabilities of medical agents into verifiable clinical value. We outline six primary categories of evaluation based on specific operational needs, with a complete summary of specific indicators and related works available in Table 5.

**Task Performance Metrics** are results-oriented and used for how good the final answer is in one medical task. In QA tasks, metrics such as accuracy, precision, and F1 score are calculated by directly comparing the model outputs with standard labels. Benchmarks such as MedQA[121] and MedMCQA [205] often rely on these exact match metrics. For text generation tasks like report summarization, Semantic Similarity assess how well the meaning of the generated text matches that of the reference text. Metrics such as ROUGE which evaluates summarization quality, and BERTScore which uses contextual embeddings to capture semantic relationships, have been applied to ensure linguistic alignment[323].

**Reasoning and Process Metrics** go beyond the final answer to assess the logical validity of the agent's cognitive path. These metrics evaluate whether the intermediate steps, such as chain-of-thought (CoT) derivations, are factually correct and logically consistent. Metrics such as Step-wise Factuality and Reasoning Gap [308] measure the fidelity of the inference process. While exact match metrics assess *what* the agent answers, reasoning metrics allow researchers to verify how the conclusion was reached, which is critical for clinical trustworthiness in complex diagnostic scenarios[145].

**Safety Metrics** are deployed to quantify the potential risks and adverse outcomes associated with medical agents. These metrics assess the frequency and severity of errors, such as Hallucination Rate [15] and Omission Rate [22]. Frameworks like MedGuard [320] utilize Attack Success Rate (ASR) and Toxicity Score to evaluate robustness against adversarial inputs.

**Agent Interaction Metrics** assess the agent's competence in executing multi-step workflows, such as retrieving patient records via APIs or scheduling appointments. Metrics such as Tool Selection Accuracy [89] are calculated by verifying the correctness of the agent's action sequence. Benchmarks such as MedAgentBench [98] apply Task Success Rate to evaluate the executable capability of agents in simulated hospital settings, ensuring they can operate effectively within digital health infrastructures.

Table 5: A summary of evaluation metrics for medical agents, categorized by dimension and specific indicators.

| **Key Metrics** | **Genre** | **Type** | **Reference** |
|---|---|---|---|
| *Task Performance Metrics* | | | |
| Accuracy, F1 Score | Static QA | Correctness | [191] |
| ROUGE-L, BERTScore | Generation | Semantic | [323] |
| Clinical Relevance | Clinical Decisions | Concordance | [266] |
| AUC-ROC, Precision, Recall | Diagnostic | Classification | [288] |
| *Reasoning and Process Metrics* | | | |
| Reasoning-step Factual Accuracy | Logic Fidelity | Factuality | [120] |
| Chain Completeness Metrics | Inference | Process | [33] |
| Self-Correction Rate, Debugging Success | Refinement | Self-Correction | [340] |
| Diagnostic Logicality etc. | clinical capabilities | Process | [168] |
| *Safety Metrics* | | | |
| Hallucination Rate, Omission Rate | Hallucination | Risk | [15] |
| Clinical Harm Score | Harm Assessment | Severity | [22] |
| Attack Success Rate | Robustness | Adversarial | [320] |
| Toxicity Score, Refusal Rate | Content Safety | Toxicity | [320] |
| Counterfactual Patient Variation Score | Fairness | Bias | [21] |
| *Agent Interaction Metrics* | | | |
| Tool Selection Accuracy, API Pass Rate | Tool Use | Competence | [89] |
| Task Success Rate | Execution | Success | [98] |
| Interaction Efficiency | Workflow | Efficiency | [232] |
| Agent-behavior metrics etc. | Workflow | Correctness | [63] |
| *Human-Centered Metrics* | | | |
| Clinical Readiness, Bedside Manner | Empathy | Assessment | [193] |
| Follow-up Consultation Willingness | Usability | Assessment | [232] |
| System Usability Scale | Usability | Assessment | [144] |
| Empathy Rating, Satisfaction Score | Empathy | Assessment | [16] |
| *LLM-as-a-Judge* | | | |
| Truthfulness, Informativeness | Scoring Agent | Assessment | [165] |
| Factuality Score, Interpretability | Scoring Agent | Assessment | [235] |
| Agent-Judge Agreement | Consistency | Agreement | [339] |

**Human-Centered Metrics** evaluate the qualitative quality of the interaction between the medical agent and human user, shifting from the model to real clinical practice. These metrics assess subjective factors including trust, usability, and empathy. Representative human-centered evaluations already appear in recent medical-agent studies. Polaris, for example, a safety-focused multi-agent architecture for nurse-style virtual care, recruits over a thousand licensed nurses and more than a hundred physicians to rate end-to-end

conversations on medical safety, clinical readiness, patient education, conversational quality, and bedside manner, demonstrating that agent performance can be benchmarked directly against human nurses along trust and empathy dimensions [193].

**LLM-as-a-Judge** employ advanced language models themselves to act as evaluators for complex, open-ended clinical tasks where static references are insufficient. This approach leverages strong models to score outputs based on nuanced factors. For example, MedMMV rely on LLM-as-a-judge panels to score truthfulness, informativeness, and cross-modal consistency of each reasoning trajectory, aggregating model ratings into reliability scores and cross-modal hallucination rates rather than only checking final-answer accuracy[165]. MedRepBench [235] utilizes LLMs to grade the factuality and reasoning quality of medical reports. Chen et al. propose a retrieval-augmented multi-agent framework that automates the generation of instance-specific evaluation rubrics by decomposing retrieved authoritative medical evidence into atomic facts and synthesising them with user interaction constraints, achieving a Clinical Intent Alignment score of 60.12% on HealthBench and nearly doubling the quality separation of GPT-4o baselines [43]. This method provides a scalable and adaptable solution for aligning agent behaviors with expert-level clinical judgment without the need for extensive human annotation.

### 8.3 Challenge and Discussion

**From benchmark gains to clinical reality: the sim-to-real gap.** Despite this rapidly expanding ecosystem, recent works on medical-agent benchmarks expose several structural limitations in how we probe real clinical utility. Environment-centric suites such as MedAgentBench and AgentClinic move beyond static QA by embedding agents in virtual EHRs and simulated clinics, with FHIR-compliant APIs and multi-step tool calls; AgentClinic additionally introduces patient-centric measures such as time-to-diagnosis and dialogue burden [98; 232]. Yet these environments remain constrained testbeds: they typically cover a single institution's data model, operate on de-identified or synthetic records, and focus on short-horizon, scripted encounters, rather than longitudinal care pathways, cross-department coordination, or safety-critical event rates in live settings [98]. Complementary efforts such as MedAgentsBench extend this paradigm toward harder cases and multi-step clinical trajectories, but still rely on repackaged datasets and pre-defined interaction graphs that cannot fully capture the heterogeneity of documentation styles, institutional policies, and patient populations across real hospitals [261]. System-level platforms like MedAgentBoard broaden task coverage and compare multi-agent systems against single-LLM and conventional baselines, revealing that multi-agent collaboration does not consistently outperform specialized pipelines on VQA or EHR prediction and that benchmark gains are highly task-dependent [360]. At the same time, early real-world studies such as EHR-MCP show that even when agents are wired into live hospital EHRs via tool protocols, evaluations still concentrate on narrow subtasks and retrospective cohorts, leaving open questions about prospective impact and robustness under temporal and distributional shift [183].

**Limitations of current metrics: from proxy scores to clinical validity.** Recent work on evaluating medical LLMs and agents suggests that the core problem is not just *which* metrics we report, but *what claims* those metrics credibly support in clinical practice. Position papers on construct validity argue that many medical LLM evaluations still rely on narrow proxies such as exam-style accuracy or task success rate, even when they are used to justify broad claims about clinical reasoning or readiness for deployment [9; 229]. For example, MedAgentBench reports a single task success rate as its primary outcome, with manually curated reference solutions and rule-based sanity checks [98]; such scalar scores capture whether an agent completes a scripted workflow, but they under-specify how severe failures are, whether near-misses are clinically acceptable, and how performance generalizes across institutions or patient subgroups. Complementary efforts like 3MDBench and AI-SCE–style frameworks move toward richer, multi-dimensional scoring, combining diagnostic accuracy with communication quality and structured checklists of clinical competencies [259; 186; 322],yet these instruments are still validated primarily against short episodes of interaction, rather than longitudinal outcomes such as complication rates, readmissions, or sustained workload reduction. At the same time, other work leverages an LLM-as-a-judge to rapidly benchmark the diagnostic ability of dozens of models on patients, dramatically reducing the human labeling burden [231]. Beyond accuracy, recent work highlights that *stochasticity* poses a critical and often overlooked barrier to real-world deployment of agentic systems. Zhai et

al. formalise the study of stochasticity in Deep Research Agents by modelling them as information-acquisition Markov Decision Processes and identifying three sources of variance—information acquisition, information compression, and inference—showing that reducing stochasticity, particularly during inference and early-stage decisions, improves research output quality and that structured output and ensemble-based query generation can reduce average stochasticity by 22% [332]. However, meta-analyses and general AI-evaluation studies caution that judge models introduce their own biases and preference leakage, and that metric validity depends on how well these automated scores track independent, domain-specific standards rather than other LLMs' opinions [229; 6]. Psychometrics-inspired frameworks for generalist medical AI further highlight that many existing scores lack clear construct validity, predictive validity, and external validity: they do not systematically relate evaluation performance to underlying clinical skills, downstream patient outcomes, or robustness in new hospitals and populations [256; 9].

## 9 Open Challenges and Future Directions

The preceding sections have surveyed the current landscape of Medical Agents from multiple angles: their core capabilities (Section 2), multi-agent system designs (Section 3), task formulations (Section 4), departmental practices (Section 5), workflow-level applications (Section 6), and the safety and evaluation frameworks(Section 7 and Section 8). While these advances demonstrate the feasibility and promise of agentic approaches in clinical settings, they also expose structural limitations and open questions. This section synthesises these observations into a small number of forward-looking themes.

### 9.1 Advancing Capabilities Under Clinical Constraints

Section 2 decomposed Medical Agents into core modules for perception, cognition, memory, planning, and tool use. Extending these capabilities in healthcare requires balancing ambition with domain-specific constraints.

First, **clinical reasoning and uncertainty handling** remain under-specified in most current systems. Many Medical Agents rely on generic chain-of-thought prompting or heuristic planning, without explicit representations of differential diagnoses, competing hypotheses, or uncertainty over patient state. Future work should explore reasoning frameworks that:

- **Structured hypothesis spaces:** represent diagnostic and management options as structured hypothesis spaces rather than flat lists;
- **Uncertainty quantification and propagation:** quantify and propagate uncertainty arising from missing data, conflicting evidence, and model limitations; and
- **Guideline- and preference-constrained planning:** integrate local guidelines, institutional policies, and patient preferences as first-class constraints on planning.

Second, **memory architectures must match the structure of care trajectories**. As discussed in Section 2, Medical Agents benefit from distinct working, episodic, semantic, and procedural memories. In clinical environments, these memories must be tuned to the granularity and timescales of care:

- **Encounter-level working memories:** encounter-level working memories that track the immediate clinical context and active decisions;
- **Admission- or episode-level episodic memories:** admission- or episode-level episodic memories that record key events, interventions, and responses;
- **Longitudinal memories:** longitudinal memories that span chronic disease trajectories across settings; and
- **Procedural memories:** procedural memories that encode reusable care pathways and operational protocols, with mechanisms for versioning and governance.

Designing such memories raises open questions about representation, retrieval, retention policies, and their interaction with privacy and consent.

