Review

Not Just One Agent: LLM-Based Multi-Agent Systems for Medicine from Answer Generation to Accountable Workflow Orchestration

Tianyi Xiong †, Hanze Guo †, Rui Sheng, Zelin Zang, Xingyin Li, Xingyu Chen, Haoyi Liu, Yue Liu, Xingrui Li, Stan Z. Li *, Yaying Du *, Shaojie Xu *

Posted Date: 10 July 2026

doi: 10.20944/preprints202606.0640.v2

Keywords: large language models; multi-agent systems; clinical medicine; system architecture; AI governance

 doi:10.20944/preprints202606.0640.v2

Review

Not Just One Agent: LLM-Based Multi-Agent Systems for Medicine from Answer Generation to Accountable Workflow Orchestration

Tianyi Xiong [1,†], Hanze Guo [2,†], Rui Sheng [3], Zelin Zang [4], Xingyin Li [5], Xingyu Chen [6], Haoyi Liu [1], Yue Liu [1], Xingrui Li [1], Stan Z. Li [4,*], Yaying Du [1,*] and Shaojie Xu [1,*]

[1] Department of Thyroid and Breast Surgery, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China
[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[3] Hong Kong University of Science and Technology, Hong Kong, China
[4] AI Lab, School of Engineering, Westlake University, Hangzhou, Zhejiang 310024, China
[5] Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen, Guangdong, China
[6] Department of Physics, Faculty of Science, National University of Singapore, Singapore, 117551, Singapore
* Correspondence: author: Stan Z Li, Yaying Du, Shaojie Xu
† These authors contributed equally to this work.

Abstract

Large language models (LLMs) have advanced medical reasoning, but static question-answering performance remains insufficient for clinical workflows that require evolving patient-state tracking, evidence integration, role coordination, and accountable decisions. LLM-based medical multi-agent systems (MAS) are being developed to move AI from isolated answer generation toward workflow-level clinical intelligence by combining role specialization, memory, tool use, retrieval, communication, and orchestration. This Review maps LLM-based medical MAS across diagnosis, treatment decision support, imaging, monitoring, surgery, hospital workflow automation, evidence synthesis, medical education, and safety governance. We further synthesize key architectures for collaboration, knowledge-augmented evidence chains, multimodal integration, privacy-preserving coordination, and adaptive optimization, together with evaluation strategies spanning outcomes, process quality, robustness, efficiency, human comparison, and temporal backtesting. We argue that medical MAS should be evaluated not as larger LLM workflows, but as clinical coordination infrastructures that redistribute evidence, responsibility, and risk across human-AI teams. Their value depends on auditable evidence chains, controllable orchestration, explicit role accountability, and clinician oversight, rather than autonomous answer generation. Before routine clinical use, future work should prioritize traceable evidence chains, human oversight, privacy-preserving collaboration, standardized reporting, regulatory readiness, and prospective clinical validation.

Keywords: large language models; multi-agent systems; clinical medicine; system architecture; AI governance

1. Introduction

Large language models and multi-agent systems

The emergence of large language models (LLMs) has substantially enhanced the capacity for medical text understanding and reasoning, and has produced remarkable performance in static evaluations such as medical examinations and complex case-based question answering [1,2]. However, high scores on examination-style tasks do not necessarily translate into clinically usable systems [3]. In real-world medical workflows, a single model is often constrained by one-shot

 doi:10.20944/preprints202606.0640.v2

reasoning, context congestion, knowledge cutoffs, factual hallucination, and the lack of executable workflows [1,4]. Clinical settings instead require systems that can organize information around an evolving patient state, invoke relevant resources, manage uncertainty, and collaborate with clinical personnel [4] (Figure 1).

Against this backdrop, the application of LLMs in medicine is moving from direct LLM answering toward agents, and further toward multi-agent systems (MAS). Medical agents are commonly defined as computational systems that operate autonomously or semi-autonomously around predefined goals and possess key capabilities such as planning, action, reflection, and memory [4]. Building on this foundation, MAS introduce communication, collaboration, debate, or orchestration among multiple specialized agents, decomposing complex tasks previously handled by a single model into a set of interrelated subtasks that can be iteratively revised [4–6]. This transition is not merely the parallel deployment of multiple models. It reframes LLMs as collaborative systems that use role specialization, communication, and workflow orchestration to decompose complex tasks, integrate multisource evidence, and generate inspectable recommendations. Early evidence suggests that, under clinically scaled mixed workloads, orchestrated MAS may maintain higher accuracy with lower resource consumption than single-agent workflows as task load accumulates [7]. MAS therefore represent not a simple aggregation of LLMs or single agents, but a systems-level design direction that may help adapt medical AI to clinical workflows (Table 1).

Table 1. Comparison of general-purpose LLMs, LLM-based single agents, and LLM-based MAS. Capabilities listed for MAS are design goals demonstrated in selected studies, not guaranteed properties of all MAS.

Feature	General-purpose LLM	LLM-based single agent	LLM-based MAS
Core positioning	General-purpose foundation model	Agent encapsulated around a single goal	System designed as role-based agents with explicit interaction or orchestration
Task execution	One-shot generation or short-chain execution	Can plan and invoke tools, but the same agent undertakes the main subtasks	May decompose complex tasks across different agents and integrate results through orchestration or negotiation when the workflow is explicitly designed and validated
Context management	Mainly single-turn or short-session context	Can maintain task-level context and short-term memory	May maintain shared states, private memories, and cross-role context, depending on implementation and evaluation
External knowledge/tool use	Mainly relies on prompting or attached knowledge bases	Can actively retrieve information and call APIs or tools	Different agents may be assigned to invoke heterogeneous tools, databases, and knowledge sources according to role
Error control	Mainly relies on user review	Limited self-checking; vulnerable to single-point errors	May reduce selected single-point errors only when cross-review, adjudication, and supervisory agents are explicitly validated
Typical medical scenarios	Medical question answering, document generation, single-step explanation	Lightweight decision support and single-process automation	Potential scenarios include MDT-like decision support, complex workflow simulation, dynamic monitoring, and evidence synthesis

 doi:10.20944/preprints202606.0640.v2

Main limitations	Lacks workflow and responsibility structures	Role mixing; easily degrades as workload increases	Complex system design, higher communication costs, more difficult evaluation and governance

Related reviews

Recent reviews have established the broader landscape of LLM-based agents and multi-agent systems, covering agent roles, memory, planning, tool use, communication, collaboration patterns, and evaluation frameworks [8–10]. In medicine, related reviews and perspectives have discussed agentic AI for clinical decision support, healthcare workflows, emergency medicine, precision oncology, biomedical data analysis, radiology ethics, perioperative care, and healthcare governance [11–19]. These articles show that agentic AI is becoming an important direction in medical AI.

However, the existing literature remains fragmented. General AI reviews rarely address clinical workflow constraints, whereas medical reviews usually focus on single specialties, single-agent systems, ethics, or adjacent biomedical applications. As a result, there is still no medicine-wide review that defines LLM-based medical MAS as a distinct clinical system architecture and maps their applications, architectures, validation strength, and evidence maturity across medical domains. This Review addresses that gap through a structured narrative evidence map of LLM-based medical MAS, emphasizing their role as clinical coordination systems rather than merely larger LLM workflows.

Evidence mapping

We searched PubMed, Web of Science, Google Scholar, medRxiv, and arXiv using combinations of terms related to large language models, multi-agent systems, agentic AI, medicine, healthcare, clinical workflows, diagnosis, treatment, and medical AI for studies published or posted from 6 October 2022 through 6 July 2026. We used 6 October 2022, the first arXiv submission date of ReAct, as a reproducible methodological anchor for the modern LLM-agent paradigm [20].

Studies were counted as core evidence if they described a medical or biomedical LLM-based MAS with at least two role-specialized agents, explicit interaction between agents, and at least one extractable evaluation outcome. Single-agent systems, conventional ensembles, non-LLM agent-based simulations, sensor-network MAS, multi-robot systems, and papers using “agent” only as a generic software term were not counted as core evidence.

Two reviewers screened titles, abstracts, and available full texts, with disagreements resolved by discussion. The final reference set included 147 records, of which 81 met the core-evidence criteria and were included in the Supplementary Evidence Table. For each core record, we extracted publication status, application domain, architecture, agent roles, data source, validation type, sample size, comparator, primary endpoint, evidence maturity, and key limitations.

Because the field is evolving rapidly, directly relevant preprints and conference papers were included but were used mainly to map architectures, applications, and evaluation designs, not to support strong claims about clinical effectiveness or deployment readiness. Among the 81 core records, 31 were peer-reviewed journal articles and 49 were preprints or conference/workshop papers. Most evaluations were benchmark-only, simulation-based, or retrospective, and no core study reported prospective clinical validation. These findings support our interpretation of LLM-based medical MAS as early-stage workflow technologies rather than systems ready for routine clinical deployment.

 doi:10.20944/preprints202606.0640.v2

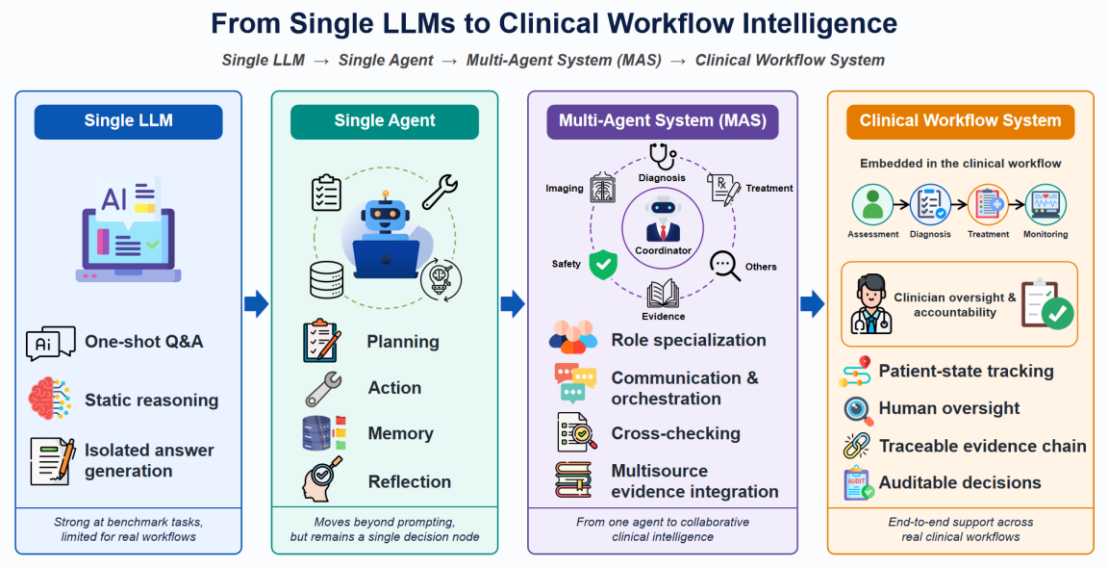

Figure 1. From Single LLMs to Clinical Workflow Intelligence.

Medical AI is progressing from isolated one-shot answer generation to agentic systems that can decompose clinical tasks, use tools, update patient context, and iteratively check outputs. Multi-agent systems further extend this transition through role specialization, communication, orchestration, cross-checking, and multisource evidence integration. When embedded into clinical workflows, these systems support patient-state tracking, clinician oversight, traceable evidence chains, and auditable decisions, shifting medical AI toward accountable workflow-level intelligence.

2. Potential Application Scenarios of Multi-Agent Systems in Clinical Medicine

2.1. Clinical Applications in Medicine

LLM-based multi-agent systems (MAS) are being explored beyond the capability boundaries of single agents, moving from early medical "chatbots" toward applications that support multi-role collaboration and task decomposition across clinical scenarios. The core feature of clinical care is multi-role collaboration: primary physicians perform initial assessment, radiology and laboratory professionals provide data support, subspecialists establish diagnoses and treatment plans, and nursing teams and pharmacists jointly implement treatment management and follow-up. Medical care thus depends heavily on the participation of different professional roles to reduce diagnostic and therapeutic uncertainty and improve the credibility of clinical decisions (Figure 3A). Unlike a single LLM agent that usually processes tasks through one dominant reasoning pathway, MAS coordinate multiple agents with dedicated roles and decompose complex medical tasks into subprocesses that aim to align with clinical workflows. In this way, MAS have been studied in clinical scenarios including disease diagnosis, treatment, imaging analysis, intraoperative assistance, and monitoring (Figure 3B).

2.1.1. Diagnosis and Differential Diagnosis

In diagnostic settings, the clinical value of MAS is often reflected in a more realistic pattern of progressive information acquisition. Several studies have used real or quasi-real clinical dialogues to emphasize that diagnosis is not a one-step answer, but a process in which conclusions converge through multiple rounds of communication, information supplementation, and dynamic revision. Through specialized role assignment and collaboration, selected MAS studies have explored iterative diagnostic optimization during multi-turn dialogues. In selected evaluation settings, this workflow has shown potential advantages over single-agent systems and human comparators in diagnostic accuracy or resource-use efficiency [21,22]. In a small set of 16 neurological diagnostic cases, the

 doi:10.20944/preprints202606.0640.v2

Gregory system reported 100% diagnostic accuracy, compared with an average of 83% for human neurologists, while reducing diagnostic cost and time [21]. A related multi-agent system from the same group, organized around cognitive-function agents (complexity assessment, interpretation, retrieval, synthesis, and validation), was evaluated separately on neurology board-style questions and MedQA neurology cases [23]. MedAgentSim, a multi-agent system involving patient, doctor, and measurement agents, reported 79.5% diagnostic accuracy on a real-world clinical dataset, higher than a single-agent system [22]. In a layered-debating MAS, multiple domain-expert agents were configured for multilayered debate, with the aim of identifying subtle disease differences; any gain in diagnostic precision should be interpreted within the benchmark setting [24]. Recent specialty-oriented systems have extended this MDT-style design to psychiatric and rheumatologic diagnosis. WiseMind paired a knowledge-guided dual-agent design with a DSM-5 knowledge graph and was evaluated on 1,206 simulated and 180 real psychiatric conversations [25]. For axial spondyloarthritis in primary care, SpAgents combined planner, data, tool, and physician agents with long-term memory and was assessed across a multicenter cohort of 596 cases, one of the larger real-world evaluations to date [26]. Beyond these, multi-agent diagnostic systems have been explored across further specialties and modalities, including multimodal Parkinson's disease diagnosis [27], multimodal Alzheimer's disease diagnosis with a self-learning critic agent [28], orchestrated primary-care diagnosis of secondary headache [29], evidence-based diagnosis of vascular anomalies with decoupled vision and reasoning agents [30], and dynamically reconfigurable diagnostic teams that adjust their composition during multi-turn interaction [31].

For rare diseases and complex cases, MAS are intended to supplement clinicians by broadening the coverage of clinical clues and strengthening the diagnostic evidence chain [32]. In rare genetic diseases, multi-agent systems have been explored for quantitatively evaluating and ranking the pathogenic likelihood of candidate genes, with the aim of prioritizing candidate disease-causing genes and reducing manual screening burden [33,34]. In the face of common challenges in rare genetic disease diagnosis—including complex candidate variants, substantial phenotypic overlap, and a persistently low overall molecular diagnostic rate of only 30%–40%—the MD2GPS multi-agent system integrates pathogenic gene prioritization with multidimensional medical knowledge reasoning through specialized roles, including data-processing agents, knowledge-reasoning agents, and verification/debate agents. This design aims to reduce the knowledge limitations of a single agent and improve pathogenic gene ranking [34]. Rare diseases involve vast candidate spaces and limited clinical experience, and therefore require stronger evidence integration and interpretability. DeepRare, a rare-disease diagnostic MAS, retrospectively processed free-text histories, phenotypic terms, and sequencing files to output candidate diseases together with evidence chains [35]. A further rare-disease system, RareCollab, integrated quantitative diagnostic reasoning with multiple specialized modules over phenotypic and molecular evidence, evaluated on Undiagnosed Diseases Network cohorts [36]. More recent studies increasingly use the clinical multidisciplinary team (MDT) as a framework for constructing medical MAS (Figure 2): a management or triage agent first summarizes case information and assigns diagnostic tasks [32,37], after which multiple specialty agents conduct collaborative diagnosis. At key points of high diagnostic uncertainty or disagreement over management, attending-physician agents or specialty-physician agents are introduced to evaluate the diagnosis and integrate consensus [38–40], with the goal of reducing information omission and cognitive bias in complex or atypical cases.

 doi:10.20944/preprints202606.0640.v2

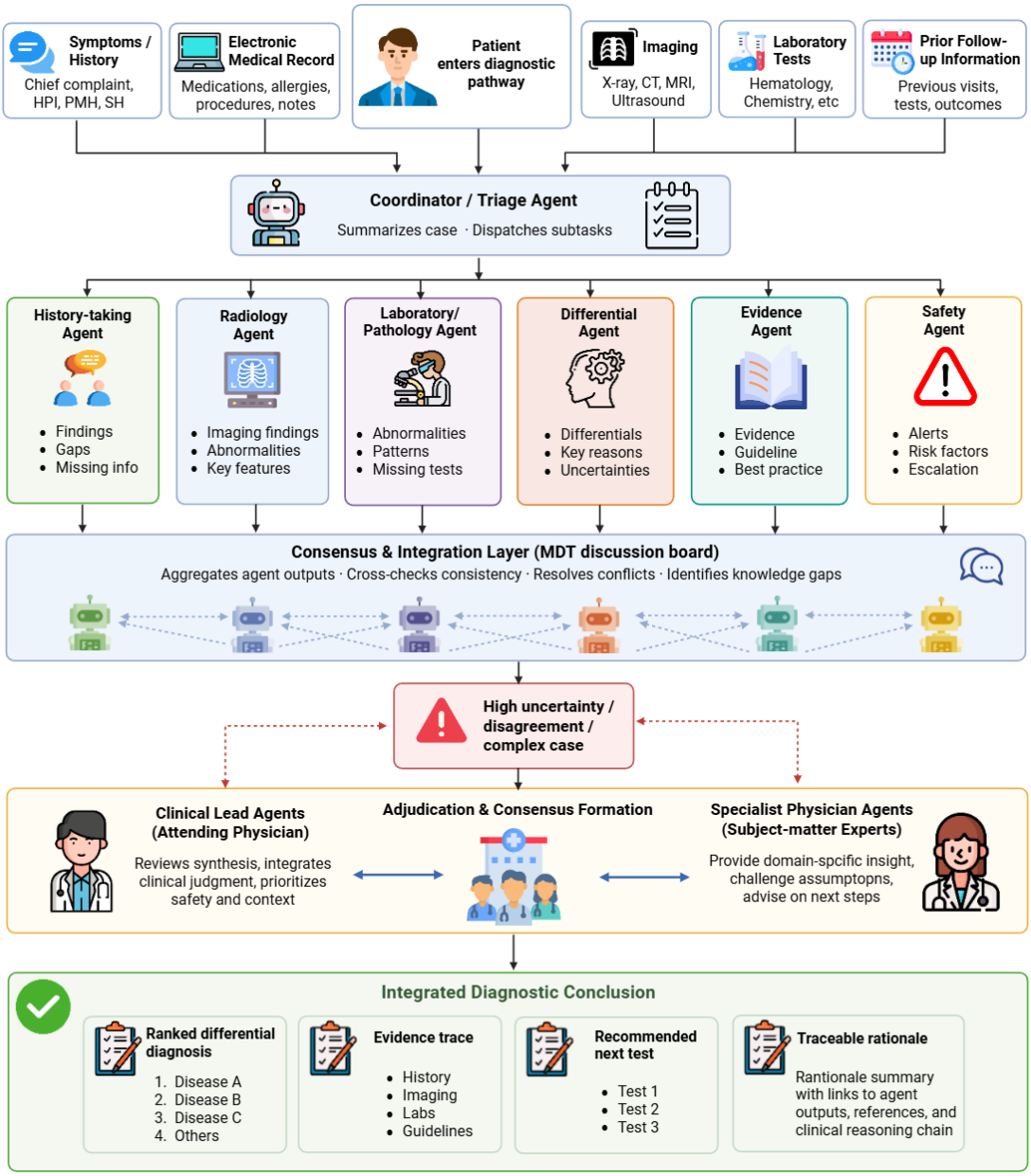

Figure 2. MDT-style multi-agent systems for diagnosis and differential diagnosis.

Patient information, including symptoms, medical history, electronic health records, imaging, laboratory tests, and follow-up data, is first summarized by a coordinator/triage agent and distributed to role-specialized agents. History-taking, radiology, laboratory/pathology, differential-diagnosis, evidence, and safety agents analyze complementary information, identify missing data or risks, and generate intermediate diagnostic outputs. These outputs are integrated through an MDT-like consensus layer that cross-checks consistency, resolves conflicts, and identifies knowledge gaps. Cases with high uncertainty, disagreement, or complexity are escalated to clinical lead and specialist physician agents for adjudication. The system ultimately produces an integrated diagnostic conclusion, including ranked differential diagnoses, evidence traces, recommended next tests, and a traceable rationale.

2.1.2. Treatment and Decision Support

MAS-driven treatment and decision support are emerging as a candidate approach for individualized precision therapy. Their advantage lies in moving beyond the limitations of the traditional static input-output pattern of LLMs and compensating for the insufficiency of single

 doi:10.20944/preprints202606.0640.v2

agents in complex clinical decision-making through multi-role collaboration, planning, and tool use [41]. Through specialized role assignment and collaborative reasoning, MAS are designed to integrate patients' multimodal clinical information, including electronic health records, imaging data, laboratory tests, and follow-up feedback, and align individual clinical characteristics with guideline recommendations and evidence-based findings. This architecture may support the generation of updated individualized treatment plans across clinical scenarios [41,42]. In a sepsis-oriented MAS, clinical tasks were divided across antibiotic recommendation, guideline checking, and full-process sepsis management; dedicated agents were assigned to regimen formulation, evidence-based validation, global condition assessment, treatment planning, and monitoring of clinical deterioration [43]. With retrieval-augmented generation, this type of system has been designed to integrate patient-specific clinical data and guideline evidence, which may shorten treatment decision time in high-risk settings such as the ICU [43]. A comparative validation study by Chen et al. [44] in ICU critical-care scenarios reported higher accuracy than a single-agent system for in-hospital mortality prediction ($p = 0.0001$) and ICU length-of-stay prediction ($p < 0.0001$), suggesting task-specific gains in critical-care outcome prediction. In oncology, the MAS constructed by Ayub et al. [45] used role-based collaboration for risk stratification and treatment recommendation, converting key stratification evidence from unstructured reports into executable treatment-pathway recommendations. In chronic diseases requiring long-term dynamic management [41], MAS may use external tool calling and retrieval augmentation to collect patient clinical data and evidence-based literature, extract clinical knowledge in structured form, and support dynamic adjustment of long-term treatment plans. Meanwhile, specialized MAS for full-cycle chronic disease management incorporate remote monitoring modules [41], connect to existing medical systems through standardized interfaces, and support continuous treatment and follow-up adjustment, thereby facilitating closed-loop management across the chronic disease life cycle. In multimorbidity management [46], MAS may link disease-specific clinical guidelines and help identify potential drug interactions or treatment contraindications arising from conflicting recommendations across guidelines, a proposed use case in multimorbidity management. In a retrospective order-set study, a MAS identified outdated, inappropriate, or missing items in order sets and generated candidate improvements to support physician review [47]. By distributing treatment decision subtasks to different agents, such as treatment recommendation, guideline checking, and follow-up monitoring, MAS may produce therapeutic recommendations that better approximate clinical team workflows and accountability structures. In oncology decision-making, the self-evolving system EvoMDT performed lesion-level reasoning and used a consensus protocol to resolve inter-agent disagreement, evaluated on several oncology question-answering benchmarks together with real tumor datasets [48]. This context supports the inclusion of oncology MAS examples, but also illustrates why a medicine-wide synthesis is still needed beyond cancer-focused perspectives. Multi-agent systems have also been applied to high-risk medication settings, such as assessing drug-induced liver injury in the intensive care unit [49].

2.1.3. Medical Imaging

MAS have been studied in medical imaging across X-ray imaging [50–53], CT [52,53], MRI [40], ultrasonography [54], and pathological slide review [55–57], with the goal of improving diagnostic efficiency, accuracy, and interpretability. In automated radiology report generation for imaging modalities such as chest X-rays, the MAS developed by Yi et al. [58] and Li et al. [59] decomposed the clinical workflow of report generation into key steps: retrieval of similar case reports, drafting of preliminary reports, extraction of key clinical findings, interpretation of visual imaging evidence, final report synthesis, and evidence-based quality control. Each step is performed independently by a specialized agent with a dedicated function, while the agents collaborate through interaction. Alam et al. [50] further combined a concept bottleneck model with a multi-agent retrieval-augmented generation (RAG) architecture, enabling key visual representations in chest radiographs to be translated into interpretable clinical concepts and their contribution weights, and to generate more

 doi:10.20944/preprints202606.0640.v2

clinically relevant imaging reports accordingly. This modular multi-role design is intended to anchor diagnostic conclusions to visual evidence, reduce generative hallucination through multistep factual verification and source tracing, and present a reasoning path from imaging features to diagnostic conclusions. In ultrasound imaging, the FetalAgents MAS uses collaborative role specialization to automate eight clinical tasks, including ultrasound-plane classification, structural segmentation, and biometric measurement, while extending analysis toward ultrasound video streams and standardized clinical report generation [54]. Automatically generated radiology reports must be evaluated across clinical accuracy, information completeness, and language standardization. GEMA-Score explored MAS-based quality control for report generation by transforming judgments of clinical usability into quantifiable scores [52]. For diagnostic efficiency, imaging MAS may help identify key image features and next-step management options around specific clinical questions, potentially reducing repeated communication and delayed decisions [51]. They may also help triage imaging workloads in primary care or resource-constrained settings and support decisions regarding further testing [53]. In pathological slide review, MAS collaboratively perform triage, navigation, description, and diagnosis across the full workflow [55,56]. Through autonomous navigation and feature extraction from pathological slides, they localize key diagnostic regions and bind textual conclusions with visual localization outputs, facilitating rapid physician verification and documentation [57]. Multi-agent designs have further been explored beyond image interpretation itself, for example to monitor and correct performance drift of deployed medical-imaging models, where a coordinating agent supervises several task agents over the model-maintenance workflow [60]. For radiotherapy planning, multi-agent simulation frameworks such as MARTP have organized several specialized agents to reproduce multidisciplinary planning workflows [61].

2.1.4. Patient Monitoring and Care Management

MAS-driven patient monitoring and care management are shifting from alerts based on a single time point or a single indicator to continuous and contextualized dynamic risk monitoring. The core capability lies in parallel parsing, verification, and integration of multisource data, including real-time vital signs, laboratory results, medication information, and progress notes, with the goal of identifying acute clinical deterioration earlier and supporting timely intervention [44,62]. In a retrospective Alzheimer's disease prediction study, a MAS combined unstructured longitudinal EHR text with collaborative specialty-agent evaluations to capture prodromal signals, suggesting possible earlier risk prediction [63]. In post-discharge follow-up and chronic disease management, MAS may support continuous monitoring across the care cycle by incorporating remote monitoring data, post-treatment response assessments, patient-reported symptoms, and medication behaviors into analysis [64]. When combined with multimodal information from electronic health records, imaging, and laboratory tests, MAS could support dynamic risk assessment, earlier warning, therapeutic adjustment, and follow-up [41]. For chronic diseases such as type 1 diabetes, in which low willingness to adopt technology and insufficient treatment adherence remain clinical barriers, the MAS developed by Yao et al. simulates evidence-based doctor-patient conversations to support personalized education, psychological support, and behavioral intervention [65]. Beyond simulated dialogue, a multi-agent patient-education system has been explored for diabetes education and limb preservation: specialized agents transcribed clinician–patient conversations into structured notes and generated conversational educational content, with medical experts evaluating accuracy, completeness, and absence of hallucination under simulated consultations rather than real patient outcomes [66]. Extending to home and community settings, patient monitoring and management often shift toward ensuring treatment adherence and responding rapidly to acute medical events. MAS integrate vital-sign monitoring, medication reminders, and emergency response to enable continuous monitoring and risk warning in home environments [64,67]. Through hospital-to-home data integration and dynamic risk assessment, MAS could support proactive patient monitoring models that move beyond passive alerts toward clinician-supervised closed-loop care. For risk prediction, mixture-of-multimodal-agents designs such as MoMA assigned specialist, aggregator,

 doi:10.20944/preprints202606.0640.v2

and predictor roles across modalities for clinical outcome prediction [68], and note-based multi-agent systems have been used to predict heart-failure 30-day readmission from clinical notes [69].

2.1.5. Surgical Assistance and Surgical Robotics

The application of MAS in intraoperative assistance and surgical robotics is being explored as a clinical intelligence framework for surgical team collaboration. Typical application scenarios include intraoperative localization and navigation with coordinated instrument control [70] and surgical plan formulation with perioperative safety management [71]. Supported by real-time information exchange, task coordination, and closed-loop feedback optimization among agents, this technical framework has the potential to reduce navigation errors, instrument-coordination mismatch, and inconsistencies in treatment-plan execution. It may thereby improve the precision, stability, and reproducibility of selected minimally invasive and complex surgical procedures, although evidence for reductions in perioperative adverse events remains limited [70,71]. In interventional surgery, Chen et al. developed a voice-control system for MRI scanners based on multi-agent collaboration. Through division of labor among agents, the system completed the full workflow of speech correction, task parsing, document retrieval, and device control, enabling contactless real-time control of intraoperative imaging equipment under sterile conditions. This addressed delays and communication errors caused by traditional reliance on assistants to relay instructions and provided a feasible solution for full-process intelligence in MRI-guided interventions [72]. In embodied multi-agent research, the transformation of dual-arm nursing robots into a multi-agent system defines the left and right robotic arms as independent agents. Through hierarchical task decomposition and collaborative execution, this approach addresses core limitations of traditional single-agent architectures, including conflicts in bimanual task planning and insufficient movement coordination. It may have application value for high-frequency scenarios in operating rooms and wards, such as instrument and material sorting, delivery, and human-robot collaborative operations [73,74]. Perioperative care has also been proposed as a multi-agent setting spanning risk assessment, decision support, and longitudinal monitoring; at present, such work should be cited mainly as application framing rather than mature clinical evidence [19]. The deeper application of MAS could support operational precision and perioperative safety, but this claim will require prospective validation and explicit human-control safeguards.