Third, **tool orchestration and learning must be clinically governed**. Sections 5 and 3 highlighted the diversity of tools that Medical Agents must coordinate: EHR modules, laboratory and imaging systems, triage platforms, simulation environments, research infrastructure, and more. Future work should:

- **Declarative workflow specifications:** move from ad hoc tool invocation to declarative specifications of workflows, including preconditions, invariants, and safety guards;
- **Verification and simulation of policies:** develop verification and simulation methods that can stress-test orchestration policies before deployment; and
- **Governed learning pipelines:** couple tool policies to *governed* learning pipelines that separate experimentation from production, support shadow deployment and staged rollouts, and allow rapid rollback when adverse patterns are detected.

Finally, **data-efficient and continual learning** must be reconciled with safety and regulatory requirements. Leveraging interaction logs, simulation traces, and operational signals to improve agents is attractive but risky. Open problems include:

- **Weak and indirect supervision:** combining weak and indirect supervision (e.g., order sets, quality metrics, audit comments) with explicit annotations to shape agent behaviour;
- **Off-policy evaluation for safety:** designing off-policy evaluation methods that estimate the counterfactual impact of policy changes without compromising patients; and
- **Privacy-preserving and federated training:** developing privacy-preserving and federated training regimes that allow cross-institutional learning without sharing raw data. Recent work demonstrates that shifting adaptation from weight space to memory space can enable federated learning with substantially reduced communication overhead [32], and that training-free frameworks can distil simulated clinical trajectories into entity-agnostic operational rules for privacy-preserving local deployment [317].

### 9.2 Trustworthy, Equitable, and Governed Medical Agents

Section 7 stated that Safety has emerged as a central concern across the Medical Agent literature, cutting across hallucination, robustness, bias, transparency, and accountability. Rather than treating these issues in isolation, future work should develop integrated frameworks that link technical guarantees to institutional governance.

On the methodological side, **comprehensive safety evaluation** requires going beyond narrow benchmarks. There is a need for evaluation suites that:

- **Simulation plus rare-event probing:** combine end-to-end simulations (e.g., multi-agent hospital environments) with targeted probes for rare but catastrophic errors;
- **Multi-agent and multi-tool stress testing:** stress-test interactions between multiple agents and tools, including potential failure cascades and feedback loops; and
- **Adversarial and red-team evaluation:** incorporate adversarial and red-team testing procedures that reflect realistic threat models in healthcare.

**Fairness and subgroup performance** must also be analysed at the level of workflows, not only predictions. For example, Medical Agents may inadvertently introduce systematic differences between patient groups in time-to-intervention, diagnostic thoroughness, or follow-up robustness. Recent work on tool-using chest X-ray agents demonstrates that tool exposure, tool transition, and model reasoning each contribute distinct

subgroup disparities not predictable from end-to-end evaluation alone, motivating process-level fairness decomposition as a first-class design concern [309]. Future research should:

- **Process-aware datasets and logs:** construct datasets and logging schemas that capture process-oriented metrics (e.g., steps taken, delays, escalations) alongside outcomes;
- **Workflow-level fairness definitions:** define fairness notions appropriate for sequential decision processes and multi-agent settings; and
- **Mitigation strategies for disparities:** develop mitigation strategies that can adjust policies or resource allocation in response to detected disparities.

At the human factors level, **hybrid human–agent workflows** require careful interaction design and training. Effective Medical Agents should:

- **Interpretable state and rationale:** expose their state, confidence, and rationale in ways that are interpretable to clinicians and allied health professionals;
- **Configurable autonomy levels:** support graded levels of autonomy from recommendation to semi-automated execution to tightly governed full automation that can be configured per task and context; and
- **Training and trust calibration:** be accompanied by training curricula that help clinicians learn how to supervise, critique, and calibrate trust in agents, avoiding both over-reliance and under-utilisation.

Finally, **multidisciplinary governance structures** will be essential as Medical Agents become embedded in the core infrastructure of care. Open questions include:

- **Governance body design:** how to structure governance bodies that include clinicians, informaticians, ethicists, patients, and operations leaders;
- **Responsibility and liability allocation:** how to align responsibilities and liability among vendors, health systems, and individual practitioners; and
- **Integration into safety and quality management:** how to integrate agent monitoring and incident response into existing clinical safety and quality management processes.

### 9.3 Environments, Deployment, and Evaluation at System Scale

Sections 5 and 3 highlighted a gap between the sophistication of agent architectures and the simplicity of many evaluation environments and deployment patterns. To understand the true impact of Medical Agents, future work must address environments, deployment, and evaluation together.

First, **richer environments and benchmarks** are needed to approximate the complexity of real health systems. Building on existing case-based datasets, virtual clinics, and multi-agent hospital simulators, future environments should:

- **Operational realism in simulators:** model realistic queues, staffing constraints, resource bottlenecks, and cross-departmental handoffs;
- **Longitudinal, cross-setting cohorts:** span longitudinal cohorts across settings (e.g., primary care, inpatient, rehabilitation, home monitoring), supporting questions about continuity of care; and
- **Standardised, interoperable data formats:** adopt standardised, interoperable data formats (e.g., FHIR-based EHR representations, harmonised imaging and annotation schemas) that facilitate reuse and comparison across studies.

Second, **deployment research must move from departmental pilots to system-level integration**. Most current reports describe narrow pilots focused on specific documentation, triage, or decision-support tasks. Future work should examine:

- **Cross-departmental architectures and APIs:** architectures and API conventions that allow agents to coordinate across departments without creating new silos or duplicating logic;
- **Gradual, controlled rollout:** mechanisms for gradual rollout, including feature flags, sandboxed environments, and staged expansions from single units to enterprise-wide deployments; and
- **Operational resilience strategies:** operational resilience strategies that ensure graceful degradation, manual fallback, and recovery from partial failures or outages.

Third, **evaluation must include system-level and organisational metrics**. Beyond traditional accuracy or AUROC measures, studies should report:

- **Operational performance metrics:** throughput, wait times, and length of stay at the service-line or institutional level;
- **Clinician experience and workload:** clinician workload, burnout indicators, and job satisfaction, particularly in relation to administrative burden; and
- **Patient-centred outcomes and equity:** patient-centred outcomes, including safety events, satisfaction, equity across subgroups, and trust in the care system.

Designing rigorous study protocols that can attribute changes in these metrics to Medical Agents, while accounting for confounding organisational changes, remains an open challenge.

### 9.4 Outlook

Medical Agents sit at the intersection of several rapidly evolving fields: agentic AI, MedAI, clinical decision support, health informatics, and organisational science. The research directions outlined above suggest that progress will depend not only on more capable models and algorithms, but also on advances in simulation, evaluation, governance, and socio-technical design.

A central hypothesis emerging from this survey is that the long-term value of Medical Agents will be determined less by isolated performance gains on narrow tasks and more by their ability to reshape clinical workflows in ways that are safe, equitable, and sustainable. Achieving this will require tightly coupled work across the levels we have highlighted: capabilities, departmental practice, task and system design, and the underlying resources that support training and evaluation. As the field matures, a complementary perspective is the systematic "diagnosis and treatment" of AI models themselves: the Model Medicine framework proposes treating AI systems analogously to biological organisms, with anatomy, physiology, and classifiable conditions, bridging the gap between current interpretability research and the systematic clinical practice that complex AI systems increasingly require [119]. Similarly, international expert perspectives on computational pathology highlight that responsible integration of agentic AI into medical practice requires connecting deployable clinical relevance with technical maturity, operational readiness, and economic and regulatory context [50]. If these efforts are successful, Medical Agents may become a core component of learning health systems, enabling continuous improvement in care delivery while respecting the ethical and professional foundations of medicine.

## 10 Conclusion

Medical Agents represent a convergence of two trends: the architectural shift from static, query-bound models to agentic systems capable of perception, cognition, and action; and the clinical imperative to move beyond narrow MedAI point solutions toward workflow-centric automation. Throughout this survey, we have

argued that this convergence is not merely a technical evolution but a structural response to the realities of contemporary healthcare, where multimodal data, temporal dynamics, and organisational constraints make single-shot prediction insufficient.

Conceptually and empirically, we characterise Medical Agents through a functional definition and three-level roadmap that link multimodal, longitudinal, stateful, tool-driven architectures to degrees of workflow integration and autonomy. This lens organises heterogeneous systems across departments, tasks, and environments, and foregrounds three imperatives for the field: integrate fragmented MedAI components into coherent workflows; embed governance that treats safety, fairness, transparency, and accountability as core design constraints; and evaluate agents at system scale rather than only on narrow benchmarks. If these imperatives are met, Medical Agents can move from experimental prototypes to stable infrastructure for learning health systems.

## References


[1] Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *NPJ Digital Medicine*, 7 (1):82, 2024.

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

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

[4] Farhad Abtahi, Fernando Seoane, Iván Pau, and Mario Vega-Barbas. Data poisoning vulnerabilities across healthcare ai architectures: A security threat analysis. *arXiv preprint arXiv:2511.11020*, 2025.

[5] Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. Medhalu: Hallucinations in responses to healthcare queries by large language models. *arXiv preprint arXiv:2409.19492*, 2024.

[6] Monica Agrawal, Irene Y Chen, Freya Gulamali, and Shalmali Joshi. The evaluation illusion of large language models in medicine. *npj Digital Medicine*, 8(1):600, 2025.

[7] Pegah Ahadian, Mingrui Yang, Eva Powlison, Xiaojuan Li, Wei Xu, and Qiang Guan. Oaagent: Multimodal llm agent for predicting knee osteoarthritis progression. In *Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, pp. 144–148, 2025.

[8] Jaerong Ahn, Andrew Wen, Nan Wang, Heling Jia, Zhiyi Yue, Sunyang Fu, and Hongfang Liu. An agentic model context protocol framework for medical concept standardization. *arXiv preprint arXiv:2509.03828*, 2025.

[9] Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Position: Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*, 2025.

[10] Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag. In *European Conference on Information Retrieval*, pp. 201–209. Springer, 2025.

[11] Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Medagentsim: Self-evolving multi-agent simulations for realistic clinical interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 362–372. Springer, 2025.

[12] Mariam ALMutairi and Hyungmin Kim. Resilient multi-agent negotiation for medical supply chains: Integrating llms and blockchain for transparent coordination. *arXiv preprint arXiv:2507.17134*, 2025.

[13] Emma Armstrong et al. Smartstate: An automated research protocol adherence system. In *AMIA Summits on Translational Science Proceedings*. AMIA, 2025.

[14] Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. Tethered to the ehr: primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426, 2017.

[15] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274, 2025.

[16] J. W. Ayers. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6):589–596, 2023.

[17] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[18] David Bani-Harouni, Nassir Navab, and Matthias Keicher. Magda: Multi-agent guideline-driven diagnostic assistance. In *International workshop on foundation models for general medical AI*, pp. 163–172. Springer, 2024.

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

[20] Hadas Ben-Atya, Naama Gavrielov, Zvi Badash, Gili Focht, Ruth Cytter-Kuint, Talar Hagopian, Dan Turner, and Moti Freiman. Agent-based uncertainty awareness improves automated radiology report labeling with an open-source large language model. *arXiv preprint arXiv:2502.01691*, 2025.

[21] Kenza Benkirane, Jackie Kay, and Maria Perez-Ortiz. How can we diagnose and treat bias in large language models for clinical decision-making? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2263–2288, 2025.

[22] Meenesh Bhimani, Alex Miller, Jonathan D Agnew, Markel Sanz Ausin, Mariska Raglow-Defranco, Harpreet Mangat, Michelle Voisard, Maggie Taylor, Sebastian Bierman-Lytle, Vishal Parikh, et al. Real-world evaluation of large language models in healthcare (rwe-llm): a new realm of ai safety & validation. *medRxiv*, pp. 2025–03, 2025.