2.2. Supporting Applications in Medicine

Compared with core clinical applications that directly participate in diagnosis, treatment, and monitoring, the supporting applications of medical MAS focus on improving the precision and operational efficiency of healthcare institutions and clinical support processes, while reducing the cognitive burden and operational cost of clinical and research work. The implementation of these supporting scenarios does not rely on a single general-purpose agent to perform full-process operational support. Instead, it depends on multiple independent agents with dedicated functional roles to achieve role-based division of labor and distributed scheduling, jointly optimizing the turnover efficiency of key processes such as outpatient visits and examinations and reducing human operational costs. Such systems may integrate multisource evidence-based information and clinical data to provide data support for refined hospital management and clinical research. In addition, supporting applications of medical MAS may extend to medical education and clinical safety governance, further strengthening the foundational infrastructure of healthcare-system operations by promoting standardized clinical training and reinforcing diagnostic and therapeutic safety control (Figure 3C).

 doi:10.20944/preprints202606.0640.v2

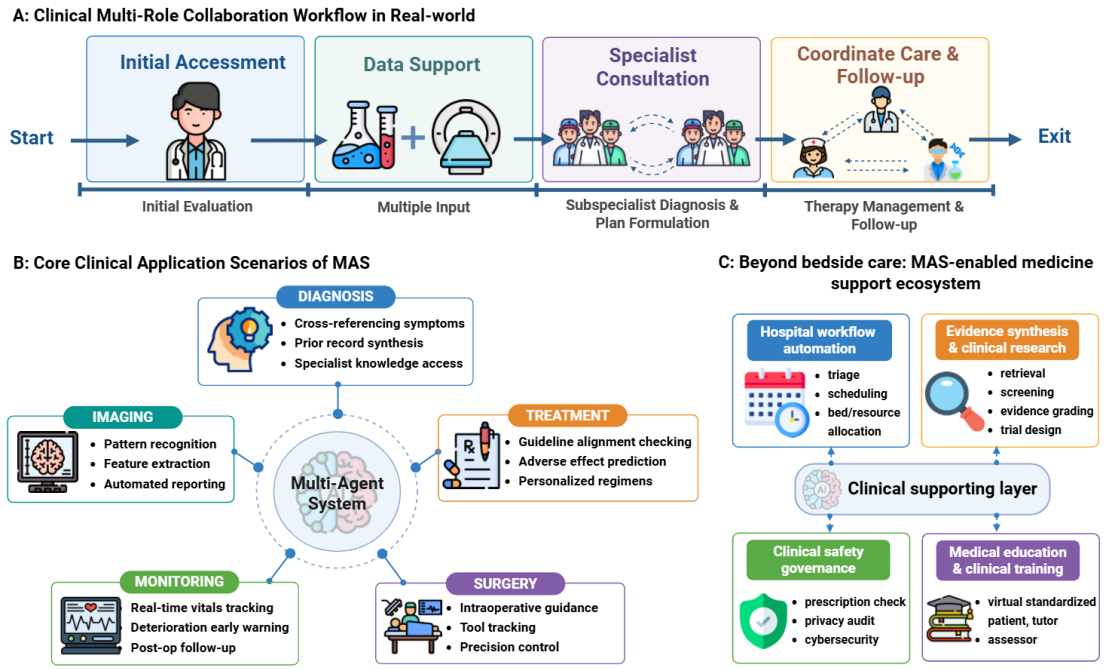

Figure 3. Major application scenarios of medical multi-agent systems. (A) MAS simulate real-world clinical teamwork, spanning initial assessment, multimodal data support, specialist consultation, and coordinated treatment follow-up. **(B)** In core clinical care, MAS support diagnosis, treatment, imaging, surgery, and patient monitoring through role-specialized collaboration around a shared patient state. **(C)** Beyond bedside care, MAS also support hospital workflow automation, evidence synthesis and clinical research, medical education, and clinical safety governance, extending multi-agent collaboration to the broader healthcare ecosystem. Taken together, these application areas should not be read as independent proof that MAS is clinically superior across medicine. Rather, they show where multi-agent coordination may become useful when clinical tasks require shared patient states, distributed evidence review, role-specific reasoning, and escalation across human-AI workflows. The more important question is therefore not simply where MAS can be applied, but which architectures, evaluation standards, and governance mechanisms make such coordination reliable enough for clinical use.

2.2.1. Hospital Workflow Automation

The core value of MAS in hospital workflow automation lies in improving the quality and efficiency of medical processes in real-world healthcare settings while reducing human error and operational risk. Existing studies commonly begin with departmental resource allocation, clinician scheduling, outpatient appointment management, and prehospital pre-triage, incorporating multidimensional core parameters such as clinicians' practice preferences, medical resource allocation, and patient priority into collaborative decision-making frameworks to enable dynamic optimization of healthcare workflows and comprehensive improvement of institutional operational efficiency [75]. In triage studies, MAS used role-based collaboration to integrate standardized triage scales, guideline retrieval, and multi-turn interactive decision-making; reported gains in grading efficiency, accuracy, or consistency should be interpreted within the benchmark or retrospective settings [76,77]. TRIAGEAGENT [76] reported improved bias control and grading precision in clinical triage through group classification, confidence aggregation, and consensus decision-making. Han et al.'s [77] multi-agent emergency decision-support system further showed that a collaborative framework simulating the role division of an emergency team not only may help generate clearer and more consistent KTAS grading results, but also improves the stability of management decisions, medication management, and resource allocation. On the patient-service side, MAS may use modular collaboration to standardize and integrate prehospital information, with the intended goal of improving continuity between prehospital and in-hospital care processes [64]. The implementation

 doi:10.20944/preprints202606.0640.v2

of clinical workflow automation must be grounded in full-process quality and safety control. Built-in evaluation and supervision agents may monitor system outputs across the workflow and are intended to reduce the risk that propagated errors or clinically infeasible recommendations enter downstream use [78,79]. Although MIRA is not a typical multi-agent system, its sandboxed EHR design provides a comparator showing how an agentic clinical system was evaluated for structured clinical actions, including test ordering, medication prescribing, procedure scheduling, and admission planning [80]. This provides a useful comparator for MAS-based workflow automation, because it highlights the importance of permission boundaries, standardized EHR interfaces, action logs, and physician-level evaluation before deployment. MAS may provide a workflow-support model for hospital workflow automation, with the potential to improve institutional operational quality and efficiency while supporting a full-cycle safety risk-control framework for clinical processes. Multi-agent designs have also been explored for emergency medical service dispatch, simulating taxonomy-driven coordination among service roles [81].

2.2.2. Evidence Synthesis and Clinical Research

Evidence-based medicine and clinical research are core upstream components of guideline revision and individualized treatment-plan development. The bottleneck in this process is not text generation itself, but the substantial labor and time costs required for large-scale literature retrieval, heterogeneous data screening, structured information extraction, statistical analysis, and evidence-quality grading, as well as the difficulty of producing standardized evidence outputs that can be directly reused by clinicians and researchers [82,83]. Through role-based division of labor among agents, MAS may automate parts of the evidence-synthesis workflow, including literature retrieval, evidence collection, data screening, information summarization, and citation/source verification. Multi-round consistency checks and result validation may improve output stability, but their ability to generate reliable systematic reviews, guideline outputs, and clinical question answering requires careful human verification [84,85]. In research contexts requiring high levels of evidence, MAS may include independent quality-assessment and supervision agents designed to follow evidence-grading standards and improve citation traceability, but these outputs still require human verification to detect fabricated or low-quality citations [86]. In key research scenarios such as clinical trial design and analysis, MAS have begun to show clear application potential. EmulatRx [87] uses five specialized agents for regulation, trial design, informatics, clinical expertise, and statistics to automatically support key steps such as protocol generation, real-world data mapping, target trial simulation, and statistical analysis, with the aim of improving the efficiency and iteration capacity of trial design. ClinicalAgent [88] uses specialized agents for planning, pharmacodynamic assessment, safety analysis, and enrollment prediction to predict clinical trial outcomes, analyze reasons for trial failure, and assess enrollment difficulty. It reported higher performance than direct prompt-based LLM methods in trial outcome prediction and may provide workflow support for evidence-based decision-making across the clinical trial life cycle. In clinical genomics and precision medicine, MAS have been proposed to integrate multisource evidence such as genetic variants and pharmacogenomics, perform rule-based evidence scoring and interpretation, and support individualized diagnosis and treatment planning [89,90]. In time-sensitive emergency and critical-care scenarios such as sepsis, MAS may align patient clinical data with treatment guidelines and research literature to generate candidate guideline-concordant decision recommendations [43]. At the same time, automated systems based on multi-agent architecture have been proposed for clinical RCT literature retrieval, PICO element extraction, and network meta-analysis statistical modeling [91]. Such systems should be treated as workflow-support tools until their screening accuracy, extraction reliability, and statistical outputs are independently validated. Related multi-agent workflows have addressed upstream information-processing steps, including ontology-enhanced zero-shot clinical named-entity recognition [92] and medical-text summarization with automated evaluation [93].

 doi:10.20944/preprints202606.0640.v2

2.2.3. Medical Education and Clinical Training

The core goal of medical education and clinical training is to provide standardized and structured formative feedback on learners' clinical reasoning, doctor-patient communication, and teamwork abilities within highly realistic and dynamically changing clinical scenarios, thereby improving clinical thinking and comprehensive diagnostic and therapeutic competence. The key advantage of MAS in this field is their ability to transform the classic clinical teaching architecture of standardized patient-medical student/resident-mentor/examiner into interactive and immersive training environments. This shifts traditional teaching away from static question banks and single-turn question answering toward dynamic training that follows clinical logic, including multi-turn history taking, decisions about ancillary tests, disclosure of clinical information, and diagnostic and therapeutic reflection [94]. Such systems may simulate aspects of clinical training environments and have been explored for improving learners' history-taking, communication skills, MDT collaboration, and integrated diagnostic competence. In a medical-education study, a MAS transformed static knowledge-point question banks into dynamic virtual clinical conversations for diagnostic training and standardized assessment [95], allowing learners to complete more clinically realistic diagnostic training and standardized assessment through real-time interaction with virtual standardized patients. A further multi-agent framework emulated standardized patients specifically for clinical communication-skills training [96]. This approach may be adapted to specialty residency training and stage-based competency assessment, pending validation in the relevant training setting. The MAARTA system developed by Awasthi et al. compared radiology trainees' eye-tracking patterns, report text, and expert diagnostic standards to identify cognitive blind spots and diagnostic biases and generate personalized teaching guidance. In diagnostic error classification, it reported higher accuracy and F1 score compared with a single agent [97]. Beyond core scenario-based clinical training, MAS may also support teaching-management tasks such as quality evaluation of question banks and assessment of teaching-feedback effectiveness [98,99]. At the same time, stable deployment in educational settings requires that safety boundaries, standardized feedback, and diagnostic/therapeutic debriefing be incorporated into the core training and evaluation process [100], so that the system meets clinical training requirements for realism and standardization.

2.2.4. Clinical Safety Governance

When medical MAS are deployed in real clinical settings, they require a safety-control system covering multiple core dimensions, including diagnostic and therapeutic decision risk, patient data privacy, and medical cybersecurity. MAS may help reduce the risk of clinical medication errors through a standardized closed loop of recommendation generation, indication verification, contraindication and interaction screening, and clinical review-summary output [101]. In an order-set optimization context, MAS proceduralized order-compliance review, evidence-based knowledge retrieval, and drug-indication verification to flag outdated or incorrect orders for clinician review [47]. A medication-counseling prototype further illustrates how structured peer review and second-confirmation mechanisms can be designed to address context drift, omitted information, and unsafe counseling trajectories [102]. Patient data privacy protection and medical cybersecurity control are core prerequisites for compliant clinical deployment of MAS, with key focuses on minimizing the exposure of sensitive data, clarifying access-control permissions, and reducing network dependence [64]. For cybersecurity protection of medical devices, MAS may link device hardware and software component information with known security vulnerabilities, use multisource knowledge integration and logical reasoning to generate structured cyber-threat models and vulnerability pathways, identify potential medical-device security vulnerabilities, and output candidate defense recommendations [103], with the goal of supporting cybersecurity and stable operation of in-hospital devices.

3. Multi-Agent System Architectures for Clinical Workflows

 doi:10.20944/preprints202606.0640.v2

Compared with single agents, the main value of MAS does not lie in increasing the number of models, but in translating clinical role division, information handoff, evidence verification, and accountability records into executable and auditable workflows. At the same time, increasing the number of nodes also introduces communication overhead, error propagation, and greater governance complexity. Therefore, the key issue for MAS in clinical settings is not whether the collaboration structure is more complex, but whether it remains aligned with real medical workflows and maintains stability and controllability under high-risk conditions.

3.1. Clinical Collaboration and Communication Design

The advantage of MAS in clinical collaborative communication first lies in distributing case-state maintenance, information dissemination, and result integration across different roles, rather than requiring a single model to assume all responsibilities within the same context. Conversational diagnostic systems in which a coordinating node maintains case state and controls information flow have shown that this organizational form more closely resembles the process of information handoff in real clinical practice [21]. Inpatient pathway support has been organized as a continuous collaboration chain [104]. SOAP-note clinical problem detection uses a collaborative format in which a manager organizes expert discussion [105]. Order-set optimization further assigns content critique, dynamic retrieval, knowledge retrieval, and medication checking to different nodes [47]. This evolution is not simple modular stacking; rather, it maps the communication mechanisms and responsibility boundaries of clinical teams into the system's internal interaction structure.

However, for MAS, more complex communication mechanisms are not necessarily better. Open natural-language interaction is highly flexible but more prone to semantic drift and task deviation. Highly structured protocols can reduce ambiguity, but may weaken the capture of weak signals and ambiguous expressions. Clinical documents, handoff notes, and consultation opinions have long used relatively standardized formats not merely as formal requirements, but because high-risk environments require traceability and clear responsibility. Accordingly, structured interaction, hierarchical handoff, and sequential orchestration in MAS should be understood as technical translations of clinical communication norms, rather than neutral engineering details [105].

Furthermore, different tasks require different depths of collaboration. For rule-based, lower-risk subtasks, lightweight routing and specialized execution nodes are often sufficient [7]. A pharmacy consultation prototype illustrates how multi-role peer review could be used for self-assessment and safety checks without substantially expanding the workflow [102]. By contrast, complex cases with high uncertainty and high consequences are better suited to multi-role checking and escalation mechanisms [106]. Tree-of-Reasoning further indicates that when diagnosis depends on the integration of evidence across sources, communication is not merely information transfer but evidence organization itself [40]. Methodologically, multi-agent debate and peer-review-style cross-checking have also been proposed to refine medical reasoning by surfacing consensus and disagreement among agents [107,108]. Thus, the core of clinical collaboration and communication design is not to increase the number of interaction rounds or roles, but to match task complexity, risk level, and collaboration depth appropriately.

3.2. Knowledge Augmentation and External Memory Architectures for Clinical Evidence Chains

The advantage of MAS in knowledge augmentation is not simply an increase in the number of retrieved results, but the ability to assign different evidence sources to different role nodes while preserving source boundaries during aggregation. Public literature and external medical knowledge can provide explicit evidence support for multi-agent diagnosis and can be further organized into verifiable evidence chains in complex cases [40]. In-hospital rules and local knowledge bases determine whether recommendations are executable [47]. A pharmacy consultation prototype suggests that drug labels, contraindications, and dialogue norms can be assigned to dedicated nodes [102]. The combination of structured EHRs and medical knowledge retrieval allows patient state and general medical evidence to be used distinctly within the same workflow [62]. Longitudinal clinical

 doi:10.20944/preprints202606.0640.v2

notes further add patient-specific information along the temporal dimension [63]. If a system cannot distinguish the hierarchy, timeliness, and applicability boundaries of these evidence sources, even strong retrieval capability may still cause general medical knowledge to be incorrectly expressed as patient-specific management recommendations.

This is a key difference between MAS and single agents: knowledge is no longer injected into the model as a one-off retrieval result, but is used by different agents around different databases, tools, and responsibilities. Evidence-tree-based frameworks for complex diagnosis emphasize explicit links between conclusions and supporting evidence. Their goal is not merely to improve answer accuracy, but to enhance the verifiability of the reasoning path [40]. Collaborative EHR-oriented frameworks attempt to integrate structured EHRs, external knowledge, and multi-role reasoning into a single workflow to improve interpretability and clinical relevance [62]. Order-set optimization studies show that even when the latest literature and guideline information are obtained, medication validation and real-world feasibility checks remain necessary before an operational clinical recommendation can be formed [47]. Thus, the real question addressed by knowledge augmentation is not whether evidence can be obtained, but whether it can be correctly integrated and used within multi-role collaboration.

A critical role of external memory in clinical settings is to preserve the temporal evolution of patient states and intermediate task results. Studies of Alzheimer's disease prediction driven by longitudinal clinical notes show that many meaningful risk signals do not appear at a single time point, but are distributed across long-term trajectories and become interpretable only after longitudinal integration [63]. Memory modules in treatment planning similarly indicate that if a system cannot continuously preserve previous strategies and intermediate results, each iteration may regress to repeatedly restarting the same case [71]. For MAS, such memory is not merely an extension of a single cache; it also involves coordination between shared state and role-specific working memory. Different agents read different information, retain different intermediate results, and subsequently hand them off to downstream nodes.

It should be emphasized that knowledge augmentation and external memory do not alter the basic generative mechanism of LLMs. As probabilistic text-generation models, LLMs may still generate fluent but insufficiently grounded explanations when faced with conflicting evidence, inconsistent sources, or noisy inputs. Hallucination, post hoc rationalization, and temporal drift are not fundamentally eliminated by introducing RAG, external memory, or knowledge graphs. In MAS, these problems can also be transmitted across multiple nodes and further amplified during downstream integration. Therefore, these mechanisms improve the lower bound of system robustness in complex tasks, rather than fundamentally eliminating sources of uncertainty.