[23] Bernd Blobel, Pekka Ruotsalainen, Mathias Brochhausen, Frank Oemig, and Gustavo A Uribe. Autonomous systems and artificial intelligence in healthcare transformation to 5p medicine–ethical challenges. *Stud Health Technol Inform*, 270:1089–1093, 2020.

[24] Matthias Blondeel, Noel Codella, Sam Preston, Hao Qiu, Leonardo Schettini, Frank Tuan, Wen-wai Yim, Smitha Saligrama, Mert Öz, Shrey Jain, et al. Healthcare agent orchestrator (hao) for patient summarization in molecular tumor boards. *arXiv preprint arXiv:2509.06602*, 2025.

[25] Mahmoud Brahimi. An edge based multi-agent model for improving hospital bed management. In *2021 International Conference on Information Systems and Advanced Technologies (ICISAT)*, pp. 1–4. IEEE, 2021.

[26] Alissa Brauneck, Louisa Schmalhorst, Mohammad Mahdi Kazemi Majdabadi, Mohammad Bakhtiari, Uwe Völker, Jan Baumbach, Linda Baumbach, and Gabriele Buchholtz. Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: scoping review. *Journal of medical Internet research*, 25:e41588, 2023.

[27] Lang Cao. Diaggpt: An llm-based and multi-agent dialogue system with automatic topic management for flexible task-oriented dialogue. *arXiv preprint arXiv:2308.08043*, 2023.

[28] Ruike Cao, Shaojie Bai, Fugen Yao, Liang Dong, Jian Xu, and Li Xiao. Atpo: Adaptive tree policy optimization for multi-turn medical dialogue. *arXiv preprint arXiv:2603.02216*, 2026.

[29] Luis Cardenas, Katherine Parajes, Ming Zhu, and Shengjie Zhai. Autohealth: Advanced llm-empowered wearable personalized medical butler for parkinson's disease management. In *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0375–0379. IEEE, 2024.

[30] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Ulfar Erlingsson, Alexandru Oprea, and Nicolas Papernot. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

[31] Ann Cavoukian et al. Privacy by design: The 7 foundational principles. *Information and privacy commissioner of Ontario, Canada*, 5(2009):12, 2009.

[32] Bowen Chen, Qiaohui Gao, Shaowen Wan, Shanhui Sun, Wei Liu, Xiang Li, Tianming Liu, and Lin Zhao. Shifting adaptation from weight space to memory space: A memory-augmented agent for medical image segmentation. *arXiv preprint arXiv:2603.05873*, 2026.

[33] Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. Cod, towards an interpretable medical agent using chain of diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14345–14368, 2025.

[34] Kai Chen, Xinfeng Li, Tianpei Yang, Hewei Wang, Wei Dong, and Yang Gao. Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856*, 2025.

[35] Kai Chen, Ji Qi, Jing Huo, Pinzhuo Tian, Fanyu Meng, Xi Yang, and Yang Gao. A self-evolving framework for multi-agent medical consultation based on large language models. In *ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

[36] Kaitao Chen, Mianxin Liu, Daoming Zong, Chaoyue Ding, Shaohao Rui, Yankai Jiang, Mu Zhou, and Xiaosong Wang. Mediator-guided multi-agent collaboration among open-source models for medical decision-making. *arXiv preprint arXiv:2508.05996*, 2025.

[37] Shan Chen, Pedro Moreira, Yuxin Xiao, Sam Schmidgall, Jeremy Warner, Hugo Aerts, Thomas Hartvigsen, Jack Gallifant, and Danielle S. Bitterman. Medbrowsecomp: Benchmarking medical deep research and computer use. *arXiv preprint arXiv:2505.14963*, 2025.

[38] Shu Chen, Zeqian Ju, Xiangyu Dong, et al. Meddialog: A large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*, 2020.

[39] Tong Chen, Zimu Wang, Yiyi Miao, et al. Medfact: A large-scale chinese dataset for evidence-based medical fact-checking of llm responses. *arXiv preprint arXiv:2509.17436*, 2025.

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

[41] Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. *arXiv e-prints*, pp. arXiv–2412, 2024.

[42] Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: Advancing rare disease care through llm-empowered multi-disciplinary team. *arXiv preprint arXiv:2412.12475*, 2024.

[43] Yinzhu Chen, Abdine Maiga, Hossein A Rahmani, and Emine Yilmaz. Automated rubrics for reliable evaluation of medical dialogue systems. *arXiv preprint arXiv:2601.15161*, 2026.

[44] Yixiong Chen, Xinyi Bai, Yue Pan, Zongwei Zhou, and Alan Yuille. Meissa: Multi-modal medical agentic intelligence. *arXiv preprint arXiv:2603.09018*, 2026.

[45] Yuxiao Cheng and Jinli Suo. Openlens ai: Fully autonomous research agent for health infomatics. *arXiv preprint arXiv:2509.14778*, 2025.

[46] Vuthea Chheang, Shayla Sharmin, Rommy Márquez-Hernández, Megha Patel, Danush Rajasekaran, Gavin Caulfield, Behdokht Kiafar, Jicheng Li, Pinar Kullu, and Roghayeh Leila Barmaki. Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments. In *2024 IEEE international conference on artificial intelligence and eXtended and virtual reality (AIxVR)*, pp. 21–30. IEEE, 2024.

[47] Jihye Choi, Nils Palumbo, Prasad Chalasani, Matthew M Engelhard, Somesh Jha, Anivarya Kumar, and David Page. Malade: orchestration of llm-powered agents with retrieval augmented generation for pharmacovigilance. *arXiv preprint arXiv:2408.01869*, 2024.

[48] Mohita Chowdhury, Yajie Vera He, Jared Joselowitz, Aisling Higham, and Ernest Lim. Astrid–an automated and scalable triad for the evaluation of rag-based clinical question answering systems. *arXiv preprint arXiv:2501.08208*, 2025.

[49] Jean-Philippe Corbeil. Iryonlp at mediqa-corr 2024: Tackling the medical error detection & correction task on the shoulders of medical agents. *arXiv preprint arXiv:2404.15488*, 2024.

[50] Qian Da, Yijiang Chen, Min Ju, Zheyi Ji, Albert Zhou, Wenwen Wang, Matthew A Abikenari, Philip Chikontwe, Guillaume Larghero, Bowen Chen, et al. Computational pathology in the era of emerging foundation and agentic ai–international expert perspectives on clinical integration and translational readiness. *arXiv preprint arXiv:2603.05884*, 2026.

[51] Kaizhong Deng, Baoru Huang, and Daniel S Elson. Deep imitation learning for automated drop-in gamma probe manipulation. *arXiv preprint arXiv:2304.14294*, 2023.

[52] Roberta Di Marino, Giovanni Dioguardi, Antonio Romano, Giuseppe Riccio, Mariano Barone, Marco Postiglione, Flora Amato, and Vincenzo Moscato. Solve-med: Specialized orchestration for leading vertical experts across medical specialties. *arXiv preprint arXiv:2511.03542*, 2025.

[53] Julia S Dollis, Iago A Brito, Fernanda B Färber, Pedro SFB Ribeiro, Rafael T Sousa, et al. When avatars have personality: Effects on engagement and communication in immersive medical training. *arXiv preprint arXiv:2509.14132*, 2025.

[54] Xuanzhao Dong, Wenhui Zhu, Hao Wang, Xiwen Chen, Peijie Qiu, Rui Yin, Yi Su, and Yalin Wang. Talk before you retrieve: Agent-led discussions for better rag in medical qa. *arXiv preprint arXiv:2504.21252*, 2025.

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

[56] Zhuoyun Du, LujieZheng LujieZheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haochao Ying. Llms can simulate standardized patients via agent coevolution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17278–17306, 2025.

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

[58] Ahmed T Elboardy, Ghada Khoriba, and Essam A Rashed. Medical ai consensus: A multi-agent framework for radiology report generation and evaluation. *arXiv preprint arXiv:2509.17353*, 2025.

[59] Taine J Elliott, Stephen P Levitt, Ken Nixon, and Martin Bekker. Data overdose? time for a quadruple shot: Knowledge graph construction using enhanced triple extraction. In *Annual Conference of South African Institute of Computer Scientists and Information Technologists*, pp. 224–240. Springer, 2025.

[60] Mariya Evtimova-Gardair. Multi-agent searching system for medical information. *arXiv preprint arXiv:2203.12465*, 2022.

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

[62] Lin Fan, Pengyu Dai, Zhipeng Deng, Haolin Wang, Xun Gong, Yefeng Zheng, and Yafei Ou. Evolving medical imaging agents via experience-driven self-skill discovery. *arXiv preprint arXiv:2603.05860*, 2026.

[63] Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10183–10213. Association for Computational Linguistics, 2025.

[64] Eman Ebrahim Fateel. Self-assessment of content, pedagogy, and technology knowledge among higher education academics in bahrain. 2025.

[65] Nima Fathi, Amar Kumar, and Tal Arbel. Aura: A multi-modal medical agent for understanding, reasoning and annotation. In *International Workshop on Agentic AI for Medicine*, pp. 105–114. Springer, 2025.

[66] Andrey Fedorov, William JR Longabaugh, David Pot, David A Clunie, Steve Pieper, et al. The nci imaging data commons as a platform for reproducible research in computational pathology. *arXiv preprint arXiv:2303.09354*, 2023.

[67] Jana Fehr, Brian Citro, Rohit Malpani, Christoph Lippert, and Vince I Madai. A trustworthy ai reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, 6:1267290, 2024.

[68] Jinghao Feng, Qiaoyu Zheng, Chaoyi Wu, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. M 3 builder: A multi-agent system for automated machine learning in medical imaging. In *International Workshop on Agentic AI for Medicine*, pp. 115–124. Springer, 2025.

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

[70] Yushi Feng, Junye Du, Yingying Hong, Qifan Wang, and Lequan Yu. Pass: Probabilistic agentic supernet sampling for interpretable and adaptive chest x-ray reasoning. *arXiv preprint arXiv:2508.10501*, 2025.

[71] Dyke Ferber, Omar SM El Nahhas, Georg Wölflein, Isabella C Wiest, Jan Clusmann, Marie-Elisabeth Leßman, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, et al. Autonomous artificial intelligence agents for clinical decision making in oncology. *arXiv preprint arXiv:2404.04667*, 2024.

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

[73] Paul Festor, Ibrahim Habli, Yan Jia, Anthony Gordon, A Aldo Faisal, and Matthieu Komorowski. Levels of autonomy and safety assurance for ai-based clinical decision systems. In *International Conference on Computer Safety, Reliability, and Security*, pp. 291–296. Springer, 2021.

[74] Tatiana Fountoukidou and Raphael Sznitman. A reinforcement learning approach for vqa validation: An application to diabetic macular edema grading. *Medical image analysis*, 87:102822, 2023.

[75] AN EVIDENCE-GROUNDED AGENTIC FRAMEWORK. Care: Towards clinical accountability in multi-modal medical reasoning with an evidence-grounded agentic framework. 2026.

[76] Joaqun M Fuster. Upper processing stages of the perception–action cycle. *Trends in cognitive sciences*, 8(4):143–145, 2004.

[77] Davide Gabrielli, Paola Velardi, Stefano Faralli, and Bardh Prenkaj. A multi-agent framework for interpreting multivariate physiological time series. *arXiv preprint arXiv:2603.04142*, 2026.

[78] Agasthya Gangavarapu and Ananya Gangavarapu. Imas: A comprehensive agentic approach to rural healthcare delivery. *arXiv preprint arXiv:2410.12868*, 2024.

[79] Bowen Gao, Yinjun Jia, Yuanle Mo, Yuyan Ni, Weiying Ma, Zhiming Ma, and Yanyan Lan. Profsa: Self-supervised pocket pretraining via protein fragment-surroundings align. *arXiv preprint arXiv:2310.07229*, 2023.

[80] Jifan Gao, Mahmudur Rahman, John Caskey, Madeline Oguss, Ann O'Rourke, Randy Brown, Anne Stey, Anoop Mayampurath, Matthew M Churpek, Guanhua Chen, et al. Moma: A mixture-of-multimodal-agents architecture for enhancing clinical prediction modelling. *arXiv preprint arXiv:2508.05492*, 2025.

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

[82] Sakharam Gawade, Shivam Akhouri, Chinmay Kulkarni, Jagdish Samant, Pragya Sahu, Jai Pahal, Saswat Meher, et al. Multi agent based medical assistant for edge devices. *arXiv preprint arXiv:2503.05397*, 2025.

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

[84] Anne Gerdes. The role of explainability in ai-supported medical decision-making. *Discover Artificial Intelligence*, 4(1):29, 2024.

[85] Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G Elmore, Ranjay Krishna, and Linda Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. *arXiv preprint arXiv:2502.08916*, 2025.

[86] Katherine J Gold, Chidimma J Udegbunam, Eve H Shikanov, Chloe Miwa, Luke J DeRoos, Amy Cohn, and Emmett Springer. In-basket message volume in primary care: A cross-sectional analysis by gender and specialty. *Journal of General Internal Medicine*, pp. 1–2, 2025.

[87] Haozhen Gong, Xiaozhong Ji, Yuansen Liu, Wenbin Wu, Xiaoxiao Yan, Jingjing Liu, Kai Wu, Jiazhen Pan, Bailiang Jian, Jiangning Zhang, Xiaobin Hu, and Hongwei Bran Li. Med-cmr: A fine-grained benchmark integrating visual evidence and clinical logic for medical complex multimodal reasoning, 2025. URL https://arxiv.org/abs/2512.00818.

[88] Ross Gore, Eranga Bandara, Sachin Shetty, Alberto E Musto, Pratip Rana, Ambrosio Valencia-Romero, Christopher Rhea, Lobat Tayebi, Heather Richter, and Atmaram Yarlagadda. Proof-of-tbi–fine-tuned vision language model consortium and openai-o3 reasoning llm-based medical diagnosis support system for mild traumatic brain injury (tbi) prediction. *arXiv preprint arXiv:2504.18671*, 2025.

[89] Alon Gorenshtein, Mahmud Omar, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. Ai agents in clinical medicine: A systematic review. *medRxiv*, pp. 2025–08, 2025.

[90] Lei Gu, Yinghao Zhu, Haoran Sang, Zixiang Wang, Dehao Sui, Wen Tang, Ewen Harrison, Junyi Gao, Lequan Yu, and Liantao Ma. Medagentaudit: Diagnosing and quantifying collaborative failure modes in medical multi-agent systems. *arXiv preprint*, 2025.

[91] Zishan Gu, Fenglin Liu, Changchang Yin, and Ping Zhang. Inquire, interact, and integrate: A proactive agent collaborative framework for zero-shot multimodal medical reasoning. *arXiv preprint arXiv:2405.11640*, 2024.

[92] Guangfu Guo, Xiaoqian Lu, and Yue Feng. Med-vragent: A framework for medical visual reasoning-enhanced agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18613–18627, 2025.

[93] Seungjun Han and Wongyung Choi. Development of a large language model-based multi-agent clinical decision support system for korean triage and acuity scale (ktas)-based triage and treatment planning in emergency departments. *arXiv preprint arXiv:2408.07531*, 2024.

[94] Yu Han and Zekun Guo. Regulator-manufacturer ai agents modeling: Mathematical feedback-driven multi-agent llm framework. *arXiv preprint arXiv:2411.15356*, 2024.

[95] Yu Han, Aaron Ceross, and Jeroen HM Bergmann. Standard applicability judgment and cross-jurisdictional reasoning: A rag-based framework for medical device compliance. *arXiv preprint arXiv:2506.18511*, 2025.

[96] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

[97] Julia Harrington et al. Nursellm: The first specialized language model for nursing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Industry Track)*, 2025.

[98] K. He. Medagentbench: Dataset for benchmarking llms as agents. *arXiv preprint arXiv:2501.14654*, 2025.

[99] Pengfei He, Yuping Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. Red-teaming llm multi-agent systems via communication attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6726–6747, 2025.

[100] Weijie He and Ting Chen. Scalable online disease diagnosis via multi-model-fused actor-critic reinforcement learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[101] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[102] Yexiao He, Ang Li, Boyi Liu, Zhewei Yao, and Yuxiong He. Medorch: Medical diagnosis with tool-augmented reasoning agents for flexible extensibility. *arXiv preprint arXiv:2506.00235*, 2025.

[103] Yufei He, Juncheng Liu, Zhiyuan Hu, Yulin Chen, Yue Liu, Yuan Sui, Yibo Li, Nuo Chen, Jun Hu, Bryan Hooi, et al. Evoclinician: A self-evolving agent for multi-turn medical diagnosis via test-time evolutionary learning. *arXiv preprint arXiv:2601.22964*, 2026.

[104] Dan Hendrycks, Collin Burns, Steven Basart, et al. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

[105] Thanh Cong Ho, Farah Kharrat, Abderrazek Abid, Fakhri Karray, and Anis Koubaa. Remoni: An autonomous system integrating wearables and multimodal large language models for enhanced remote health monitoring. In *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6. IEEE, 2024.

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

[107] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2023.

[108] Md Imran Hossain, Ghada Zamzmi, Peter R Mouton, Md Sirajus Salekin, Yu Sun, and Dmitry Goldgof. Explainable ai for medical data: Current methods, limitations, and future directions. *ACM Computing Surveys*, 57(6):1–46, 2025.

[109] Wenlong Hou, Guangqian Yang, Ye Du, Yeung Lau, Lihao Liu, Junjun He, Ling Long, and Shujun Wang. Adagent: Llm agent for alzheimer's disease analysis with collaborative coordinator. In *International Workshop on Agentic AI for Medicine*, pp. 23–32. Springer, 2025.

[110] Zefa Hu, Haozhi Zhao, Yuanyuan Zhao, Shuang Xu, and Bo Xu. T-agent: A term-aware agent for medical dialogue generation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2024.

[111] Zheang Huai, Honglong Yang, and Xiaomeng Li. Which tool response should i trust? tool-expertise-aware chest x-ray agent with multimodal agentic learning. *arXiv preprint arXiv:2602.21517*, 2026.

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

[113] Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael R Lyu, and Maarten Sap. On the resilience of llm-based multi-agent collaboration with faulty agents. *arXiv preprint arXiv:2408.00989*, 2024.

[114] A Human-AI. Image segmentation using only" better or worse" expert feedback. In *Human-AI Collaboration: First International Workshop, HAIC 2025, Held in Conjunction with MICCAI 2025, Daejeon, South Korea, September 27, 2025, Proceedings*, pp. 3. Springer Nature, 2025.

[115] Yoshitaka Inoue, Tianci Song, Xinling Wang, Augustin Luna, and Tianfan Fu. Drugagent: Multi-agent large language model-based reasoning for drug-target interaction prediction. *ArXiv*, pp. arXiv–2408, 2025.

[116] Ruba Islayem, Senay Gebreab, Walaa AlKhader, Ahmad Musamih, Khaled Salah, Raja Jayaraman, and Muhammad Khurram Khan. Using large language models for enhanced fraud analysis and detection in blockchain based health insurance claims. *Scientific Reports*, 15(1):29763, 2025.

[117] Akshay Jagatap, Srujana Merugu, and Prakash Mandayam Comar. 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)*, pp. 822–832, 2025.

[118] Won Seok Jang, Hieu Tran, Manav Mistry, SaiKiran Gandluri, Yifan Zhang, Sharmin Sultana, Sunjae Kown, Yuan Zhang, Zonghai Yao, and Hong Yu. Chatbot to help patients understand their health. *arXiv preprint arXiv:2509.05818*, 2025.

[119] Jihoon Jeong. Model medicine: A clinical framework for understanding, diagnosing, and treating ai models. *arXiv preprint arXiv:2603.04722*, 2026.

[120] Rui Jiao, Yue Zhang, and Jinku Li. Trustworthy reasoning: Evaluating and enhancing factual accuracy in llm intermediate thought processes. *arXiv preprint arXiv:2507.22940*, 2025.

[121] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

[122] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[123] Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, Nikhil Khandekar, Nicholas Wan, Xuguang Ai, and W John Wilbur. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nature Communications*, 16(1):9377, 2025.

[124] Manuel Joy. Agentic workflows in healthcare: Advancing clinical efficiency through ai integration. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11:567–575, 03 2025. doi: 10.32628/CSEIT25112396.

[125] Benjamin Kane, Catherine Giugno, Lenhart Schubert, Kurtis Haut, Caleb Wohn, and Ehsan Hoque. A flexible schema-guided dialogue management framework: From friendly peer to virtual standardized cancer patient. *arXiv preprint arXiv:2207.07276*, 2022.

[126] Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, et al. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*, 2022.

[127] Ishan Kavathekar, Hemang Jain, Ameya Rathod, Ponnurangam Kumaraguru, and Tanuja Ganu. Tamas: Benchmarking adversarial risks in multi-agent llm systems. *arXiv preprint arXiv:2511.05269*, 2025.

[128] Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. Enhancing diagnostic accuracy through multi-agent conversations: using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*, 2024.

[129] 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. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *Journal of Medical Internet Research*, 26: e59439, 2024.

[130] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[131] Ho-Jung Kim, Dogeun Park, Jae-Jun Lee, Jin-Pyeong Jeon, and Dong-Ok Won. A proposed llm-based supported treatment framework for intracerebral hemorrhage. In *2025 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–4. IEEE, 2025.

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

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

[134] Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, et al. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*, 2025.

[135] Yubin Kim, Hyewon Jeong, Chanwoo Park, Eugene Park, Haipeng Zhang, Xin Liu, Hyeonhoon Lee, Daniel McDuff, Marzyeh Ghassemi, Cynthia Breazeal, et al. Tiered agentic oversight: A hierarchical multi-agent system for ai safety in healthcare. *arXiv preprint arXiv:2506.12482*, 2025.

[136] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

[137] Kabir Kumar. Benchmarking automatic speech recognition coupled llm modules for medical diagnostics. *arXiv preprint arXiv:2502.13982*, 2025.

[138] Yi Lai, Atreyi Kankanhalli, and Desmond Ong. Human-ai collaboration in healthcare: A review and research agenda. 2021.

[139] Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. *arXiv preprint arXiv:2510.04284*, 2025.

[140] Yunghwei Lai, Weizhi Ma, and Yang Liu. Patient-zero: A unified framework for real-record-free patient agent generation. *arXiv preprint arXiv:2509.11078*, 2025.

[141] Jason J. Lau, Soumick Gayen, Avi Ben-Cohen, et al. A dataset of clinically generated visual questions and answers about radiology images. *Radiology: Artificial Intelligence*, 2018.

[142] Chanseo Lee, Sonu Kumar, Kimon A. Vogt, and Sam Meraj. Improving clinical documentation with AI: A comparative study of sporo AI scribe and GPT-4o mini. *arXiv preprint arXiv:2410.15528*, 2024.

[143] Kwanhyung Lee, Sungsoo Hong, Joonhyung Park, Jeonghyeop Lim, Juhwan Choi, Donghwee Yoon, and Eunho Yang. EMR-AGENT: Automating cohort and feature extraction from EMR databases. *CoRR*, abs/2510.00549, 2025.