3.3. Multimodal Information Integration

Multimodal information in clinical care includes not only imaging and text, but also laboratory indicators, pathological results, structured EHRs, progress notes, and treatment parameters. Multimodal information integration is therefore almost unavoidable for hospital-level MAS. The advantage of MAS in this regard is not simply to input different types of information into a single model, but to allow different roles to separately process structured records, external knowledge, and clinical reasoning tasks before integrating them at the patient level [62]. This organization more closely resembles cross-departmental collaboration in real hospitals and better preserves intermediate judgments, providing a basis for subsequent review and accountability.

The key practice of MAS in multimodal settings is to decompose "perception-interpretation-integration-verification" into sequential steps handled by different agents, rather than asking a single model to directly generate final conclusions from raw multimodal inputs. Specifically, vision-related agents extract abnormal signs or quantitative features from images; structured-data agents interpret laboratory indicators and EHR variables; text-generation or report agents translate these results into clinically readable diagnostic statements; and coordinating or quality-control agents aggregate and review intermediate conclusions from different modalities. MedChat exemplifies the further

 doi:10.20944/preprints202606.0640.v2

translation of visual results into clinical language [109]. MAS for radiotherapy planning and focused ultrasound treatment planning organize task decomposition, parameter optimization, outcome assessment, and plan revision into continuous feedback chains [71,110]. Therefore, the core contribution of MAS in multimodal scenarios is not simply adding more modalities, but making cross-modal correspondences explicit through role specialization, thereby enabling cross-validation among different evidence sources.

However, multimodal integration does not inherently imply higher reliability. For MAS, the truly difficult issue is not modality access itself, but alignment among multiple modality-specific agents. Different agents may make judgments based on information from different time points, resolutions, or granularities. If signs extracted by a vision agent are not strictly aligned with the clinical context used by a text agent, they may be overinterpreted during downstream integration. Contradictions between structured indicators and free text may also be weakened or even erased at the aggregation stage. The final output may appear semantically complete and logically coherent, while its internal evidence is not genuinely aligned. Thus, the focus of multimodal information architecture should not be limited to enhancing fusion capability; it should also include source annotation across multimodal agents, retention of intermediate results, cross-node contradiction detection, and explicit expression of uncertainty. Otherwise, multimodal systems may deliver not more reliable judgments, but only stronger surface-level consistency.

3.4. Cross-Institutional Collaboration and Privacy-Preserving Coordination

The advantage of MAS in cross-institutional collaboration is that different functions can be deployed and executed within different system boundaries, with collaboration achieved through controlled interfaces, rather than centralizing all data and capabilities into a single model. Inpatient pathway support requires connectivity with EHR systems and clinical pathway platforms [104]. Order-set optimization must be embedded in clinical decision-support environments [47]. Radiotherapy treatment planning depends on specialized software interfaces such as treatment planning systems (TPS) [110]. FUAS-Agents are likewise built on specific toolchains [71]. Therefore, for MAS, cross-institutional collaboration is first an interface-orchestration problem, not merely a model-reasoning problem.

A more feasible direction is not to centralize all data in a single model, but to organize controlled collaboration across system boundaries. Structured EHRs, external knowledge, and physician-like discussions can be placed within the same workflow [62]. Privacy-preserving data interaction architectures further show that natural-language parsing and conversion into structured requests can be handled by outer-layer nodes, whereas key steps involving interpretation and write-back of sensitive medical records are better retained within the hospital [111]. In this context, privacy protection is primarily reflected in how deployment boundaries constrain collaborative orchestration, rather than as an isolated ethical topic.

Once MAS enter real deployment, the main difficulties manifest as the simultaneous rise of interface complexity and governance cost. Quality evaluation before clinical use, human review, and continuous post-deployment safety monitoring should all be incorporated into the governance framework [47,112]. More importantly, malicious nodes, communication contamination, and structural vulnerabilities in multi-agent topologies may expand local errors into systemic risks [112]. Therefore, the emphasis here is not to repeat privacy and ethics discussions, but to determine whether interface contracts, permission boundaries, log tracing, exception rollback, and human takeover mechanisms can be designed together with the collaborative architecture.

3.5. Reinforcement Learning–Based Optimization and Evolution of Medical Agents

Unlike the relatively static architectural designs discussed above, this section focuses on how medical agents adjust their behavior through feedback. For MAS, the significance of reinforcement learning and evolution is not merely to improve the response quality of individual nodes, but to optimize collaborative behaviors such as role division, communication order, tool invocation, and

 doi:10.20944/preprints202606.0640.v2

termination conditions. Existing studies generally follow two paths. One uses language-based reflection or expert feedback to correct local behaviors within established workflows [87,113]. The other formalizes multi-turn inquiry or multimodal reasoning as an optimizable sequential decision-making process, allowing the system to learn more effective collaborative strategies through interaction [114,115]. This indicates that medical MAS are no longer limited to executing predefined workflows, but are beginning to acquire limited feedback-adaptive capabilities.

Compared with single agents, the distinctive feature of MAS in this direction is that the optimization target is not only "how to answer," but also "who should answer, when to hand off, and how to integrate disagreements." Recent studies have therefore begun to extend optimization objectives from the local strategies of individual agents to the collaborative structure itself, moving beyond prompt adjustment or single-step decisions toward optimization of role boundaries, collaboration order, and information-routing patterns [37,116]. From this perspective, the "evolution" of medical agents is gradually extending from parameter-level capability improvement to process-level and structure-level adjustment.

However, optimization in MAS differs markedly from optimization in single agents. First, final performance is often determined by multiple nodes and multiple interaction rounds, making it difficult to assign rewards accurately to specific agents and communication decisions; this creates a credit-assignment problem [114]. Second, simultaneous updates across multiple agents introduce clear non-stationarity: a change in one role's strategy alters the optimal responses of other roles, increasing uncertainty in training and evaluation [115]. Third, existing studies rely heavily on synthetic patients and proxy evaluators [114]. Architecture search is still mainly validated on controlled benchmarks [116]. Improvements in self-evolving consultation frameworks are also largely based on limited case accumulation [37]. Under these conditions, systems may learn to fit evaluators or workflow templates rather than improve real clinical value. More importantly, optimized collaborative strategies do not mean that LLMs have escaped basic risks such as hallucination, drift, and shared bias. Therefore, reinforcement learning–based optimization and evolution of medical agents have clear research value, but at the current stage they are better regarded as tools for improving collaboration efficiency, process adaptability, and structural robustness, rather than as direct evidence that autonomous clinical intelligence is mature.

Overall, what clinical MAS change is not the decomposition of one answer into multiple answers, but the explicit incorporation of hospital role structures, evidence flow, and accountability chains into system design. Their value lies in improving the analyzability, auditability, and governability of clinical processes; their risk lies in the possibility that originally local errors may expand into multi-step, multi-node cascades. The future entry of these systems into hospitals should not be judged by the best score on a single benchmark, but by whether they satisfy workflow alignment, evidence traceability, risk governability, accountable explanations, and regulatory readiness (Figure 4).

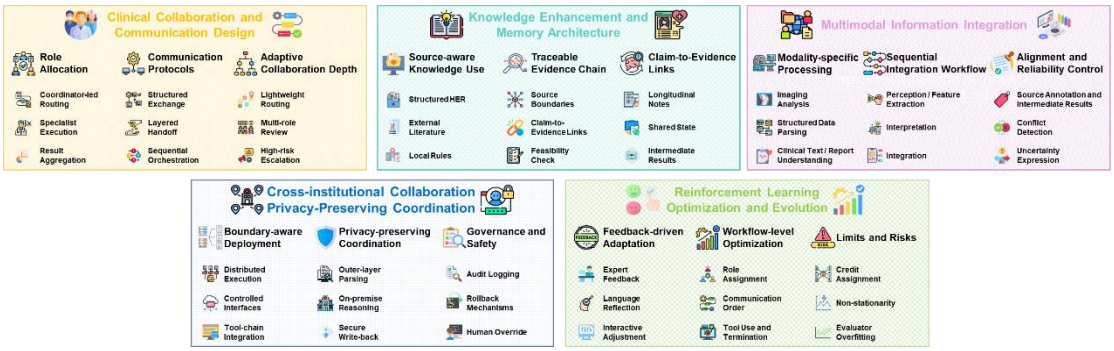

Figure 4. Overview of multi-agent system architecture for clinical workflows.

The framework organizes clinical multi-agent systems into five interconnected dimensions: collaboration and communication design, knowledge-enhanced evidence and memory architecture,

 doi:10.20944/preprints202606.0640.v2

multimodal information integration, cross-institutional and privacy-preserving coordination, and reinforcement learning-driven optimization and evolution.

4. Evaluation Systems and Clinical Validation Benchmarks

For medical MAS, evaluation should not be compressed into whether a one-off output is correct. Instead, it should be understood as the full process by which multiple role-based agents complete task decomposition, information acquisition, state updating, conflict resolution, and result release around the same clinical task. Unlike single agents, the potential benefits of MAS derive from role specialization, collaborative orchestration, and intermediate verification. Their performance evaluation should therefore test the whole collaboration chain, including whether it delivers stable judgments, clear evidence chains, and governable risks after adding communication and scheduling costs. If evaluation remains centered on static vignettes, single-turn answers, and endpoint matching, it will tend to underestimate the importance of the information-acquisition process and overestimate transferability to real clinical workflows. Therefore, evaluation should move beyond static accuracy toward a combined assessment of outcome performance, collaboration quality, and deployment attributes [7] (Figure 5).

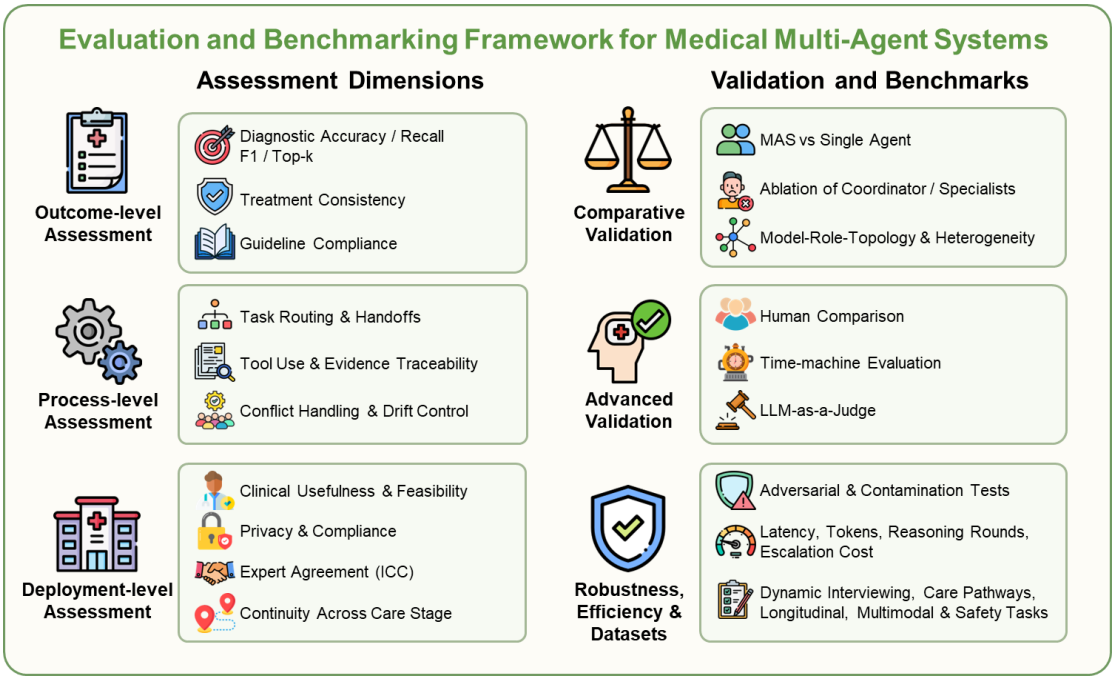

Figure 5. Overview of the evaluation and benchmarking framework for medical multi-agent systems.

The left side summarizes outcome-, process-, and deployment-level assessment dimensions. The right side highlights comparative and advanced validation methods together with robustness, efficiency, and representative benchmark datasets and tasks.

4.1. Evaluation Metrics and Methods

Evaluation of medical MAS should first retain basic outcome-level metrics, such as diagnostic accuracy, recall, F1 score, top-k hit rate, consistency of treatment recommendations, and guideline concordance. These metrics indicate whether system conclusions are clinically acceptable, but they do not explain how those conclusions are formed. For MAS, the more critical question is whether the collaborative process itself is reliable, including whether task routing is appropriate, information handoff between agents is complete, external tool use is correct, evidence sources are clear, and conclusions drift without justification after multiple rounds of collaboration [7]. At the same time, evaluation standards are moving closer to the dimensions used in real hospital review. Order-set

 doi:10.20944/preprints202606.0640.v2

optimization studies did not simply classify system outputs as "right" or "wrong," but asked clinicians to rate them across accuracy, usefulness, feasibility, and impact. This approach more closely resembles how hospitals actually review MAS before deployment [47]. In inpatient pathway tasks, evaluation further extends to the continuity and completeness of pathway stages, asking whether the system can maintain stable connections across multistage tasks such as triage, testing, diagnosis, and treatment [104]. In addition, automated medical text review studies introduce consistency metrics such as the intraclass correlation coefficient (ICC) to measure alignment between multi-agent review and expert review. This suggests that MAS evaluation is shifting from "whether the system is close to the standard answer" to "whether the system is close to the judgment pattern of expert groups" [117]. Medical MAS are therefore better evaluated through a framework that links outcome-level, process-level, and deployment-level metrics.

4.2. Comparative Analysis of Multi-Agent Collaboration Effectiveness

The collaboration effectiveness of medical MAS should not be compared only with general LLM baselines. More informative comparisons should include a single-agent version under the same base model, simplified versions with the coordinator or specialty agents removed, and human or semi-human workflows under the same task constraints. Such comparisons help distinguish improvements due to base-model capability from gains attributable to collaborative structure. Clinically scaled workload studies suggest that assigning retrieval, extraction, and calculation to different workers can improve stability under high mixed-task pressure. These gains appear to come from task isolation and orchestration structure, rather than simply from a larger model backbone [7]. Inpatient pathway studies also suggest that the advantage of MAS lies not only in endpoint scores, but also in tighter alignment with clinical workflows across triage, examination, diagnosis, and treatment recommendations [104].

In base-model comparison, medical MAS should not focus on whether the strongest model is used, but on whether the appropriate model is assigned to the appropriate role. EmulatRx compared GPT-4o, Phi-4, DeepSeek-R1:14B, and Gemma-3:12B, but ultimately used GPT-4o as the core node because it was more balanced across tasks such as concept extraction, SQL generation, causal inference, and report organization [87]. MMedAgent-RL did not assign all roles to the same model; instead, it configured different nodes according to distinct requirements for image understanding, cross-modal integration, and final adjudication, indicating that model selection is fundamentally governed by task structure rather than parameter scale alone [115]. More broadly, existing evidence suggests that there is no simple one-directional hierarchy among general-purpose base models, medical-specialized base models, and reasoning-oriented models. Rather, their relative value is task dependent. Current studies indicate that strong general-purpose base models are still frequently used as core coordinating or primary reasoning nodes in medical MAS, but their advantage lies more in overall robustness than in being unidirectionally optimal for every role and task. For example, EmulatRx used GPT-4o as its core node, and Gregory used o1 as its primary reasoning backbone, reflecting the fact that complex role coordination depends more on comprehensive robustness than on peak performance in a single capability [21,87]. Meanwhile, reasoning-oriented models are beginning to show potential in highly structured decision tasks. Conversational evaluation on JAMA Clinical Challenges showed that O1 and DeepSeek-R1-distill-LLaMA3-70B, with stronger built-in reasoning optimization, exhibited capability differences from general models within dynamic inquiry frameworks. In automated medical text review, GPT-o3-mini achieved the highest human-machine consistency. These findings suggest that reasoning-oriented models may be particularly useful for MAS nodes requiring evidence integration, judgment calibration, and multi-round deliberation. However, existing results remain largely task-specific and do not yet support comprehensive replacement of strong general-purpose base models [98,117].