[144] M. Lee. Large language models in real-world clinical workflows: a systematic review of applications and implementation. *Journal of Medical Internet Research*, 27(1):e88888, 2025.

[145] Wonjun Lee, Riley CW O'Neill, Dongmian Zou, Jeff Calder, and Gilad Lerman. Geometry-preserving encoder/decoder in latent generative models. *arXiv preprint arXiv:2501.09876*, 2025.

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

[147] Chenjun Li, Laurin Lux, Alexander H Berger, Martin J Menten, Mert R Sabuncu, and Johannes C Paetzold. Fine-tuning vision language models with graph-based knowledge for explainable medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 198–207. Springer, 2025.

[148] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[149] Guoyi Li, Shihao Xu, Jiatong Ma, Yunyun Han, Jianhua Chen, and Yafeng Deng. Mind: Unified inquiry and diagnosis rl with criteria grounded clinical supports for psychiatric consultation. *arXiv preprint arXiv:2603.03677*, 2026.

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

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

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

[153] Sijing Li, Tianwei Lin, Lingshuai Lin, Wenqiao Zhang, Jiang Liu, Xiaoda Yang, Juncheng Li, Yucheng He, Xiaohui Song, Jun Xiao, et al. Eyecaregpt: Boosting comprehensive ophthalmology understanding with tailored dataset, benchmark and model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3893–3902, 2025.

[154] Songhao Li, Jonathan Xu, Tiancheng Bao, Yuxuan Liu, Yuchen Liu, Yihang Liu, Lilin Wang, Wenhui Lei, Sheng Wang, Yinuo Xu, et al. A co-evolving agentic ai system for medical imaging analysis. *arXiv preprint arXiv:2509.20279*, 2025.

[155] Wenjie Li, Yujie Zhang, Haoran Sun, Xingqi He, Hongcheng Gao, Chenglong Ma, Ming Hu, Guankun Wang, Shiyi Yao, Renhao Yang, et al. Medscope: Incentivizing "think with videos" for clinical reasoning via coarse-to-fine tool calling. *arXiv preprint arXiv:2602.13332*, 2026.

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

[157] Xiang Li, Huizi Yu, Wenkong Wang, Yiran Wu, Jiayan Zhou, Wenyue Hua, Xinxin Lin, Wenjia Tan, Lexuan Zhu, Bingyi Chen, et al. Dispatchmas: Fusing taxonomy and artificial intelligence agents for emergency medical services. *arXiv preprint arXiv:2510.21228*, 2025.

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

[159] Xueyang Li, Mingze Jiang, Gelei Xu, Jun Xia, Mengzhao Jia, Danny Chen, and Yiyu Shi. At-cxr: Uncertainty-aware agentic triage for chest x-rays. *arXiv preprint arXiv:2508.19322*, 2025.

[160] Yuxuan Li, Hirokazu Shirado, and Sauvik Das. Actions speak louder than words: Agent decisions reveal implicit biases in language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3303–3325, 2025.

[161] Zheng Li, Jiayi Xu, Zhikai Hu, Hechang Chen, Lele Cong, Yunyun Wang, and Shuchao Pang. Medcorag: Interpretable hepatology diagnosis via hybrid evidence retrieval and multispecialty consensus. *arXiv preprint arXiv:2603.05129*, 2026.

[162] Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. Reflectool: Towards reflection-aware tool-augmented clinical agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[163] Yusheng Liao, Chuan Xuan, Yutong Cai, Lina Yang, Zhe Chen, Yanfeng Wang, and Yu Wang. Agentehr: Advancing autonomous clinical decision-making via retrospective summarization. *arXiv preprint arXiv:2601.13918*, 2026.

[164] Junfan Lin, Keze Wang, Ziliang Chen, Xiaodan Liang, and Liang Lin. Towards causality-aware inferring: A sequential discriminative approach for medical diagnosis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13363–13375, 2023.

[165] Hongjun Liu, Yinghao Zhu, Yuhui Wang, Yitao Long, Zeyu Lai, Lequan Yu, and Chen Zhao. Medmmv: A controllable multimodal multi-agent framework for reliable and verifiable clinical reasoning. *arXiv preprint arXiv:2509.24314*, 2025.

[166] Jie Liu et al. Medchain: Bridging the gap between llm agents and clinical practice through interactive sequential benchmarking. *arXiv preprint arXiv:2412.01605*, 2024.

[167] Jinsong Liu, Yuhang Jiang, Ramayya Krishnan, Rema Padman, Yiye Zhang, and Jiang Bian. Closing reasoning gaps in clinical agents with differential reasoning learning. *arXiv preprint arXiv:2602.09945*, 2026.

[168] Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu, Ming Zhang, Xiaowei Ma, Xiangguo Lyu, Liya Ma, Zhiqiang Zhang, Wei Xue, Yiran Huang, and Jinjie Gu. Towards automatic evaluation for LLMs' clinical capabilities: Metric, data, and algorithm. *arXiv preprint arXiv:2403.16446*, 2024.

[169] Philip R Liu, Sparsh Bansal, Jimmy Dinh, Aditya Pawar, Ramani Satishkumar, Shail Desai, Neeraj Gupta, Xin Wang, and Shu Hu. Medchat: A multi-agent framework for multimodal diagnosis with large language models. *arXiv preprint arXiv:2506.07400*, 2025.

[170] Shengyuan Liu, Liuxin Bao, Qi Yang, Wanting Geng, Boyun Zheng, Chenxin Li, Wenting Chen, Houwen Peng, and Yixuan Yuan. Medsam-agent: Empowering interactive medical image segmentation with multi-turn agentic reinforcement learning. *arXiv preprint arXiv:2602.03320*, 2026.

[171] Wei Liu, Tianyang Tang, Peng Zhu, et al. Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *Knowledge Science, Engineering and Management*. Springer, 2022.

[172] Wenge Liu, Yi Cheng, Hao Wang, Jianheng Tang, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. "my nose is running." "are you also coughing?": Building a medical diagnosis agent with interpretable inquiry logics. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, 2022.

[173] Ziji Liu, Liang Xiao, Rujun Zhu, Hang Yang, and Miaomiao He. Medgen: An explainable multi-agent architecture for clinical decision support through multisource knowledge fusion. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 6474–6481. IEEE, 2024.

[174] Chang Han Low, Ziyue Wang, Tianyi Zhang, Zhitao Zeng, Zhu Zhuo, Evangelos B Mazomenos, and Yueming Jin. Surgraw: Multi-agent workflow with chain-of-thought reasoning for surgical intelligence. *arXiv preprint arXiv:2503.10265*, 2025.

[175] Chen Lyu and Gabriele Pergola. Society of medical simplifiers. *arXiv preprint arXiv:2410.09631*, 2024.

[176] Chaofan Ma, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. Boundary-aware supervoxel-level iteratively refined interactive 3d image segmentation with multi-agent reinforcement learning. *IEEE Transactions on Medical Imaging*, 40(10):2563–2574, 2020.

[177] Siqi Ma, Jiajie Huang, Bolin Yang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. Medla: A logic-driven multi-agent framework for complex medical reasoning with large language models. *arXiv preprint arXiv:2509.23725*, 2025.

[178] Rakesh Raj Madavan, Akshat Kaimal, Hashim Faisal, and S Chandrakala. Med-grim: Enhanced zero-shot medical vqa using prompt-embedded multimodal graph rag. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4081–4091, 2025.

[179] Cesare Magnetti, Hadrien Reynaud, and Bernhard Kainz. Cross modality 3d navigation using reinforcement learning and neural style transfer. *arXiv preprint arXiv:2111.03485*, 2021.

[180] Arjun Mahajan and Dylan Powell. Transforming healthcare delivery with conversational ai platforms. *NPJ Digital Medicine*, 8(1):581, 2025.

[181] Jecia ZY Mao, Francis X Creighton, Russell H Taylor, and Manish Sahu. Scope: Speech-guided collaborative perception framework for surgical scene segmentation. In *International Workshop on Emerging LLM/LMM Applications in Medical Imaging*, pp. 71–78. Springer, 2025.

[182] Yuren Mao, Wenyi Xu, Yuyang Qin, and Yunjun Gao. Ct-agent: A multimodal-llm agent for 3d ct radiology question answering. *arXiv preprint arXiv:2505.16229*, 2025.

[183] Kanato Masayoshi, Masahiro Hashimoto, Ryoichi Yokoyama, Naoki Toda, Yoshifumi Uwamino, Shogo Fukuda, Ho Namkoong, and Masahiro Jinzaki. Ehr-mcp: Real-world evaluation of clinical information retrieval by large language models via model context protocol. *arXiv preprint arXiv:2509.15957*, 2025.

[184] Oded Medina, Liora Kleinburd, and Nir Shvalb. Navigation through endoluminal channels using q-learning. *arXiv preprint arXiv:2309.03615*, 2023.

[185] Nikita Mehandru, Brenda Y Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J Butte, and Ahmed Alaa. Large language models as agents in the clinic. *arXiv preprint arXiv:2309.10895*, 2023.

[186] Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J. Butte, and Ahmed Alaa. Evaluating large language models as agents in the clinic. *npj Digital Medicine*, 7(1): 84, 2024.

[187] Viraj Mehta, Abhinav Komanduri, Rishabh Singh Bhadouriya, Vilina Mehta, Michael David Johnson, Priyanka Shrestha, Margaret Nikolov, Bhav Jain, Nigam Shah, and Kevin Schulman. Evaluating transparency in ai/ml model characteristics for fda-reviewed medical devices. *npj Digital Medicine*, 8 (1):673, 2025.

[188] Yang Meng, Zewen Pan, Yandi Lu, Ruobing Huang, Yanfeng Liao, and Jiarui Yang. Cataractsurg-80k: Knowledge-driven benchmarking for structured reasoning in ophthalmic surgery planning. *arXiv preprint arXiv:2508.20014*, 2025.

[189] Marziyeh Mohammadi, Mohsen Vejdanihemmat, Mahshad Lotfinia, Mirabela Rusu, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. Differential privacy for medical deep learning: methods, tradeoffs, and deployment implications. *arXiv preprint arXiv:2506.00660*, 2025.

[190] Abdul M Mohammed, Iqtidar Mansoor, Sarah Blythe, and Dennis Trujillo. Developing an artificial intelligence tool for personalized breast cancer treatment plans based on the nccn guidelines. *arXiv preprint arXiv:2502.15698*, 2025.

[191] E. Mornin. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. *npj Digital Medicine*, 8(1):12–24, 2025.

[192] Andreas Motzfeldt, Joakim Edin, Casper L Christensen, Christian Hardmeier, Lars Maaløe, and Anna Rogers. Code like humans: A multi-agent solution for medical coding. *arXiv preprint arXiv:2509.05378*, 2025.

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

[194] Tom Nadarzynski, Nicky Knights, Deborah Husbands, Cynthia A Graham, Carrie D Llewellyn, Tom Buchanan, Ian Montgomery, and Damien Ridge. Achieving health equity through conversational ai: A roadmap for design and implementation of inclusive chatbots in healthcare. *PLOS Digital Health*, 3(5): e0000492, 2024.

[195] Matthew B Neeley, Guantong Qi, Guanchu Wang, Ruixiang Tang, Dongxue Mao, Chaozhong Liu, Sasidhar Pasupuleti, Bo Yuan, Fan Xia, Pengfei Liu, et al. Survey and improvement strategies for gene prioritization with large language models. *Bioinformatics Advances*, pp. vbaf148, 2025.

[196] Huy Hoang Nguyen, Matthew B Blaschko, Simo Saarakkala, and Aleksei Tiulpin. Clinically-inspired multi-agent transformers for disease trajectory forecasting from multimodal data. *IEEE transactions on medical imaging*, 43(1):529–541, 2023.

[197] Andrei Niculae, Adrian Cosma, Cosmin Dumitrache, and Emilian Radoi. Dr. copilot: A multi-agent prompt optimized assistant for improving patient-doctor communication in romanian. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1780–1792, 2025.

[198] Yingpeng Ning, Yuanyuan Sun, Ling Luo, Yanhua Wang, Yuchen Pan, and Hongfei Lin. Medtrust-rag: Evidence verification and trust alignment for biomedical question answering. *arXiv preprint arXiv:2510.14400*, 2025.

[199] Sota Nishisako, Takahiro Higashi, and Fumihiko Wakao. Reducing hallucinations and trade-offs in responses in generative ai chatbots for cancer information: Development and evaluation study. *JMIR cancer*, 11(1):e70176, 2025.

[200] Mengjia Niu, Hao Li, Jie Shi, Hamed Haddadi, and Fan Mo. Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval. *arXiv preprint arXiv:2405.06545*, 2024.

[201] Humza Nusrat, Bing Luo, Ryan Hall, Joshua Kim, Hassan Bagher-Ebadian, Anthony Doemer, Benjamin Movsas, and Kundan Thind. Autonomous radiotherapy treatment planning using dola: A privacy-preserving, llm-based optimization agent. *arXiv preprint arXiv:2503.17553*, 2025.

[202] Çağatay Umut Öğdü, Kübra Arslanoğlu, and Mehmet Karaköse. An adaptive multi-agent llm-based clinical decision support system integrating biomedical rag and web intelligence. *IEEE Access*, 2025.

[203] Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang, Wasswa William, Atul J Butte, Nigam H Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, et al. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, 6(6):e428–e432, 2024.

[204] OpenAI. Openai o1 system card, 2024.

[205] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 248–260. PMLR, 2022.

[206] Satrio Pambudi and Filippo Menolascina. Bridging clinical narratives and acr appropriateness guidelines: A multi-agent rag system for medical imaging decisions. *arXiv preprint arXiv:2510.04969*, 2025.

[207] Jiazhen Pan, Bailiang Jian, Paul Hager, et al. Beyond benchmarks: Dynamic, automatic and systematic red-teaming agents for trustworthy medical language models. *arXiv preprint arXiv:2508.00923*, 2025.

[208] Xiaoyu Pan, Yang Bai, Ke Zou, Yang Zhou, Jun Zhou, Huazhu Fu, Yih-Chung Tham, and Yong Liu. Eh-benchmark: Ophthalmic hallucination benchmark and agent-driven top-down traceable reasoning workflow. *Information Fusion*, pp. 103631, 2025.

[209] Himanshu Gautam Pandey, Akhil Amod, and Shivang Kumar. Advancing healthcare automation: Multi-agent system for medical necessity justification. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 39–49, 2024.

[210] Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. *arXiv preprint arXiv:2502.14302*, 2025.

[211] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

[212] Qi Peng, Jiankun Liu, Quan Zou, Xing Chen, Zheng Zhong, Zefeng Wang, Jiayuan Xie, Yi Cai, and Qing Li. Integration of multi-source medical data for medical diagnosis question answering. *IEEE Transactions on Medical Imaging*, 2024.

[213] Qi Peng, Jialin Cui, Jiayuan Xie, Yi Cai, and Qing Li. Tree-of-reasoning: Towards complex medical diagnosis via multi-agent reasoning with evidence tree. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1744–1753, 2025.

[214] Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. Aligning (medical) LLMs for (counterfactual) fairness. *arXiv preprint arXiv:2408.12055*, 2024.

[215] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15174–15186, 2024.

[216] Yunhang Qian, Xiaobin Hu, Jiaquan Yu, Siyang Xin, Xiaokun Chen, Jiangning Zhang, Peng-Tao Jiang, Jiawei Liu, and Hongwei Bran Li. Medmaslab: A unified orchestration framework for benchmarking multimodal medical multi-agent systems. *arXiv preprint arXiv:2603.09909*, 2026.

[217] Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 615–636, 2024.

[218] Jianing Qiu, Lin Li, Jiankai Sun, Hao Wei, Zhe Xu, Kyle Lam, and Wu Yuan. Emerging cyber attack risks of medical ai agents. *arXiv preprint arXiv:2504.03759*, 2025.

[219] Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343, 2025.

[220] Zhan Qu and Michael Färber. Trace: Temporal reasoning via agentic context evolution for streaming electronic health records (ehrs). *arXiv preprint arXiv:2602.12833*, 2026.

[221] Ngoc Bui Lam Quang, Nam Le Nguyen Binh, Thanh-Huy Nguyen, Le Thien Phuc Nguyen, Quan Nguyen, and Ulas Bagci. Gmat: Grounded multi-agent clinical description generation for text encoder in vision-language mil for whole slide image classification. In *International Workshop on Emerging LLM/LMM Applications in Medical Imaging*, pp. 1–9. Springer, 2025.

[222] Zhiyao Ren, Yibing Zhan, Baosheng Yu, Liang Ding, Pingbo Xu, and Dacheng Tao. Healthcare agent: Eliciting the power of large language models for medical consultation. *npj Artificial Intelligence*, 1(24), 2025.

[223] David Restrepo, Chenwei Wu, Zhengxu Tang, Zitao Shuai, Thao Nguyen Minh Phan, Jun-En Ding, Cong-Tinh Dao, Jack Gallifant, Robyn Gayle Dychiao, Jose Carlo Artiaga, André Hiroshi Bando, Carolina Pelegrini Barbosa Gracitelli, Vincenz Ferrer, Leo Anthony Celi, Danielle Bitterman, Michael G. Morley, and Luis Filipe Nakayama. Multi-ophthalingua: A multilingual benchmark for assessing and debiasing llm ophthalmological qa in LMICs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[224] Mohammad Reza Rezaei, Reza Saadati Fard, Jayson L Parker, Rahul G Krishnan, and Milad Lankarany. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between llms and evolving medical knowledge. *arXiv preprint arXiv:2502.13010*, 2025.

[225] Harry Robertshaw, Han-Ru Wu, Alejandro Granados, and Thomas C Booth. World model for ai autonomous navigation in mechanical thrombectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 680–690. Springer, 2025.

[226] Daniel Philip Rose, Chia-Chien Hung, Marco Lepri, Israa Alqassem, Kiril Gashteovski, and Carolin Lawrence. Meddxagent: A unified modular agent framework for explainable automatic differential diagnosis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13803–13826, 2025.

[227] Shaheer U Saeed, Yunguan Fu, Vasilis Stavrinides, Zachary MC Baum, Qianye Yang, Mirabela Rusu, Richard E Fan, Geoffrey A Sonn, J Alison Noble, Dean C Barratt, et al. Adaptable image quality assessment using meta-reinforcement learning of task amenability. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pp. 191–201. Springer, 2021.

[228] Pramit Saha, Joshua Strong, Divyanshu Mishra, Cheng Ouyang, and J. Alison Noble. Fedagentbench: Towards automating real-world federated medical image analysis with server–client llm agents. *arXiv preprint arXiv:2509.23803*, 2025.

[229] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

[230] Reza Samimi, Aditya Bhattacharya, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. Visual-conversational interface for evidence-based explanation of diabetes risk prediction. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pp. 1–18, 2025.

[231] Peter Sarvari and Zaid Al-Fagih. Rapidly benchmarking large language models for diagnosing comorbid patients: comparative study leveraging the llm-as-a-judge method. *JMIRx Med*, 6:e67661, 2025.

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

[233] MD Ragib Shahriyear. Guidelineguard: An agentic framework for medical note evaluation with guideline adherence. *arXiv preprint arXiv:2411.06264*, 2024.

[234] Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, Hong-Ning Dai, Feng Zhao, and Jianming Yong. Adaptive multi-agent deep reinforcement learning for timely healthcare interventions. *arXiv preprint arXiv:2309.10980*, 2023.

[235] Fangxin Shang, Yuan Xia, Dalu Yang, Yahui Wang, and Binglin Yang. Medrepbench: A comprehensive benchmark for medical report interpretation. *arXiv preprint arXiv:2508.16674*, 2025.

[236] Naman Sharma. Cxr-agent: Vision-language models for chest x-ray interpretation with uncertainty aware radiology reporting. *arXiv preprint arXiv:2407.08811*, 2024.

[237] Mohammed A. Shehab. Agentic-ai healthcare: Multilingual, privacy-first framework with MCP agents. *CoRR*, abs/2510.02325, 2025.

[238] Chen Shen, Wanqing Zhang, Kehan Li, Erwen Huang, Haitao Bi, Aiying Fan, Yiwen Shen, Hongmei Dong, Ji Zhang, Yuming Shao, et al. Feat: A multi-agent forensic ai system with domain-adapted large language model for automated cause-of-death analysis. *arXiv preprint arXiv:2508.07950*, 2025.

[239] Chuyun Shen, Wenhao Li, Qisen Xu, Bin Hu, Bo Jin, Haibin Cai, Fengping Zhu, Yuxin Li, and Xiangfeng Wang. Interactive medical image segmentation with self-adaptive confidence calibration. *Frontiers of Information Technology & Electronic Engineering*, 24(9):1332–1348, 2023.

[240] Chuyun Shen, Wenhao Li, Ya Zhang, Yanfeng Wang, and Xiangfeng Wang. Temporally-extended prompts optimization for sam in interactive medical image segmentation. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3550–3557. IEEE, 2023.

[241] Baorong Shi, Bo Cui, Boyuan Jiang, Deli Yu, Fang Qian, Haihua Yang, Huichao Wang, Jiale Chen, Jianfei Pan, Jieqiong Cao, et al. Medxiaohe: A comprehensive recipe for building medical mllms. *arXiv preprint arXiv:2602.12705*, 2026.

[242] Danli Shi, Xiaolan Chen, Bingjie Yan, Weiyi Zhang, Pusheng Xu, Jiancheng Yang, Ruoyu Chen, Siyu Huang, Bowen Liu, Xinyuan Wu, et al. A multimodal ai agent for clinical decision support in ophthalmology. *arXiv preprint arXiv:2511.09394*, 2025.

[243] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024.

[244] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[245] Satya Narayan Shukla and Benjamin M Marlin. Modeling irregularly sampled clinical time series. *arXiv preprint arXiv:1812.00531*, 2018.

[246] Roma Shusterman, Allison C Waters, Shannon O'Neill, Marshall Bangs, Phan Luu, and Don M Tucker. An active inference strategy for prompting reliable responses from large language models in medical practice. *npj Digital Medicine*, 8(1):119, 2025.

[247] Md Kamrul Siam, Md Jobair Hossain Faruk, Bofan He, Jerry Q Cheng, and Huanying Gu. Multimodal models in healthcare: Methods, challenges, and future directions for enhanced clinical decision support. *Information*, 16(11):971, 2025.

[248] Karanbir Singh and William Ngu. Bias-Aware Agent: Enhancing fairness in AI-driven knowledge retrieval. *arXiv preprint arXiv:2503.21237*, 2025.

[249] Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of internal medicine*, 165(11):753–760, 2016.

[250] Moran Sorka, Alon Gorenshtein, Dvir Aran, and Shahar Shelly. A multi-agent approach to neurological clinical reasoning. *arXiv preprint arXiv:2508.14063*, 2025.

[251] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*, 2023.

[252] Peter D Stetson, Suzanne Bakken, Jesse O Wrenn, and Eugenia L Siegler. Assessing electronic note quality using the physician documentation quality instrument (pdqi-9). *Applied clinical informatics*, 3 (02):164–174, 2012.