By contrast, the value of medical-specialized base models may be more prominent in scenarios with dense specialty knowledge, limited data access, or stronger privacy constraints, rather than as a universal advantage across all medical MAS tasks. Under VHA constraints, CARE-AD used a fine-

 doi:10.20944/preprints202606.0640.v2

tuned LLaMA 3.1 8B model for information extraction from clinical notes and LLaMA 3 70B as specialty and aggregation agents, demonstrating the practical value of open-source or local models for building MAS in restricted hospital environments [63]. Inpatient pathway research has included medical-specialized models such as HuatuoGPT2, Clinical-Camel, and Meditron in comparisons, indicating that domain-specific base models have become an important class of candidates in MAS evaluation. However, based on current evidence, they are better understood as candidates worth comparing and adapting, rather than as models that have already proven consistently superior to strong general-purpose base models in multi-agent settings [104]. Thus, a more prudent discussion of base-model comparison should not presume that any model class is necessarily optimal, but should compare trade-offs in specialty knowledge, reasoning capacity, cost, privacy-oriented deployment conditions, and role fit.

In addition, automated screening studies for systematic reviews show that model heterogeneity itself may be a source of collaborative benefit. If all agents use the strongest and most expensive model from the same family, collaborative diversity may decline and computational cost may rise without necessarily achieving the overall optimum. Weaker models may partially compensate for lower individual capability through voting, debate, and adjudication mechanisms within role collaboration. In some workflows, therefore, system performance improvement may not depend entirely on replacing every node with the strongest model, but may arise from functional complementarity and cost-stratified configuration among stronger and weaker models [82]. Thus, in discussions of collaboration effectiveness, the critical comparison is not which single model is strongest, but which "model-role-topology" combination best supports the complementarity, stability, and deployability of MAS [116].

4.3. Advanced Validation Methods: Human Comparison and Temporal Backtesting

For medical MAS, advanced validation is important because these systems are not ordinary text-generation tools, but systems intended to enter clinical workflows. The value of human comparison is not merely to display model scores alongside physician scores, but to examine whether MAS can perform information collection, evidence integration, opinion coordination, and result release in a manner similar to clinical teams. The human-machine comparison of Gregory in complex neurological cases shows that higher-level evaluation should approximate real consultation logic. Neurosurgical multi-agent interaction evaluation further indicates that comparison with residents within a dynamic dialogue framework can test diagnostic organization and interaction performance more realistically than static clinical vignettes [21,95]. Similarly, MIRA, an EHR-integrated autonomous medical AI agent rather than a classical MAS, provides an important boundary example for workflow-level validation. In simulated emergency-department workflows based on more than 500 MIMIC-IV cases, it interacted with a patient agent, used structured EHR tools to order tests and treatments, and was compared with physician cohorts across diagnosis, guideline concordance, medication safety, and admission decisions [80]. This study suggests that future MAS evaluation should move beyond final-answer accuracy and explicitly test whether agentic systems can safely acquire information, execute structured clinical actions, and remain auditable within controlled EHR environments. Order-set optimization studies used expert scoring, filtering mechanisms, and comparisons with real ordering behavior to validate output quality, suggesting that evaluation of MAS cannot remain at the level of text similarity, but must address whether outputs can be clinically executed [47]. In addition, LLM-as-a-Judge may provide a scalable supplement for evaluating medical text generation. It can score factual correctness, completeness, and clinical utility, and metrics such as ICC can quantify consistency between system ratings and expert ratings [117]. However, such automated review should be calibrated against expert assessment rather than treated as a substitute for it.

The key idea of temporal backtesting is not retrospectively redoing a case, but placing MAS back at a historical time point before the case has concluded and allowing it to access only the information that was genuinely available at that time. This assesses whether shared states and intermediate

 doi:10.20944/preprints202606.0640.v2

evidence were sufficient to support early decision-making [63]. Conversational evaluation on JAMA has shown that critical clinical information is often not provided all at once, but is gradually revealed through interactions between patient agents and system agents. Therefore, the inquiry process must be explicitly incorporated into evaluation to avoid overoptimistic assessments of MAS capability [98]. This also means that dynamic trajectory evaluation of medical MAS should not only compare whether the final diagnosis is correct, but should also examine convergence speed, efficiency in acquiring key information, and interaction length across the full process of history taking, physical examination, testing, and diagnosis. Related studies have begun to use dialogue rounds, number of diagnoses, interaction length, and time efficiency as supplementary metrics, promoting a shift from static endpoint scores to process-level dynamic performance [21,98]. For systems that depend on longitudinal illness trajectories and continuous state updating, time slicing is especially necessary because it prevents information leakage from disguising "retrospective correctness" as "real-time collaborative capability" [63]. Traceable rare-disease diagnosis studies further require that candidate conclusions and evidence chains be stored together, enabling evaluators to verify whether final conclusions arise from intermediate evidence [35].

4.4. Robustness, Operational Efficiency, and Dataset Summary

Compared with single agents, medical MAS are more prone to chained fragility: an erroneous retrieval step, inappropriate routing decision, or distorted intermediate summary may be inherited, amplified, and solidified as a team conclusion in downstream nodes. Robustness evaluation should therefore not look only at endpoint error rates, but should examine whether the system can absorb errors, expose conflicts, trace evidence, and roll back exceptions. MedSentry's adversarial evaluation shows that different topologies differ in how risks diffuse under malicious prompts, information contamination, and internal mismatch. This indicates that safety should be treated as part of performance in medical MAS, not as an issue appended after deployment [112]. Building on this, evaluation can further include stress tests aimed at real deployment environments, such as adversarial prompts, information contamination, and topology perturbations, to avoid systems appearing stable only under ideal conditions [112]. Although Tree-of-Reasoning improves interpretability through evidence trees and cross-verification, such structures also introduce higher reasoning complexity; evaluation must therefore report both resource costs and workflow benefits [40].

Operational efficiency should also be a core evaluation dimension because the deployment value of MAS depends on whether additional collaboration costs generate net benefits. Clinically scaled workload studies have already shown that accuracy, latency, and token consumption must be reported together; otherwise, it is impossible to determine whether multi-agent orchestration truly outperforms single-agent processing [7]. At the same time, the number of reasoning rounds should be reported separately as a key operational indicator to more fully characterize the integrated trade-off among collaboration depth, real-time response latency, and resource consumption [21,98]. Cost-sensitive diagnosis studies further suggest that not all cases require full multi-agent deliberation. Using lightweight pathways for simple cases and escalating only difficult cases to MAS collaboration is itself an important component of operational-efficiency optimization [118].

In terms of dataset and benchmark composition, current evaluations of medical MAS are broadly organized around six categories of tasks. Dynamic inquiry and progressive information-disclosure tasks test information acquisition and interaction organization [98]. Inpatient pathway tasks assess stability across multistage role transitions [104]. Longitudinal disease-course tasks evaluate shared states and early prediction ability [63]. Safety evaluation frameworks test vulnerability and risk diffusion across different topologies [112]. Cost-sensitive diagnostic tasks emphasize trade-offs between operational cost and escalation strategies [118]. Multimodal reasoning tasks further assess cross-modal integration [115]. These datasets have pushed medical MAS evaluation from static question answering toward process, time, and deployment dimensions. However, task definitions remain inconsistent, process annotations are insufficient, and reporting standards are scattered. A

 doi:10.20944/preprints202606.0640.v2

minimum reporting framework should preserve outcomes, routing logs, tool invocations, evidence chains, disagreement states, and resource consumption, so that the added value of MAS can be compared directly rather than inferred from final scores alone (Table 2).

Table 2. Representative evaluation datasets and benchmarks for medical MAS.

Dataset/ benchmark	*Data source and type*	*Main task scenario*	*Primary capability evaluated*	*Common metrics*	*Representative references*
JAMA Clinical Challenges	*JAMA clinical challenge cases (815 selected from 1,519 cases)*	*Dynamic diagnosis and progressive information disclosure*	*Inquiry organization, information acquisition, dynamic diagnostic convergence*	*Accuracy, dialogue rounds*	*[98]*
MIMIC series (MIMIC-III/IV)	*Real inpatient EHR, ICU data, and progress notes*	*Inpatient pathways, task orchestration, entity extraction, causal analysis, cost-sensitive diagnosis*	*Workflow continuity, shared-state modeling, operational efficiency*	*Accuracy, F1, AUC, token use, latency*	*[7,87,104,118]*
PDSQI-9 + MIMIC-III/ProbSum	*Medical document summaries and quality scales*	*Automated review of medical text-generation quality*	*Expert agreement and reliability of automated review*	*ICC, Krippendorff's alpha, Gwet's AC2*	*[117]*
VHA longitudinal clinical notes (CARE-AD)	*Long-term medical records and clinical notes from the U.S. Veterans Health Administration*	*Early prediction of Alzheimer's disease*	*Long-term memory, shared state, time-sliced prediction*	*Accuracy, Precision, Recall, F1*	*[63]*
VUMC real-world order sets	*Real order sets and knowledge base from Vanderbilt University Medical Center*	*Order-set optimization and assessment of clinical executability*	*Output executability, human-machine agreement, expert filtering*	*Expert ratings, Cohen's kappa*	*[47]*
Clinically scaled mixed-task set	*PubMed literature, EHR discharge summaries, and dosage-calculation tasks*	*Mixed workload involving retrieval, extraction, and calculation*	*Stability, scalability, operational efficiency*	*Accuracy, latency, token cost*	*[7]*

5. Challenges and Ethics

5.1. Technical Summary and Challenges

The key challenge for medical MAS is not whether an individual model is sufficiently strong, but whether multiple role-based agents can form a stable, traceable, and rollback-capable collaboration structure within the same clinical task. Compared with single agents, MAS must simultaneously handle task decomposition, shared-state updating, cross-node messaging, tool invocation, and result integration. Their failure modes therefore no longer primarily appear as single-point errors, but more often as chain propagation, stage-wise accumulation, and inter-role

 doi:10.20944/preprints202606.0640.v2

amplification. For this reason, the technical bottleneck of medical MAS is fundamentally a problem of collaborative governance, rather than merely a problem of model capability.

5.1.1. Medical Hallucination and Cascaded Error Amplification

In medical MAS, the greatest risk is often not the first inaccurate judgment made by an agent, but the possibility that this judgment will be accepted as a premise by downstream nodes and gradually solidified through paraphrasing, summarization, and adjudication. Research on cognitive bias correction shows that multi-role discussion can reduce anchoring bias and premature closure only when explicit mechanisms for rebuttal, correction, and escalation are in place. Otherwise, multi-round discussion may cause errors to appear as collective consensus [106]. Order-set optimization studies also indicate that content critique, knowledge retrieval, medication validation, and summary generation need to be separated precisely to prevent a single unverified error from directly entering the final recommendation [47]. Safety evaluations further show that once upstream nodes are contaminated or misled, local errors may diffuse along the collaborative topology into systemic risks [112].

5.1.2. Bottlenecks in Collaborative Scheduling and Risk of Consensus Bias

The performance gains of MAS arise from division of labor, but the system burden also arises from that same division of labor. Clinically scaled workload studies show that appropriate task isolation and lightweight orchestration can improve stability; however, as the number of communication rounds and roles increases, latency, token consumption, and context redundancy also increase [7]. In multimodal reasoning scenarios, specialty agents may produce inconsistent local conclusions. If the system uses centralized adjudication, the adjudication node may also be affected by model-family bias, weakening the diversity advantage that heterogeneous collaboration is meant to provide [82]. Meanwhile, if a system relies excessively on majority opinion or uses inflexible static collaboration workflows, it may mistake surface-level agreement for high-quality consensus and thereby obscure genuine conflicts among cross-modal evidence [115]. As research moves further toward automated architecture search, the challenge shifts from “how to answer” to “who should answer, in what order handoff should occur, and where the process should terminate.” This indicates that the bottleneck of medical MAS is not an insufficient number of agents, but whether topology, model, and task are appropriately matched [116].

5.1.3. Clinical Memory Management and Pressure on Privacy Boundaries

The memory required by clinical MAS is not simply long context in a general sense, but continuous coordination among shared case state, role-specific private memory, and external medical-record interfaces. CARE-AD shows that the value of longitudinal clinical reasoning comes from preserving long-term disease-course clues. If outdated summaries or erroneous intermediate judgments are mixed into the shared state, multiple downstream agents may continue reasoning around a distorted patient profile [63]. Privacy-preserving EHR collaboration systems suggest that interpretation and write-back of sensitive information are better retained within in-hospital private nodes, while outer-layer agents handle only structured requests and controlled message exchange. Otherwise, the stronger the collaborative capability, the larger the potential exposure surface [111]. Traceable diagnostic studies further show that intermediate evidence and decision pathways must be preserved together, so that shared memory is not only information retention but also a collaborative foundation for review, correction, and audit [35].

5.2. Ethical and Privacy Issues

Compared with single-agent systems, medical MAS do not reduce existing ethical and privacy challenges. Instead, through task decomposition, role collaboration, shared memory, and tool orchestration, they transform risks that previously mainly existed at input and output boundaries

 doi:10.20944/preprints202606.0640.v2

into systemic issues running through the entire collaboration chain [10,119–121]. In this framework, sensitive information, erroneous judgments, and biased reasoning may appear not only in final responses, but also continue to spread through inter-agent communication, shared states, and tool invocations. Governance therefore no longer targets only a single model, but the entire orchestration process and its life cycle [119,121–123].

5.2.1. Expansion of Privacy Boundaries

In single-agent systems, privacy risks mainly arise during input invocation and final output. MAS further expand the privacy exposure surface: patient information may be copied, cached, and reused in inter-agent messages, shared memory, tool arguments, and runtime logs [119,120,122]. AgentLeak reports that the overall privacy exposure of MAS is approximately 1.6 times that of single-agent systems, indicating that risk may extend from output-level leakage to continuous exposure within internal data flows [122]. Current research is moving governance upstream into the internal links themselves, including full-process control of agent communication, shared memory, tool invocation, and logs; minimum necessary access, memory isolation, and permission boundaries for different agents; and, in cross-institutional collaboration, the use of federated learning, differential privacy, and secure aggregation to reduce centralized exposure of raw patient data [122–126].

5.2.2. Complexification of Accountability Structures

Errors in single-agent systems can usually be traced relatively easily to the model, training data, or deployment entity. In MAS, once planning, retrieval, reasoning, validation, and execution are distributed across different roles, the accountability chain becomes substantially longer [119,120]. An inappropriate clinical recommendation may simultaneously involve task allocation, evidence retrieval, specialty reasoning, result verification, and workflow design [44,119,123]. Radiology-specific discussions have emphasized that MAS may introduce governance issues beyond those of single AI systems, including distributed accountability, auditability, and responsibility across interacting agents [18]. This structure more closely resembles real clinical collaboration, but it also makes transparency, interpretability, and accountability boundaries more ambiguous [119,121]. To address this problem, current research no longer treats accountability merely as post hoc attribution, but attempts to embed it into system design. Examples include validation agents, traceable logs, evidence-alignment mechanisms, and medical algorithmic auditing approaches that can help localize errors, explain outputs, and monitor vulnerable links or post-deployment drift [44,123,127,128].

5.2.3. Chain Amplification of Bias Propagation

MAS are not inherently fairer or more reliable than single agents. On the contrary, a biased judgment, hallucinated content, or incomplete retrieval result from an upstream agent may be accepted and amplified by downstream agents, ultimately forming a false consensus [44,119,121]. The WHO has noted that generative AI in health may produce inaccurate or biased information, and that clinicians may overtrust system outputs and fail to recognize errors, bias, or incomplete evidence in time. In multi-agent environments, this risk is more likely to propagate along workflow chains and become harder for clinicians to detect because of apparent mutual corroboration [121]. Mitigation strategies include independent validation and cross-review, stronger evidence-based intermediate checks, disclosure of training data and applicability boundaries, and continuous clinician involvement in post-deployment supervision and fairness evaluation, so that overreliance on automated outputs does not replace professional judgment [44,120,127,128].

5.3. Regulation, Standardization, and Real-World Deployment

The regulatory challenge of medical MAS is not simply whether an individual model performs well, but whether a dynamic, multi-role, tool-using workflow can be governed safely across its full life cycle. Existing frameworks for software as a medical device, clinical decision-support software,

 doi:10.20944/preprints202606.0640.v2

AI-enabled medical devices, and quality-management systems provide useful foundations for intended-use definition, risk management, human oversight, change control, cybersecurity, and post-market monitoring [120,121,127–136]. However, MAS require these principles to be applied to the orchestration chain, not only to final outputs.