[253] Haoyang Su, Shaoting Zhang, and Xiaosong Wang. Lammi-pathology: A tool-centric bottom-up lvlm-agent framework for molecularly informed medical intelligence in pathology. *arXiv preprint arXiv:2602.18773*, 2026.

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

[255] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.

[256] Luning Sun, Christopher Gibbons, José Hernández-Orallo, Xiting Wang, Liming Jiang, David Stillwell, Fang Luo, and Xing Xie. Beyond benchmarks: Evaluating generalist medical artificial intelligence with psychometrics. *Journal of medical Internet research*, 27:e70901, 2025.

[257] Zhoujian Sun, Ziyi Liu, Cheng Luo, Jiebin Chu, and Zhengxing Huang. Improving interactive diagnostic ability of a large language model agent through clinical experience learning. *arXiv preprint*, 2025.

[258] Vishal Sunder, Prashant Serai, and Eric Fosler-Lussier. Building an ASR error robust spoken virtual patient system in a highly class-imbalanced scenario without speech data. *arXiv preprint arXiv:2204.05183*, 2022.

[259] Ivan Sviridov, Amina Miftakhova, Tereshchenko Artemiy Vladimirovich, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3mdbench: Medical multimodal multi-agent dialogue benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26625–26665, 2025.

[260] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621. Association for Computational Linguistics, 2024.

[261] Xiangru Tang, Daniel Shao, Jiwoong Sohn, et al. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*, 2025.

[262] Karishma Thakrar, Shreyas Basavatia, and Akshay Daftardar. Architecting clinical collaboration: Multi-agent reasoning systems for multimodal medical vqa. *arXiv preprint arXiv:2507.05520*, 2025.

[263] Sankara Reddy Thamma. Agentic ai for clinical decision support: Real-time diagnosis, triage, and treatment planning. *International Journal of Scientific Research in Science, Engineering and Technology*, 12(3):428–433, 2025.

[264] Matthew Toles, Nikhil Balwani, Rattandeep Singh, Valentina Giulia Sartori Rodriguez, and Zhou Yu. Program synthesis dialog agents for interactive decision-making. *arXiv preprint arXiv:2502.19610*, 2025.

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

[266] Joshua Yi Min Tung, Quan Le, Jinxuan Yao, Yifei Huang, Daniel Yan Zheng Lim, Gerald Gui Ren Sng, Rachel Shu En Lau, Yu Guang Tan, Kenneth Chen, Kae Jack Tay, et al. Performance of retrieval-augmented generation large language models in guideline-concordant prostate-specific antigen testing: Comparative study with junior clinicians. *Journal of Medical Internet Research*, 27:e78393, 2025.

[267] Md Shahab Uddin, Ahsan Ahmed, Md Aktarujjaman, Mohammad Moniruzzaman, Mumtahina Ahmed, MF Mridha, and Md Jakir Hossen. A hybrid reinforcement learning and knowledge graph framework for financial risk optimization in healthcare systems. *Scientific Reports*, 15(1):29057, 2025.

[268] Alyssa Unell, Noel CF Codella, Sam Preston, Peniel Argaw, Wen-wai Yim, Zelalem Gero, Cliff Wong, Rajesh Jena, Eric Horvitz, Amanda K Hall, et al. Cancerguide: Cancer guideline understanding via internal disagreement estimation. *arXiv preprint arXiv:2509.07325*, 2025.

[269] Akhil Vaid, Joshua Lampert, Juhee Lee, Ashwin Sawant, Donald Apakama, Ankit Sakhuja, Ali Soroush, Sarah Bick, Ethan Abbott, Hernando Gomez, Michael Hadley, Denise Lee, Isotta Landi, Son Q. Duong, Nicole Bussola, Ismail Nabeel, Silke Muehlstedt, Robert Freeman, Patricia Kovatch, Brendan Carr, Fei Wang, Benjamin Glicksberg, Edgar Argulian, Stamatios Lerakis, Rohan Khera, David L. Reich, Monica Kraft, Alexander Charney, and Girish Nadkarni. Natural language programming in medicine: Administering evidence based clinical workflows with autonomous agents powered by generative large language models. *arXiv preprint arXiv:2401.02851*, 2024.

[270] Shrish Shrinath Vaidya, Gowthamaan Palani, Sidharth Ramesh, Velmurugan Balasubramanian, Minmini Selvam, Gokulraja Srinivasaraja, and Ganapathy Krishnamurthi. Medpao: A protocol-driven agent for structuring medical reports. In *International Workshop on Agentic AI for Medicine*, pp. 33–45. Springer, 2025.

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

[272] Parth Vashisht, Abhilasha Lodha, Mukta Maddipatla, Zonghai Yao, Avijit Mitra, Zhichao Yang, Junda Wang, Sunjae Kwon, and Hong Yu. Umass-bionlp at mediqa-m3g 2024: Dermprompt–a systematic exploration of prompt engineering with gpt-4v for dermatological diagnosis. *arXiv preprint arXiv:2404.17749*, 2024.

[273] Shubham Vatsal, Harsh Dubey, and Aditi Singh. Agentic ai in healthcare and medicine: A seven-dimensional taxonomy for empirical evaluation of llm-based agents. *IEEE Access*, 14:4840–4863, 2026.

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

[275] Junda Wang, Zonghai Tao, Hansi Zeng, Zhichao Yang, Hamed Zamani, and Hong Yu. Tarse: Test-time adaptation via retrieval of skills and experience for reasoning agents. *arXiv preprint arXiv:2603.01241*, 2026.

[276] Lehan Wang, Yi Qin, Honglong Yang, and Xiaomeng Li. Proactive reasoning-with-retrieval framework for medical multimodal large language models. *arXiv preprint arXiv:2510.18303*, 2025.

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

[278] Pengyu Wang, Shuchang Ye, Usman Naseem, and Jinman Kim. Mrgagents: A multi-agent framework for improved medical report generation with med-lvlms. *arXiv preprint arXiv:2505.18530*, 2025.

[279] Sheng Wang, Fangyuan Zhao, Dechao Bu, Yunwei Lu, Ming Gong, Hongjie Liu, Zhaohui Yang, Xiaoxi Zeng, Zhiyuan Yuan, and Baoping Wan. Lins: A general medical q&a framework for enhancing the quality and credibility of llm-generated responses. *Nature Communications*, 16(1):9076, 2025.

[280] Shouju Wang, Fenglin Yu, Xirui Liu, Xiaoting Qin, Jue Zhang, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. Privacy in action: Towards realistic privacy mitigation and evaluation for llm-powered agents. *arXiv preprint arXiv:2509.17488*, 2025.

[281] Shuyue Wang, Liujie Ren, Tianyao Zhou, Lili Chen, Tianyu Zhang, Yaoyao Fu, and Shuo Wang. Large language model-enhanced interactive agent for public education on newborn auricular deformities. *arXiv preprint arXiv:2409.12984*, 2024.

[282] Xiaoyang Wang and Christopher C. Yang. Balancing fairness and performance in healthcare AI: A gradient reconciliation approach. *arXiv preprint arXiv:2504.14388*, 2025.

[283] Yihan Wang, Qiao Yan, Zhenghao Xing, Lihao Liu, Junjun He, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. Silence is not consensus: Disrupting agreement bias in multi-agent llms via catfish agent for clinical decision making. *arXiv preprint arXiv:2505.21503*, 2025.

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

[285] Zhijun Wang, Ling Luo, Dinghao Pan, Huan Zhuang, Lejing Yu, Yuanyuan Sun, and Hongfei Lin. Rexdrug: Reliable multi-drug combination extraction through reasoning-enhanced llms, 2026.

[286] Ziqing Wang, Chengsheng Mao, Xiaole Wen, Yuan Luo, and Kaize Ding. Amanda: Agentic medical knowledge augmentation for data-efficient medical visual question answering. *arXiv preprint arXiv:2510.02328*, 2025.

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

[288] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[289] Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*, 2025.

[290] Ziyue Wang, Linghan Cai, Chang Han Low, Haofeng Liu, Junde Wu, Jingyu Wang, Rui Wang, Lei Song, Jiang Bian, Jingjing Fu, et al. 3dmedagent: Unified perception-to-understanding for 3d medical analysis. *arXiv preprint arXiv:2602.18064*, 2026.

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

[292] Yuxiang Wei et al. Medaide: Towards an omni medical aide via specialized LLM-based multi-agent collaboration. *arXiv preprint arXiv:2410.12532*, 2024.

[293] Hui Min Wong, Philip Heesen, Pascal Janetzky, Martin Bendszus, and Stefan Feuerriegel. Medclarify: An information-seeking ai agent for medical diagnosis with case-specific follow-up questions. *arXiv preprint arXiv:2602.17308*, 2026.

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

[295] Jun-Yu Wu and Min-Yuh Day. Med-tamara: Trust-aware multi-agent risk assessment in medical ai dialogue. In *2025 IEEE 26th International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 240–245. IEEE, 2025.

[296] Junwei Wu, Runze Yan, Hanqi Luo, Darren Liu, Minxiao Wang, Kimberly L Townsend, Lydia S Hartwig, Derek Milketinas, Xiao Hu, and Carl Yang. Nutriorion: A hierarchical multi-agent framework for personalized nutrition intervention grounded in clinical guidelines. *arXiv preprint arXiv:2602.18650*, 2026.

[297] Shuai Wu, Yilong Chang, Sophie Leanza, Jay Sim, Lu Lu, Qi Li, Diego Stone, and Ruike Renee Zhao. Magnetic milli-spinner for robotic endovascular surgery. *Advanced Materials*, pp. e08180, 2025.

[298] Wenhao Wu, Zhentao Tang, Yafu Li, Shixiong Kai, Mingxuan Yuan, Zhenhong Sun, Chunlin Chen, and Zhi Wang. From conflict to consensus: Boosting medical reasoning via multi-round agentic rag. *arXiv preprint arXiv:2603.03292*, 2026.

[299] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Yanyuan Qiao, Imran Razzak, and Yutong Xie. A knowledge-driven adaptive collaboration of llms for enhancing medical decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33483–33500, 2025.

[300] Yicong Wu, Ting Chen, Irit Hochberg, Zhoujian Sun, Ruth Edry, Zhengxing Huang, and Mor Peleg. Lessons learned from evaluation of llm based multi-agents in safer therapy recommendation. *arXiv preprint arXiv:2507.10911*, 2025.

[301] Yue Wu, Xiaolan Chen, Weiyi Zhang, Shunming Liu, Wing Man Rita Sum, Xinyuan Wu, Xianwen Shang, Chea-su Kee, Mingguang He, and Danli Shi. Chatmyopia: An ai agent for pre-consultation education in primary eye care settings. *arXiv preprint arXiv:2507.19498*, 2025.

[302] Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Shujie Liu, Yan Lu, et al. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. *arXiv preprint arXiv:2506.00555*, 2025.

[303] Ying Xiao, Jie Huang, Ruijuan He, et al. Amqa: An adversarial dataset for benchmarking bias of llms in medicine and healthcare. *arXiv preprint arXiv:2505.19562*, 2025.

[304] Ziyan Xiao, Yinghao Zhu, Liang Peng, and Lequan Yu. Experienceweaver: Optimizing small-sample experience learning for llm-based clinical text improvement. *arXiv preprint arXiv:2602.00740*, 2026.

[305] Yuzhang Xie, Hejie Cui, Ziyang Zhang, Jiaying Lu, Kai Shu, Fadi Nahab, Xiao Hu, and Carl Yang. Kerap: A knowledge-enhanced reasoning approach for accurate zero-shot diagnosis prediction using multi-agent llms. *arXiv preprint arXiv:2507.02773*, 2025.