5.3.1. Intended Use and Regulatory Classification

Medical MAS should be classified according to intended use, user role, and clinical consequence rather than by the generic label of "agentic AI." FDA guidance on clinical decision-support software distinguishes functions that allow clinicians to independently review the basis for recommendations from software functions that may require device-level oversight [131]. The FDA database of AI/ML-enabled medical devices similarly shows that authorized AI systems are evaluated as concrete medical functions, organized by clinical area, intended use, and regulatory pathway [129]. For MAS, the intended use should specify which agent performs each function, whether the system summarizes evidence or recommends action, whether clinicians can inspect the evidence chain, and whether outputs influence diagnosis, triage, treatment planning, surgery, monitoring, or administrative workflow. Under the EU AI Act, health-related AI systems in safety-critical contexts may fall within high-risk obligations for risk management, data governance, logging, transparency, human oversight, robustness, accuracy, and cybersecurity [132]. A MAS used for patient-facing triage or treatment planning therefore has a very different regulatory profile from one used only for literature retrieval or internal scheduling.

5.3.2. Lifecycle Standards and Change Control

Medical MAS are difficult to govern as static software because their behavior may change when prompts, retrieval corpora, tools, routing rules, memory policies, base models, or external APIs are updated. Existing standards can provide a scaffold: IEC 62304 addresses medical-device software life-cycle processes [133]; ISO 14971 addresses risk management [134]; ISO 13485 addresses quality-management systems [135]; and ISO/IEC 42001 addresses organization-level AI management systems [136]. For MAS, these principles should extend from individual modules to collaborative workflows. Each agent should have a defined role, input boundary, output contract, permission scope, escalation rule, and logging requirement. FDA guidance on predetermined change-control plans is especially relevant because planned modifications, validation methods, and impact assessments should be specified before deployment [130]. In MAS, change control should cover not only model updates, but also changes in agent topology, tool access, retrieval sources, shared memory, and coordination policy. Otherwise, a seemingly minor update to one agent may alter downstream clinical behavior in ways that final-answer testing alone cannot detect.

5.3.3. Deployment Accountability and Post-Market Monitoring

Real-world deployment also requires clearer accountability across developers, hospitals, clinicians, system integrators, data providers, and tool vendors [119,120,123]. Medical MAS should report the accountable organization, intended users, clinical setting, human-override pathway, monitoring owner, incident-reporting process, cybersecurity controls, rollback conditions, and audit-log design. Post-market monitoring should evaluate not only diagnostic accuracy or recommendation concordance, but also routing failures, unsafe tool calls, privacy leakage, escalation delays, automation bias, and disagreement between system suggestions and clinician actions [121,127,128]. Deployment should proceed through staged evaluation, including silent-mode testing, clinician-supervised use, bounded pilot implementation, continuous audit, and explicit withdrawal criteria. Until these safeguards are in place, medical MAS should be framed as clinician-supervised workflow-support systems rather than autonomous clinical actors(Figure 6).

Commercial and regulatory deployment remains limited. Although FDA guidance, AI-enabled device listings, the EU AI Act, and IEC/ISO standards provide frameworks for intended use,

 doi:10.20944/preprints202606.0640.v2

oversight, change control, and post-market monitoring, the included LLM-based medical MAS studies are still mainly academic prototypes or retrospective/simulated evaluations. We did not identify mature evidence of routine commercial deployment, FDA-cleared or CE-marked status, or prospective post-market performance for an LLM-based medical MAS workflow system.

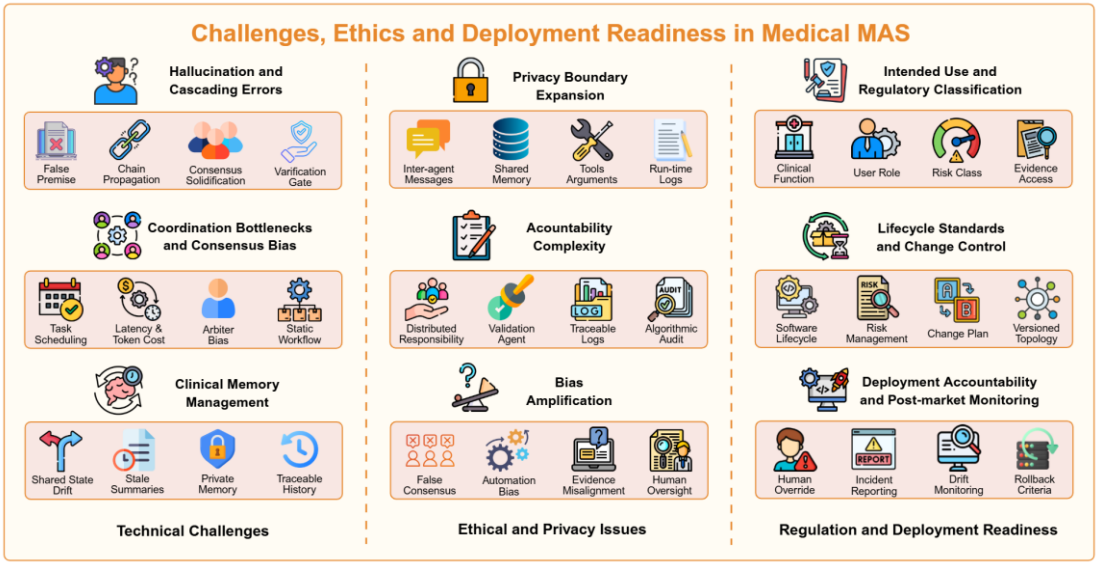

Figure 6. Overview of the challenges and ethics for medical multi-agent systems.

Medical MAS face both technical and ethical challenges. Key technical risks include hallucination and cascading errors, coordination bottlenecks, consensus bias, and clinical memory management difficulties. Ethical concerns include expanded privacy boundaries, more complex accountability, and bias amplification across collaborative workflows, underscoring the need for stronger governance and human oversight.

6. Future Directions

6.1. Paths for Technical Innovation

Future technical innovation in medical MAS should not be understood as continued expansion of the parameter scale of individual foundation models, but as improving the verifiability, controllability, and evolvability of the collaboration chain itself. Once MAS enter clinical settings, the central question is no longer whether an answer can be produced. It is who provides what evidence and when, how the system handles disagreement, and when escalation or human takeover should be triggered. The next stage of innovation should therefore focus primarily on self-verification, bounded collaboration, and knowledge constraints.

First, medical MAS need to build metacognitive capabilities into the architectural layer, rather than placing error correction only after final output. MD-PIE suggests that MAS can introduce specialty agents for querying, feedback, and bridge validation in addition to the attending agent, allowing the system to continuously assess during reasoning whether current evidence is sufficient, whether candidate diagnoses should be retained, and whether higher-level knowledge support is needed [38].

Second, at the privacy and deployment levels, this self-verification mechanism should be combined with controlled collaboration boundaries. A more feasible path is not to allow all nodes to share complete data, but to establish hierarchical collaboration among external coordination nodes, in-hospital medical-record interpretation nodes, and local execution nodes, keeping high-risk data processing within private environments as much as possible [111]. Thus, the emphasis of future "federated self-evolving collaboration" lies not only in using distributed updating and parameter-

 doi:10.20944/preprints202606.0640.v2

sharing mechanisms to alleviate cross-institutional data silos [124], but also in combining differential privacy and secure aggregation to reduce raw patient-data exposure while maintaining system performance and collaboration efficiency as much as possible [125]. At the same time, role boundaries, message interfaces, and escalation rules need continuous optimization to ensure the controllability and governability of MAS in real deployment [111].

Third, reliable reasoning in medical MAS requires both structured knowledge constraints and controllable human-AI collaboration. Knowledge graphs, evidence-based relations, and patient-specific information should not merely be passively input as retrieved fragments, but should become shared structured scaffolds across agents that explicitly constrain reasoning paths. Human-AI collaborative knowledge-graph systems show that structured knowledge can improve information consistency and interpretability during collaboration [137]. Traceable rare-disease diagnostic systems further demonstrate that MAS become genuinely auditable only when conclusions and intermediate evidence are preserved together [35]. Conversely, physicians should not merely provide final signatures; they should act as high-authority nodes in conflict adjudication, human escalation, and workflow reconstruction. Order-set optimization research has shown that the value of human-AI collaboration lies not in simple review of outputs, but in embedding human judgment into high-risk nodes [47]. The human-in-the-loop (HITL) mechanism embodied by EmulatRx further suggests that real feedback is better used to continuously revise role division, collaboration order, and output constraints, rather than serving only as an offline evaluation signal [87]. Accordingly, future medical MAS are more likely to evolve into clinical systems constrained by knowledge and governed jointly by humans and AI, rather than fully autonomous closed agent clusters.

6.2. Clinical Application Expansion

Compared with the application scenarios summarized above, future clinical expansion should not simply replicate existing capabilities across more specialties. Instead, it should embed AI more deeply into broader healthcare service networks and translational medicine domains: on the one hand, into home, community, and public health networks; on the other, into translational processes such as drug discovery, digital twins, and clinical trials [138–145]. For medical MAS, this trend shifts the focus toward cross-institutional, cross-scale, and cross-modal data collaboration, while requiring future studies to distinguish the contribution of agentic coordination from that of the underlying AI models, data platforms, and digital-health infrastructure.

Post-discharge home monitoring can improve safety and adherence and shows trends toward reducing readmissions, shortening hospital stays, and decreasing certain types of healthcare use [138]. Remote cognitive assessment based on smartphones and smartwatches has been validated as feasible in large populations, suggesting that early screening may be performed in daily-life settings [146]. Precision public health emphasizes the integration of genetic, behavioral, environmental, and AI-derived data to implement more precise prevention, diagnosis, and intervention at the population level [139]. Population-level studies have also extended applications to infection control, immunization planning, long-term health modeling, and resource optimization [140]. If medical MAS can connect data from wearable devices, home monitoring, internet hospitals, and community health centers, their value may lie in high-risk population identification, stratified referral, and public health services rather than merely automating individual follow-up.

Medical MAS can further extend from clinical service support to medical knowledge generation and translational research. AI applications in drug development now span target identification, molecule generation, clinical research, and post-marketing surveillance [141]. Studies have shown that the generative AI-discovered TNIK inhibitor rentosertib has entered human studies in patients with idiopathic pulmonary fibrosis [142]. Digital twins are being defined as individualized, dynamically updated, and predictive patient models, but only a minority of current studies truly meet this standard [143]. In type 1 diabetes, one randomized clinical trial used digital twin technology for biweekly parameter optimization and increased time in range from 72% to 77% [144], indicating that digital twins are beginning to move from conceptual models into the clinical frontier of

 doi:10.20944/preprints202606.0640.v2

intervention optimization and trial evaluation [143,144]. At the clinical-trial level, recent studies have proposed frameworks for dynamic deployment and continuous clinical validation [145], and AI is increasingly being applied to the assessment of trial safety, efficacy, and operational risks [147]. These advances suggest that future MAS could extend beyond decision support into medical discovery and clinical translation, especially if agentic coordination is used to organize evidence integration, iterative validation, and cross-domain feedback.

7. Conclusion

Medical MAS are moving medical AI from model applications oriented toward single tasks toward systems-level collaboration embedded in clinical workflows. This review establishes a conceptual framework for medical MAS. Compared with single agents, these systems are better able to incorporate clinical role division, evidence integration, workflow orchestration, and risk control into a unified operational framework. Accordingly, they have shown task adaptability and translational potential in diagnosis and differential diagnosis, treatment decision-making, medical imaging analysis, patient monitoring, intraoperative assistance, hospital workflow optimization, evidence-based research, medical education, and clinical safety governance. In the future, medical MAS are likely to develop toward proactive monitoring, individualized continuous management, cross-institutional collaboration, multimodal integration, and real-world deployment under human-AI co-governance. However, communication overhead, error cascades, privacy exposure, increasingly complex accountability, regulatory uncertainty, and insufficient evaluation standards remain key constraints on large-scale clinical adoption. The next stage will require continued refinement of knowledge constraints, verification mechanisms, governance frameworks, and human takeover pathways. Overall, medical MAS provide an important technical direction for integrating medical AI into healthcare systems, but their clinical impact will depend on rigorous validation, governance, and clinician-supervised deployment.

Author Contributions: S.J.X., Y.Y.D. and S.Z.L. conceived the idea. T.Y.X. and H.Z.G. are co-first authors who contributed equally to this work. T.Y.X. and H.Z.G. collected and analyzed relevant literature and data, T.Y.X. and H.Z.G. wrote the manuscript and created and generated the figures and tables. R.S., X.R.L., Z.L.Z., X.Y.L., H.Y.L., Y.L., S.Z.L., Y.Y.D. and S.J.X. commented on and revised the manuscript.

Acknowledgments: This work was supported by the Hubei Chutian Talents Entrepreneurship and Innovation Team (N20252453), the Natural Science Foundation of Hubei Province (2026AFB788), the Medical Artificial Intelligence General Program of Tongji Hospital 2025 (No. AI2025A02), and the Qingyan Fund — Young Scholars Development Program (QYJJ-QNXZ-11).

Declaration of interests: The authors declare no financial competing interests. To address potential citation-related conflicts, the authors audited the core evidence table for direct author overlap between the manuscript authors and the included core empirical studies. No direct author overlap was identified in the current core evidence table.

Declaration of generative AI and AI-assisted technologies in the writing process: Large language models were used solely for grammatical refinement and linguistic polishing. After utilizing such tools, the authors conducted necessary review and editing, and take full responsibility for the final published content.:

References

1. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184
2. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI*. 2024;1(1). doi:10.1056/AIp2300031
3. Bean AM, Payne RE, Parsons G, et al. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nat Med*. 2026;32(2):609-615. doi:10.1038/s41591-025-04074-y

 doi:10.20944/preprints202606.0640.v2

4. Liu F, Niu Y, Zhang Q, et al. A foundational architecture for AI agents in healthcare. *Cell Reports Medicine*. 2025;6(10):102374. doi:10.1016/j.xcrm.2025.102374
5. Fan W, Chen P, Shi D, Guo X, Kou L. Multi-agent modeling and simulation in the AI age. *Tsinghua Sci Technol*. 2021;26(5):608-624. doi:10.26599/TST.2021.9010005
6. Gao S, Fang A, Huang Y, et al. Empowering biomedical discovery with AI agents. *Cell*. 2024;187(22):6125-6151. doi:10.1016/j.cell.2024.09.022
7. Klang E, Omar M, Raut G, et al. Orchestrated multi agents sustain accuracy under clinical-scale workloads compared to a single agent. *npj Health Syst*. 2026;3(1):23. doi:10.1038/s44401-026-00077-0
8. Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. *Front Comput Sci*. 2024;18(6):186345. doi:10.1007/s11704-024-40231-1
9. Guo T, Chen X, Wang Y, et al. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In: Vol 9. 2024:8048-8057. doi:10.24963/ijcai.2024/890
10. Wang W, Ma Z, Wang Z, et al. A Survey of LLM-based Agents in Medicine: How far are we from Baymax? In: Che W, Nabende J, Shutova E, Pilehvar MT, eds. *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics; 2025:10345-10359. doi:10.18653/v1/2025.findings-acl.539
11. Zou J, Topol EJ. The rise of agentic AI teammates in medicine. *The Lancet*. 2025;405(10477):457. doi:10.1016/S0140-6736(25)00202-8
12. Moritz M, Topol E, Rajpurkar P. Coordinated AI agents for advancing healthcare. *Nat Biomed Eng*. 2025;9(4):432-438. doi:10.1038/s41551-025-01363-2
13. Kim H, Jo S, Lim MH, Choi DH. From non-agentic large language models to multi-agent systems in emergency medicine: a scoping review. *CEEM*. Published online April 8, 2026. doi:10.15441/ceem.26.136
14. Spieser J, Balapour A, Meller J, Patra KC, Shamsaei B. A Review of Multi-Agent AI Systems for Biological and Clinical Data Analysis. *Methods and Protocols*. 2026;9(2):33. doi:10.3390/mps9020033
15. Branda F, Ahmed MM, Ciccozzi M, Guzzi PH, Scarpa F. The next paradigm in bioinformatics: a review of multi-agent systems and foundational models for end-to-end scientific discovery. *Brief Bioinform*. 2026;27(3):bbag245. doi:10.1093/bib/bbag245
16. Guo X, Chen J, Zhang Y, et al. Precision oncology: from large language models to multi-agent systems. *Front Oncol*. 2026;16. doi:10.3389/fonc.2026.1828507
17. Xie Z, Wang H, Dai L, Wang Z, Song H, Qian J. Ethical issues in multi-agent AI systems for healthcare: a narrative review. *Front Public Health*. 2026;14. doi:10.3389/fpubh.2026.1792627
18. Salehi S, Singh Y, Habibi P, Erickson BJ. Beyond Single Systems: How Multi-Agent AI Is Reshaping Ethics in Radiology. *Bioengineering*. 2025;12(10):1100. doi:10.3390/bioengineering12101100
19. Kim J, Lui B, Goldstein PA, Rubin JE, White RS, Jotwani R. From Data to Decisions: Harnessing Multi-Agent Systems for Safer, Smarter, and More Personalized Perioperative Care. *Journal of Personalized Medicine*. 2025;15(11):540. doi:10.3390/jpm15110540
20. Yao S, Zhao J, Yu D, et al. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv*. Preprint posted online March 10, 2023:arXiv:2210.03629. doi:10.48550/arXiv.2210.03629
21. Sorka M, Gorenshtein A, Abramovitch H, Soontrapa P, Shelly S, Aran D. AI vs Human Performance in Conversational Hospital-Based Neurological Diagnosis. *medRxiv*. Preprint posted online August 14, 2025. doi:10.1101/2025.08.13.25333529
22. Almansoori M, Kumar K, Cholakkal H. Self-Evolving Multi-Agent Simulations for Realistic Clinical Interactions. *arXiv*. Preprint posted online 2025. doi:10.48550/ARXIV.2503.22678
23. Sorka M, Gorenshtein A, Aran D, Shelly S. A multi-agent approach to neurological clinical reasoning. *PLOS Digital Health*. 2025;4(12):e0001106. doi:10.1371/journal.pdig.0001106
24. Zhao Y, Wang H, Zheng Y, Wu X. A Layered Debating Multi-Agent System for Similar Disease Diagnosis. In: Chiruzzo L, Ritter A, Wang L, eds. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Association for Computational Linguistics; 2025:539-549. doi:10.18653/v1/2025.naacl-short.46

 doi:10.20944/preprints202606.0640.v2

25. Wu Y, Wan G, Li J, et al. WiseMind: a knowledge-guided multi-agent framework for accurate and empathetic psychiatric diagnosis. *npj Digit Med*. Published online March 25, 2026. doi:10.1038/s41746-026-02559-9
26. Ji X, Li Z, Zeng L, et al. Early diagnosis of axial spondyloarthritis in primary care using multi-agent systems. *npj Digit Med*. 2026;9(1):185. doi:10.1038/s41746-026-02372-4
27. Junior EP, de Sousa FC, Chen J, Camacho D, Benjamin SR, de Albuquerque VHC. Multimodal diagnosis of Parkinson's disease with an internet-based collaborative agent architecture of medical language models. *Computers in Biology and Medicine*. 2026;203:111468. doi:10.1016/j.compbiomed.2026.111468
28. Zhang M, Cui Q, Lü Y, Pan Y, Yu W, Li W. A large language model-based self-learning and critical agent framework for multimodal Alzheimer's disease diagnosis. *Neuropsychology*. Published online March 26, 2026. doi:10.1037/neu0001079
29. Wu X, Zhang H, Garduno-Rapp NE, et al. Orchestrator multi-agent clinical decision support system for secondary headache diagnosis in primary care. *J Am Med Inform Assoc*. Published online June 27, 2026:ocag111. doi:10.1093/jamia/ocag111
30. Zhang Y, Qiu Y, Lin X. Bridging Perception and Reasoning: An Evidence-Based Agentic System for Diagnosis and Treatment Recommendations of Vascular Anomalies. *Diagnostics*. 2026;16(4):621. doi:10.3390/diagnostics16040621
31. Shang T, He W, Zheng C, Li L, Shen L, Zhao B. DynamiCare: A Dynamic Multi-Agent Framework for Interactive and Open-Ended Medical Decision-Making. *AMIA Jt Summits Transl Sci Proc*. 2026;2026:408-417.
32. Chen X, Yi H, You M, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *npj Digit Med*. 2025;8(1):159. doi:10.1038/s41746-025-01550-0
33. Neeley MB, Qi G, Wang G, et al. Survey and improvement strategies for gene prioritization with large language models. Zhu S, ed. *Bioinformatics Advances*. 2024;5(1):vbaf148. doi:10.1093/bioadv/vbaf148
34. Zhou X, Ren Y, Zhao Q, et al. An LLM-Driven Multi-Agent Debate System for Mendelian Diseases. *arXiv*. Preprint posted online April 11, 2025:arXiv:2504.07881. doi:10.48550/arXiv.2504.07881
35. Zhao W, Wu C, Fan Y, et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*. 2026;651(8106):775-784. doi:10.1038/s41586-025-10097-9
36. Qi G, Wang J, Chong ML, et al. RareCollab -- An Agentic System Diagnosing Mendelian Disorders with Integrated Phenotypic and Molecular Evidence. *arXiv*. Preprint posted online February 3, 2026:arXiv:2602.04058. doi:10.48550/arXiv.2602.04058
37. Chen K, Li X, Yang T, Wang H, Dong W, Gao Y. MDTeamGPT: A Self-Evolving LLM-based Multi-Agent Framework for Multi-Disciplinary Team Medical Consultation. *arXiv*. Preprint posted online March 18, 2025:arXiv:2503.13856. doi:10.48550/arXiv.2503.13856
38. Esteitieh Y, Mandal S, Laliotis G. Towards Metacognitive Clinical Reasoning: Benchmarking MD-PIE Against State-of-the-Art LLMs in Medical Decision-Making. *Health Informatics*. Preprint posted online January 29, 2025. doi:10.1101/2025.01.28.25321282
39. Chen K, Qi J, Huo J, et al. 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)*. IEEE; 2025:1-5. doi:10.1109/ICASSP49660.2025.10889517
40. Peng Q, Cui J, Xie J, Cai Y, Li Q. Tree-of-Reasoning: Towards Complex Medical Diagnosis via Multi-Agent Reasoning with Evidence Tree. *arXiv*. Preprint posted online August 5, 2025:arXiv:2508.03038. doi:10.48550/arXiv.2508.03038
41. Madrid-García A, Benavent D, Merino-Barbancho B. From chat to act: large language model agents and agentic AI as the next frontier of AI in rheumatology. *EULAR Rheumatology Open*. 2025;1(3):147-156. doi:10.1016/j.ero.2025.06.012
42. Xu G, Meng Y, Wang R, Qi G. Collaborating LLMs and PLMs for Medical Tasks. In: *2024 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE; 2024:428-431. doi:10.1109/ICKG63256.2024.00060
43. Iapascurta V, Fiodorov I, Belii A, Bostan V. Multi-Agent Approach for Sepsis Management. *Healthc Inform Res*. 2025;31(2):209-214. doi:10.4258/hir.2025.31.2.209

 doi:10.20944/preprints202606.0640.v2

44. Chen YJ, Albarqawi A, Chen CS. Enhancing Clinical Decision-Making: Integrating Multi-Agent Systems with Ethical AI Governance. *arXiv*. Preprint posted online September 22, 2025:arXiv:2504.03699. doi:10.48550/arXiv.2504.03699
45. Ayub U, Naqvi SAA, Jajja SA, et al. A large language model (LLM)-based multi-agent framework for risk stratification and treatment recommendations in localized prostate cancer (locPCa). *J Clin Oncol*. 2025;43(16_suppl):5108-5108. doi:10.1200/JCO.2025.43.16_suppl.5108
46. Liu Z, Xiao L, He M, Zhu R, Yang H, Chen J. PICOAS: a clinical knowledge linking model for delivering up-to-date, interrelated, and personalized decision support. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2024:6589-6596. doi:10.1109/BIBM62325.2024.10821913
47. Liu S, Huang SS, McCoy AB, Wright AP, Horst S, Wright A. Optimizing Order Sets With a Large Language Model–Powered Multiagent System. *JAMA Network Open*. 2025;8(9):e2533277-e2533277. doi:10.1001/jamanetworkopen.2025.33277
48. Liu Q, Hu Z, Huang T, et al. EvoMDT: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *npj Digit Med*. 2026;9(1):124. doi:10.1038/s41746-025-02304-8
49. Ho AZT, Law JZF, Wang A, Lim DYZ, Ong JCL. DILIConsult: A Multi-Agent Large Language Model Framework for Evaluating Drug-Induced Liver Injury in ICU Settings. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*. 2026;46(4):e70131. doi:10.1002/phar.70131
50. Alam HMT, Srivastav D, Kadir MA, Sonntag D. Towards Interpretable Radiology Report Generation via Concept Bottlenecks using a Multi-Agentic RAG. In: Vol 15574. 2025:201-209. doi:10.1007/978-3-031-88714-7_18
51. Yi Z, Liu J, Xiao T, Albert MV. A Multi-Agent System for Complex Reasoning in Radiology Visual Question Answering. *arXiv*. Preprint posted online August 4, 2025:arXiv:2508.02841. doi:10.48550/arXiv.2508.02841
52. Zhang Z, Lee K, Jing P, et al. GEMA-Score: Granular Explainable Multi-Agent Scoring Framework for Radiology Report Evaluation. *arXiv*. Preprint posted online 2025. doi:10.48550/ARXIV.2503.05347
53. Bani-Harouni D, Navab N, Keicher M. MAGDA: Multi-agent Guideline-Driven Diagnostic Assistance. In: Deng Z, Shen Y, Kim HJ, et al., eds. *Foundation Models for General Medical AI*. Vol 15184. Lecture Notes in Computer Science. Springer Nature Switzerland; 2025:163-172. doi:10.1007/978-3-031-73471-7_17
54. Hu X, Huang J, Liu M, et al. FetalAgents: A Multi-Agent System for Fetal Ultrasound Image and Video Analysis. *arXiv*. Preprint posted online March 10, 2026:arXiv:2603.09733. doi:10.48550/arXiv.2603.09733
55. Ghezloo F, Seyfioglu MS, Soraki R, et al. PathFinder: A Multi-Modal Multi-Agent System for Medical Diagnostic Decision-Making Applied to Histopathology. *arXiv*. Preprint posted online February 13, 2025:arXiv:2502.08916. doi:10.48550/arXiv.2502.08916
56. Seyfioglu MS. Towards Autonomous Histopathological Diagnosis: An End-to-End Multi-Agent AI Framework for Diagnostic Decision-Making and Interpretation. https://hdl.handle.net/1773/52978
57. Weishaupt LL, Chen C, Williamson DFK, et al. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. *arXiv*. Preprint posted online March 26, 2026:arXiv:2506.20964. doi:10.48550/arXiv.2506.20964
58. Yi Z, Xiao T, Albert MV. A Multimodal Multi-Agent Framework for Radiology Report Generation. *arXiv*. Preprint posted online May 14, 2025:arXiv:2505.09787. doi:10.48550/arXiv.2505.09787
59. Li J, Zhou T, Zhou Z, et al. Experience-guided multi-agent interpretable framework for radiology report summarization. *Comput Methods Programs Biomed*. 2026;273:109078. doi:10.1016/j.cmpb.2025.109078
60. Tzanis E, Klontzas ME. ReclAIm: A Multi-Agent Framework for Monitoring and Correcting Performance Decline in Medical Imaging AI. *Radiol Artif Intell*. Published online June 3, 2026:e250923. doi:10.1148/ryai.250923
61. Wang D, Hu Z, Li Y, Xu D, Yang R, Zhuge C. MARTP: a multi-agent simulation framework for automated radiation therapy planning based on LLMs. *Phys Med Biol*. 2026;71(12):125037. doi:10.1088/1361-6560/ae6af6
62. Wang Z, Zhu Y, Zhao H, et al. ColaCare: Enhancing Electronic Health Record Modeling through Large Language Model-Driven Multi-Agent Collaboration. In: *Proceedings of the ACM on Web Conference 2025*. 2025:2250-2261. doi:10.1145/3696410.3714877

 doi:10.20944/preprints202606.0640.v2

63. Li R, Wang X, Berlowitz D, Mez J, Lin H, Yu H. CARE-AD: a multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes. *npj Digit Med*. 2025;8(1):541. doi:10.1038/s41746-025-01940-4
64. Gawade S, Akhouri S, Kulkarni C, et al. Multi Agent based Medical Assistant for Edge Devices. *arXiv*. Preprint posted online March 7, 2025:arXiv:2503.05397. doi:10.48550/arXiv.2503.05397
65. Yao Z, Chafekar T, Wang J, et al. ChatCLIDS: Simulating Persuasive AI Dialogues to Promote Closed-Loop Insulin Adoption in Type 1 Diabetes Care. *medRxiv*. Preprint posted online September 4, 2025:2025.09.02.25334973. doi:10.1101/2025.09.02.25334973
66. Mashatian S, Wung SF, Ritter A, et al. Responsible AI for Personalized Patient Education and Engagement Across Medical Conditions: Leveraging Multi-Agent LLMs, Ambient Technology, and NotebookLM—A Case Study in Diabetes Education and Limb Preservation. *Journal of the American Podiatric Medical Association*. 2026;116(3):30. doi:10.3390/japma116030030
67. Sun B, Die, Hu. CTG-Insight: A Multi-Agent Interpretable LLM Framework for Cardiotocography Analysis and Classification. *arXiv*. Preprint posted online 2025. doi:10.48550/ARXIV.2507.22205
68. Gao J, Rahman M, Caskey J, et al. MoMA: a mixture-of-multimodal-agents architecture for enhancing clinical prediction modelling. *npj Digit Med*. 2025;9(1):46. doi:10.1038/s41746-025-02219-4
69. Zhou R, Li C, Yang C, Lu J. ClinNoteAgents: An LLM Multi-Agent System for Predicting and Interpreting Heart Failure 30-Day Readmission from Clinical Notes. *AMIA Jt Summits Transl Sci Proc*. 2026;2026:594-603.
70. Yao T, Xu Y, Wang H, Qiu X, Althoefer K, Qi P. Multiagent Fuzzy Reinforcement Learning With LLM for Cooperative Navigation of Endovascular Robotics. *IEEE Trans Fuzzy Syst*. 2026;34(4):1109-1119. doi:10.1109/TFUZZ.2025.3585934
71. Zhao L, Bai J, Bian Z, et al. Autonomous Multi-Modal LLM Agents for Treatment Planning in Focused Ultrasound Ablation Surgery. *arXiv*. Preprint posted online July 15, 2025:arXiv:2505.21418. doi:10.48550/arXiv.2505.21418
72. Chen H, Gutt M, Belker OA, et al. Proof of concept for voice based MRI scanner control using large language models in real time guided interventions. *Sci Rep*. 2025;15(1):31206. doi:10.1038/s41598-025-11290-6
73. Fang C, Yue X, Zhao Z, Guo S. The Multi-Agentization of a Dual-Arm Nursing Robot Based on Large Language Models. *Bioengineering*. 2025;12(5):448. doi:10.3390/bioengineering12050448
74. Zhao Z, Yue X, Xie J, Fang C, Shao Z, Guo S. A Dual-Agent Collaboration Framework Based on LLMs for Nursing Robots to Perform Bimanual Coordination Tasks. *IEEE Robot Autom Lett*. 2025;10(3):2942-2949. doi:10.1109/LRA.2025.3533476
75. Ruiz Mejia JM, Rawat DB. MedScrubCrew: A Medical Multi-Agent Framework for Automating Appointment Scheduling Based on Patient-Provider Profile Resource Matching. *Healthcare*. 2025;13(14):1649. doi:10.3390/healthcare13141649
76. Lu M, Ho B, Ren D, Wang X. TriageAgent: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics; 2024:5747-5764. doi:10.18653/v1/2024.findings-emnlp.329
77. Han S, Choi W. 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. *Adv Artif Intell Mach Learn*. 2025;5(1):3261-3275.
78. Kim Y, Jeong H, Park C, et al. Tiered Agentic Oversight: A Hierarchical Multi-Agent System for Healthcare Safety. *arXiv*. Preprint posted online September 28, 2025:arXiv:2506.12482. doi:10.48550/arXiv.2506.12482
79. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. *Nature*. 2025;642(8067):442-450. doi:10.1038/s41586-025-08866-7
80. Ferber D, Hilgers L, Höper C, et al. Towards autonomous medical artificial intelligence agents. *Nature*. Published online June 17, 2026. doi:10.1038/s41586-026-10675-5
81. Li X, Yu H, Wang W, et al. DispatchMAS: fusing taxonomy and artificial intelligence agents for emergency medical services. *BMC Emerg Med*. 2026;26(1):78. doi:10.1186/s12873-026-01540-9