[306] Shan Xu, Zhaokun Yan, Chengxiao Dai, and Fan Wu. Mega-rag: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of llms in public health. *Frontiers in Public Health*, 13:1635381, 2025.

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

[308] Zhe Xu, Cheng Jin, Yihui Wang, Ziyi Liu, and Hao Chen. Discovering pathology rationale and token allocation for efficient multimodal pathology reasoning. *arXiv preprint arXiv:2505.15687*, 2025.

[309] Zikang Xu, Ruinan Jin, and Xiaoxiao Li. Ducx: Decomposing unfairness in tool-using chest x-ray agents. *arXiv preprint arXiv:2603.00777*, 2026.

[310] Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*, 2024.

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

[312] Xinxin Yan and Ndapandula Nakashole. A grounded well-being conversational agent with multiple interaction modes: Preliminary results. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pp. 143–151, 2021.

[313] Zhiling Yan, Dingjie Song, Zhe Fang, Yisheng Ji, Xiang Li, Quanzheng Li, and Lichao Sun. Livemedbench: A contamination-free medical benchmark for llms with automated rubric evaluation. *arXiv preprint arXiv:2602.10367*, 2026.

[314] Dongrong Yang, Xin Wu, Yibo Xie, Xinyi Li, Qiuwen Wu, Jackie Wu, and Yang Sheng. Zero-shot large language model agents for fully automated radiotherapy treatment planning. *arXiv preprint arXiv:2510.11754*, 2025.

[315] Guang-Zhong Yang, James Cambias, Kevin Cleary, Eric Daimler, James Drake, Pierre E Dupont, Nobuhiko Hata, Peter Kazanzides, Sylvain Martel, Rajni V Patel, et al. Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy, 2017.

[316] Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. Llm-medqa: Enhancing medical question answering through case studies in large language models. *arXiv preprint arXiv:2501.05464*, 2024.

[317] Wanrong Yang, Zhengliang Liu, Yuan Li, Bingjie Yan, Lingfang Li, Mingguang He, Dominik Wojtczak, Yalin Zheng, and Danli Shi. Empowering locally deployable medical agent via state enhanced logical skills for fhir-based clinical tasks. *arXiv preprint arXiv:2603.06902*, 2026.

[318] Yahe Yang and Chengyue Huang. Tree-based rag-agent recommendation system: A case study in medical test data. *arXiv preprint arXiv:2501.02727*, 2025.

[319] Yifan Yang, Qiao Jin, Furong Huang, and Zhiyong Lu. Adversarial attacks on large language models in medicine. *ArXiv*, pp. arXiv–2406, 2024.

[320] Yifan Yang, Qiao Jin, Robert Leaman, Xiaoyu Liu, Guangzhi Xiong, Maame Sarfo-Gyamfi, Changlin Gong, Santiago Ferrière-Steinert, W John Wilbur, Xiaojun Li, et al. Ensuring safety and trust: Analyzing the risks of large language models in medicine. *arXiv preprint arXiv:2411.14487*, 2024.

[321] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

[322] Zonghan Yao et al. Medqa-cs: Benchmarking large language models' clinical skills using an ai-sce framework. *arXiv preprint arXiv:2410.01553*, 2024.

[323] Ziruo Yi, Ting Xiao, and Mark V Albert. A multimodal multi-agent framework for radiology report generation. *arXiv preprint arXiv:2505.09787*, 2025.

[324] Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, et al. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*, 2025.

[325] Chengzhang Yu, Yiming Zhang, Zhixin Liu, Zenghui Ding, Yining Sun, and Zhanpeng Jin. Frame: Feedback-refined agent methodology for enhancing medical research insights. *arXiv preprint arXiv:2505.04649*, 2025.

[326] Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Jingxian He, Wenyue Hua, and Mingyu Jin. Simulated patient systems are intelligent when powered by large language model-based ai agents. *arXiv preprint arXiv:2409.18924*, 10, 2024.

[327] Yuechun Yu, Han Ying, Haoan Jin, Wenjian Jiang, Dong Xian, Binghao Wang, Zhou Yang, and Mengyue Wu. Medkgeval: A knowledge graph-based multi-turn evaluation framework for open-ended patient interactions with clinical llms. *arXiv preprint arXiv:2510.12224*, 2025.

[328] Zhaolin Yu, Litao Yang, Ben Babicka, Ming Hu, Jing Hao, Anthony Huang, James Huang, Yueming Jin, Jiasong Wu, and Zongyuan Ge. Opgagent: An agent for auditable dental panoramic x-ray interpretation. *arXiv preprint arXiv:2603.00462*, 2026.

[329] Grace Chang Yuan, Xiaoman Zhang, Sung Eun Kim, and Pranav Rajpurkar. Do mixed-vendor multi-agent llms improve clinical diagnosis? *arXiv preprint arXiv:2603.04421*, 2026.

[330] Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2024.

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

[332] Haotian Zhai, Elias Stengel-Eskin, Pratik Patil, and Liu Leqi. Evaluating stochasticity in deep research agents. *arXiv preprint arXiv:2602.23271*, 2026.

[333] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2345–2354, 2020.

[334] Yuqi Zhan, Xinyue Wu, Tianyu Lin, Yutong Bao, Xiaoyu Wang, Weihao Cheng, Huangwei Chen, Feiwei Qin, and Zhu Zhu. Medcollab: Causal-driven multi-agent collaboration for full-cycle clinical diagnosis via ibis-structured argumentation. *arXiv preprint arXiv:2603.01131*, 2026.

[335] Duzhen Zhang, Zixiao Wang, Zhong-Zhi Li, Yahan Yu, Shuncheng Jia, Jiahua Dong, Haotian Xu, Xing Wu, Yingying Zhang, Tielin Zhang, et al. Medkgent: A large language model agent framework for constructing temporally evolving medical knowledge graph. *arXiv preprint arXiv:2508.12393*, 2025.

[336] Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025.

[337] Hang Zhang, Qian Lou, and Yanshan Wang. Towards safe ai clinicians: A comprehensive study on large language model jailbreaking in healthcare. *arXiv preprint arXiv:2501.18632*, 2025.

[338] Hang Zhang, Ruheng Wang, Yuelyu Ji, Mingu Kwak, Xizhi Wu, Chenyu Li, Li Zhang, Wenqi Shi, Yifan Peng, and Yanshan Wang. Scaling medical reasoning verification via tool-integrated reinforcement learning. *arXiv preprint arXiv:2601.20221*, 2026.

[339] J. Zhang. Multi-agent-as-judge: Aligning llm-agent-based automated evaluation with multi-dimensional human evaluation. *arXiv preprint arXiv:2507.21028*, 2025.

[340] K. Zhang. Audited reasoning refinement: Fine-tuning language models via llm-guided step-wise evaluation and correction. *arXiv preprint arXiv:2509.12476*, 2025.

[341] Yilun Zhang and Dexing Kong. Haibu mathematical-medical intelligent agent: Enhancing large language model reliability in medical tasks via verifiable reasoning chains. *arXiv preprint arXiv:2510.07748*, 2025.

[342] Zhenxuan Zhang, Kinhei Lee, Weihang Deng, Huichi Zhou, Zihao Jin, Jiahao Huang, Zhifan Gao, Dominic C Marshall, Yingying Fang, and Guang Yang. Gema-score: Granular explainable multi-agent score for radiology report evaluation. *arXiv preprint arXiv:2503.05347*, 2025.

[343] Guoshenghui Zhao and Eric Song. Privacy-preserving large language models: Mechanisms. *Applications, and Future Directions*, 2024.

[344] Lina Zhao, Jiaxing Bai, Zihao Bian, Qingyue Chen, Yafang Li, Guangbo Li, Min He, Huaiyuan Yao, and Zongjiu Zhang. Autonomous multi-modal llm agents for treatment planning in focused ultrasound ablation surgery. *arXiv preprint arXiv:2505.21418*, 2025.

[345] Weike Zhao, Chaoyi Wu, Yanjie Fan, Xiaoman Zhang, Pengcheng Qiu, Yuze Sun, Xiao Zhou, Yanfeng Wang, Xin Sun, Ya Zhang, et al. An agentic system for rare disease diagnosis with traceable reasoning. *arXiv preprint arXiv:2506.20430*, 2025.

[346] Xinjie Zhao, Moritz Blum, Rui Yang, Boming Yang, Luis Márquez Carpintero, Mónica Pina-Navarro, Tony Wang, Xin Li, Huitao Li, and Yanran Fu. Agentigraph: An interactive knowledge graph platform for llm-based chatbots utilizing private data. *arXiv preprint arXiv:2410.11531*, 2024.

[347] Xinjie Zhao, Moritz Blum, Fan Gao, Yingjian Chen, Boming Yang, Luis Marquez-Carpintero, Mónica Pina-Navarro, Yanran Fu, So Morikawa, Yusuke Iwasawa, et al. Agentigraph: A multi-agent knowledge graph framework for interactive, domain-specific llm chatbots. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 6757–6761, 2025.

[348] Yu Zhao, Hao Guan, Yongcheng Jing, Ying Zhang, and Dacheng Tao. Medcog: Maximizing llm inference density in medical reasoning via meta-cognitive regulation. *arXiv preprint arXiv:2602.07905*, 2026.

[349] Zhendong Zhao, Xiaotian Yue, Jiexin Xie, Chuanhong Fang, Zhenzhou Shao, and Shijie Guo. A dual-agent collaboration framework based on llms for nursing robots to perform bimanual coordination tasks. *IEEE Robotics and Automation Letters*, 2025.

[350] Guangyao Zheng, Michael A Jacobs, Vladimir Braverman, and Vishwa S Parekh. Asynchronous decentralized federated lifelong learning for landmark localization in medical imaging. *arXiv preprint arXiv:2303.06783*, 2023.

[351] Hongjie Zheng, Zesheng Shi, Ping Yi, et al. Medcoact: Confidence-aware multi-agent collaboration for complete clinical decision. *arXiv preprint*, 2025.

[352] Yiming Zheng et al. Mmlu-pro: A more robust benchmark for multi-task language understanding. *arXiv preprint arXiv:2406.01574*, 2024.

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

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

[355] Yucheng Zhou, Lingran Song, and Jianbing Shen. Mam: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. *arXiv preprint arXiv:2506.19835*, 2025.

[356] Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. Medeyes: Learning dynamic visual focus for medical progressive diagnosis. *arXiv preprint arXiv:2511.22018*, 2025.

[357] Jared H. Zhu and Junde Wu. Medicalos: An LLM agent based operating system for digital healthcare. *arXiv preprint arXiv:2509.11507*, 2025.

[358] Wang Bill Zhu, Tianqi Chen, Xinyan Velocity Yu, Ching Ying Lin, Jade Law, Mazen Jizzini, Jorge J. Nieva, Ruishan Liu, and Robin Jia. Cancer-myth: Evaluating large language models on patient questions with false presuppositions, 2025.

[359] Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Shaoting Zhang, and Xiaofan Zhang. Menti: Bridging medical calculator and llm agent with nested tool calling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5097–5116, 2025.

[360] Yinghao Zhu, Ziyi He, Haoran Hu, et al. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. *arXiv preprint arXiv:2505.12371*, 2025.

[361] Jiamin Zhuang. Enhancing medical lung x-ray diagnosis through multi-agent vision-language model collaboration. In *2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, pp. 238–241. IEEE, 2025.

[362] Yangyang Zhuang, Wenjia Jiang, Jia-Yu Zhang, Ze Yang, Joey Tianyi Zhou, and Chi Zhang. Learning to be a doctor: Searching for effective medical agent architectures. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6996–7005, 2025.

[363] Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Liò. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis. *arXiv preprint arXiv:2412.16833*, 2024.

[364] Yuxin Zuo, Shang Qu, Yifei Li, et al. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.