 doi:10.20944/preprints202606.0640.v2

82. Akinseloyin O, Jiang X, Palade V. Large language model-based multiagent collaboration for abstract screening toward automated systematic reviews. *Biology Methods and Protocols*. 2026;11(1):bpag006. doi:10.1093/biomethods/bpag006
83. Wu H, Zhu Y, Wang Z, et al. EHRFlow: A Large Language Model-Driven Iterative Multi-Agent Electronic Health Record Data Analysis Workflow. In: *KDD'24 Workshop: Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*. 2024. https://openreview.net/forum?id=lvycHSrFgk
84. Angulo J, Yeste V. AQAMS and AQAMS2: Multi Agent Systems for Biomedical Question Answering. In: 2025. https://api.semanticscholar.org/CorpusID:282385539
85. Israni M, Renuse S, V P. AutoMed: Multi-Agent AI System for Personalized Medical Knowledge Retrieval and Summarization. In: *2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*. IEEE; 2025:1-6. doi:10.1109/ICDSAAI65575.2025.11011656
86. Gorenshtein A, Shihada K, Sorka M, Aran D, Shelly S. LITERAS: Biomedical literature review and citation retrieval agents. *Computers in Biology and Medicine*. 2025;192:110363. doi:10.1016/j.compbiomed.2025.110363
87. Li H, Pan W, Rajendran S, Zang C, Wang F. EmulatRx: Empowering Clinical Trial Design with Agentic Intelligence and Real World Data. *medRxiv*. Preprint posted online March 3, 2026. doi:10.1101/2025.04.17.25326033
88. Yue L, Xing S, Chen J, Fu T. 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*. ACM; 2024:1-10. doi:10.1145/3698587.3701359
89. Furlan V, Moran J, Salazar NB, et al. EAGLE-AI: A large language model workflow for automated extraction and scoring of literature evidence linking genes to autism spectrum disorder. *medRxiv*. Preprint posted online June 22, 2026:2025.09.10.25334730. doi:10.1101/2025.09.10.25334730
90. Wysocki O, Wysocka M, Jacobo M, Unsworth H, Freitas A. Biomedical reasoning in action: Multi-agent System for Auditable Biomedical Evidence Synthesis. *arXiv*. Preprint posted online 2025. doi:10.48550/ARXIV.2510.05335
91. Livieratos A, Kudela M, Zhao Y, et al. MetaMind: A multi-agent transformer-driven framework for automated network meta-analyses. Wang C, ed. *PLoS One*. 2026;21(2):e0342895. doi:10.1371/journal.pone.0342895
92. Tao X, Dong X, Zhu Q, Zhou X. OEMA: ontology-enhanced multi-agent collaboration framework for zero-shot clinical named entity recognition. *Jamia Open*. 2026;9(3):ooag049. doi:10.1093/jamiaopen/ooag049
93. Chen Y, Wen B, Zulkernine F. A Multiagent Summarization and Auto-Evaluation Framework for Medical Text: Development and Evaluation Study. *JMIR AI*. 2025;4:e75932-e75932. doi:10.2196/75932
94. Wei H, Qiu J, Yu H, Yuan W. MEDCO: Medical Education Copilots Based on a Multi-agent Framework. In: Del Bue A, Canton C, Pont-Tuset J, Tommasi T, eds. *Computer Vision – ECCV 2024 Workshops*. Springer Nature Switzerland; 2025:119-135. doi:10.1007/978-3-031-91813-1_8
95. Sangwon KL, Zhang J, Steele R, et al. A Multi-AI Agent Framework for Interactive Neurosurgical Education and Evaluation: From Vignettes to Virtual Conversations. *Neurosurgery Practice*. 2026;7(2). doi:10.1227/neuprac.0000000000000217
96. Qu Y, Xu X, Long Y, Wang Y, Li J, Lu X. A Large Language Model–Powered Multiagent Framework Emulating Standardized Patients in Clinical Communication Skills Training: Development and Evaluation Study. *J Med Internet Res*. 2026;28:e84747-e84747. doi:10.2196/84747
97. Awasthi A, Chung BV, Vu AM, et al. MAARTA:Multi-agentic Adaptive Radiology Teaching Assistant. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part V*. Springer-Verlag; 2025:358-368. doi:10.1007/978-3-032-04971-1_34
98. Sangwon KL, Zhang J, Steele R, et al. Evaluating Large Language Model Diagnostic Performance on JAMA Clinical Challenges via a Multi-Agent Conversational Framework. *medRxiv*. Preprint posted online August 24, 2025:2025.08.20.25334087. doi:10.1101/2025.08.20.25334087

 doi:10.20944/preprints202606.0640.v2

99. Altermatt FR, Neyem A, Sumonte N, Mendoza M, Villagran I, Lacassie HJ. Performance of single-agent and multi-agent language models in Spanish language medical competency exams. *BMC Med Educ*. 2025;25(1):666. doi:10.1186/s12909-025-07250-3
100. Lim E, He YV, Joselowitz J, et al. MATRIX: Multi-Agent simulaTion fRamework for safe Interactions and conteXtual clinical conversational evaluation. *arXiv*. Preprint posted online 2025. doi:10.48550/ARXIV.2508.19163
101. Giuffre M, Kresevic S, Ajcevic M, Crocè L, Shung D. Large Language Model Agent-Based Framework for automated Treatment Prescription in Patients with Chronic Hepatitis C Virus Infection. *Digestive and Liver Disease*. 2025;57:S46-S47. doi:10.1016/j.dld.2025.01.088
102. Sabel J, Wingren M, Lundell A, et al. Medication counseling with large language models: balancing flexibility and rigidity. In: *2025 IEEE International Conference on Agentic AI (ICA)*. 2025:178-183. doi:10.1109/ICA67499.2025.00047
103. Stein S, Pilgermann M, Weber S, Sedlmayr M. Leveraging MDS2 and SBOM data for LLM-assisted vulnerability analysis of medical devices. *Comput Struct Biotechnol J*. 2025;28:267-280. doi:10.1016/j.csbj.2025.07.012
104. Chen Z, Peng Z, Liang X, et al. MAP: evaluation and multi-agent enhancement of large language models for inpatient pathways. *npj Health Syst*. 2026;3(1):37. doi:10.1038/s44401-026-00085-0
105. Lee Y, Wang X, Yang CC. Automated Clinical Problem Detection from SOAP Notes using a Collaborative Multi-Agent LLM Architecture. *arXiv*. Preprint posted online August 29, 2025:arXiv:2508.21803. doi:10.48550/arXiv.2508.21803
106. Ke Y, Yang R, Lie SA, et al. Mitigating Cognitive Biases in Clinical Decision-Making Through Multi-Agent Conversations Using Large Language Models: Simulation Study. *J Med Internet Res*. 2024;26:e59439. doi:10.2196/59439
107. Miao Y, Wen J, Luo Y, Li J. MedARC: Adaptive multi-agent refinement and collaboration for enhanced medical reasoning in large language models. *International Journal of Medical Informatics*. 2026;206:106136. doi:10.1016/j.ijmedinf.2025.106136
108. Zhan Z, Zhou S, Zhang R. Let large language models judge each other: multi-agent peer-reviewed reasoning for medical question answering. *J Am Med Inform Assoc*. Published online June 15, 2026:ocag042. doi:10.1093/jamia/ocag042
109. Liu PR, Bansal S, Dinh J, et al. MedChat: A Multi-Agent Framework for Multimodal Diagnosis with Large Language Models. In: *2025 IEEE 8th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. 2025:456-462. doi:10.1109/MIPR67560.2025.00078
110. Wang Q, Wang Z, Li M, et al. A feasibility study of automating radiotherapy planning with large language model agents. *Phys Med Biol*. 2025;70(7):075007. doi:10.1088/1361-6560/adbff1
111. De Maio C, Fenza G, Furno D, Grauso T, Loia V. Privacy-Preserving Healthcare Data Interactions: A Multi-Agent Approach Using LLMs. *JCOMSS*. 2025;21(1):13-22. doi:10.24138/jcomss-2024-0119
112. Chen K, Zhen T, Wang H, et al. MedSentry: Understanding and Mitigating Safety Risks in Medical LLM Multi-Agent Systems. *arXiv*. Preprint posted online May 27, 2025:arXiv:2505.20824. doi:10.48550/arXiv.2505.20824
113. Chan TK, Dinh ND. ENTAgents: AI Agents for Complex Knowledge Otolaryngology. *Otolaryngology*. Preprint posted online January 7, 2025. doi:10.1101/2025.01.01.25319863
114. Feng Y, Wang J, Zhou L, Lei Z, Li Y. DoctorAgent-RL: A Multi-Agent Collaborative Reinforcement Learning System for Multi-Turn Clinical Dialogue. *arXiv*. Preprint posted online October 14, 2025:arXiv:2505.19630. doi:10.48550/arXiv.2505.19630
115. Xia P, Wang J, Peng Y, et al. MMedAgent-RL: Optimizing Multi-Agent Collaboration for Multimodal Medical Reasoning. *arXiv*. Preprint posted online January 26, 2026:arXiv:2506.00555. doi:10.48550/arXiv.2506.00555
116. Zhuang Y, Jiang W, Zhang J, Yang Z, Zhou JT, Zhang C. Learning to Be A Doctor: Searching for Effective Medical Agent Architectures. *arXiv*. Preprint posted online August 15, 2025:arXiv:2504.11301. doi:10.48550/arXiv.2504.11301

 doi:10.20944/preprints202606.0640.v2

117. Croxford E, Gao Y, First E, et al. Automating Evaluation of AI Text Generation in Healthcare with a Large Language Model (LLM)-as-a-Judge. *medRxiv*. Preprint posted online October 15, 2025:2025.04.22.25326219. doi:10.1101/2025.04.22.25326219
118. Zhao H, Zhu Y, Wang Z, Wang Y, Gao J, Ma L. ConfAgents: A Conformal-Guided Multi-Agent Framework for Cost-Efficient Medical Diagnosis. *arXiv*. Preprint posted online August 6, 2025:arXiv:2508.04915. doi:10.48550/arXiv.2508.04915
119. Qiu J, Lam K, Li G, et al. LLM-based agentic systems in medicine and healthcare. *Nat Mach Intell*. 2024;6(12):1418-1420. doi:10.1038/s42256-024-00944-1
120. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*. 2024;6(6):e428-e432. doi:10.1016/S2589-7500(24)00061-X
121. *Ethics and Governance of Artificial Intelligence for Health: Large Multi-Modal Models. WHO Guidance*. 1st ed. World Health Organization; 2024. https://www.who.int/publications/i/item/9789240084759
122. Yagoubi FE, Mallah RA, Badu-Marfo G. AgentLeak: A Full-Stack Benchmark for Privacy Leakage in Multi-Agent LLM Systems. *arXiv*. Preprint posted online February 12, 2026:arXiv:2602.11510. doi:10.48550/arXiv.2602.11510
123. Prakash C, Lind M, Sisodia A. Agentic AI Governance and Lifecycle Management in Healthcare. *arXiv*. Preprint posted online January 22, 2026:arXiv:2601.15630. doi:10.48550/arXiv.2601.15630
124. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*. 2020;2(6):305-311. doi:10.1038/s42256-020-0186-1
125. Pati S, Kumar S, Varma A, et al. Privacy preservation for federated learning in health care. *Patterns*. 2024;5(7):100974. doi:10.1016/j.patter.2024.100974
126. Juneja G, Pasupulati JNS, Albalak A, Hua W, Wang WY. MAGPIE: A benchmark for Multi-AGent contextual PrIvacy Evaluation. *arXiv*. Preprint posted online October 16, 2025:arXiv:2510.15186. doi:10.48550/arXiv.2510.15186
127. Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *The Lancet Digital Health*. 2022;4(5):e384-e397. doi:10.1016/S2589-7500(22)00003-6
128. Nouis SC, Uren V, Jariwala S. Evaluating accountability, transparency, and bias in AI-assisted healthcare decision- making: a qualitative study of healthcare professionals' perspectives in the UK. *BMC Med Ethics*. 2025;26(1):89. doi:10.1186/s12910-025-01243-z
129. U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices. June 16, 2026. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
130. U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. August 18, 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence
131. U.S. Food and Drug Administration. Clinical Decision Support Software. January 29, 2026. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software
132. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
133. International Electrotechnical Commission. IEC 62304:2006+AMD1:2015 CSV: Medical device software - Software life cycle processes. 2015. https://webstore.iec.ch/en/publication/22794
134. International Organization for Standardization. ISO 14971:2019: Medical devices - Application of risk management to medical devices. ISO. 2019. https://www.iso.org/standard/72704.html
135. International Organization for Standardization. ISO 13485:2016: Medical devices - Quality management systems - Requirements for regulatory purposes. ISO. 2016. https://www.iso.org/standard/59752.html
136. International Organization for Standardization. ISO/IEC 42001:2023: Information technology - Artificial intelligence - Management system. ISO. 2023. https://www.iso.org/standard/42001

 doi:10.20944/preprints202606.0640.v2

137. Li H, Cheng X, Zhang X. Accurate Insights, Trustworthy Interactions: Designing a Collaborative AI-Human Multi-Agent System with Knowledge Graph for Diagnosis Prediction. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM; 2025:1-15. doi:10.1145/3706598.3713526
138. Tan SY, Sumner J, Wang Y, Wenjun Yip A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. *npj Digit Med*. 2024;7(1):192. doi:10.1038/s41746-024-01182-w
139. Roberts MC, Holt KE, Del Fiol G, Baccarelli AA, Allen CG. Precision public health in the era of genomics and big data. *Nat Med*. 2024;30(7):1865-1873. doi:10.1038/s41591-024-03098-0
140. Rehan MW, Rehan MM. Survey, taxonomy, and emerging paradigms of societal digital twins for public health preparedness. *npj Digit Med*. 2025;8(1):520. doi:10.1038/s41746-025-01737-5
141. Zhang K, Yang X, Wang Y, et al. Artificial intelligence in drug development. *Nat Med*. 2025;31(1):45-59. doi:10.1038/s41591-024-03434-4
142. Xu Z, Ren F, Wang P, et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nat Med*. 2025;31(8):2602-2610. doi:10.1038/s41591-025-03743-2
143. Tudor BH, Shargo R, Gray GM, et al. A scoping review of human digital twins in healthcare applications and usage patterns. *npj Digit Med*. 2025;8(1):587. doi:10.1038/s41746-025-01910-w
144. Kovatchev BP, Colmegna P, Pavan J, et al. Human-machine co-adaptation to automated insulin delivery: a randomised clinical trial using digital twin technology. *npj Digit Med*. 2025;8(1):253. doi:10.1038/s41746-025-01679-y
145. Rosenthal JT, Beecy A, Sabuncu MR. Rethinking clinical trials for medical AI with dynamic deployments of adaptive systems. *npj Digit Med*. 2025;8(1):252. doi:10.1038/s41746-025-01674-3
146. Butler PM, Yang J, Brown R, et al. Smartwatch- and smartphone-based remote assessment of brain health and detection of mild cognitive impairment. *Nat Med*. 2025;31(3):829-839. doi:10.1038/s41591-024-03475-9
147. Teodoro D, Naderi N, Yazdani A, Zhang B, Bornet A. A scoping review of artificial intelligence applications in clinical trial risk assessment. *npj Digital Medicine*. 2025;8(1):486. doi:10.1038/s41746-025-01886-7