Review

Not Just One Agent: 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: 9 June 2026

doi: 10.20944/preprints202606.0640.v1

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

 doi:10.20944/preprints202606.0640.v1

Review

Not Just One Agent: 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, Hong Kong
[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: stan.zq.li@westlake.edu.cn (S.Z.L.); yayingdu@hust.edu.cn (Y.D.); d202582769@hust.edu.cn (S.X.)
† 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. Medical multi-agent systems (MAS) shift AI from isolated answer generation toward workflow-level clinical intelligence by combining role specialization, memory, tool use, retrieval, communication, and orchestration. This Review maps 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 MAS should be validated not merely as answer engines, but as auditable, controllable workflow systems. Future work should prioritize traceable evidence chains, human oversight, privacy-preserving collaboration, standardized reporting, and prospective clinical validation.

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

1. Introduction

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]. When confronted with real-world medical workflows, a single model is still frequently constrained by one-shot reasoning, context congestion, knowledge cutoffs, factual hallucination, and the lack of executable workflows[1,4].What clinical settings truly require is a system that can

 doi:10.20944/preprints202606.0640.v1

continuously organize information around a patient's evolving 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; rather, it transforms LLMs from prompt-responsive tools into collaborative systems that use role specialization, communication, and workflow orchestration to decompose complex tasks, integrate multisource evidence, and generate executable conclusions. Evidence suggests that under clinically scaled mixed workloads, orchestrated MAS can maintain substantially higher accuracy with lower resource consumption, whereas single-agent performance deteriorates rapidly as task load accumulates[7]. MAS therefore represent not a simple aggregation of LLMs or single agents, but a key systems-level innovation that enables medical AI to enter clinical workflows (Table 1).

Table 1. Comparison of general-purpose LLMs, LLM-based single agents, and LLM-based MAS.

Feature	General-purpose LLM	LLM-based single agent	LLM-based MAS
Core positioning	General-purpose foundation model	Agent encapsulated around a single goal	Collaborative system composed of multiple role-based agents
Task execution	One-shot generation or short-chain execution	Can plan and invoke tools, but the same agent undertakes the main subtasks	Decomposes complex tasks across different agents and integrates results through orchestration/negotiation
Context management	Mainly single-turn or short-session context	Can maintain task-level context and short-term memory	Can maintain shared states, private memories, and cross-role context
External knowledge/tool use	Mainly relies on prompting or attached knowledge bases	Can actively retrieve information and call APIs or tools	Different agents can 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	Can reduce single-point errors through cross-review, voting, adjudication, and supervisory agents
Typical medical scenarios	Medical question answering, document	Lightweight decision support	MDT-like diagnosis and treatment, complex workflows,

 doi:10.20944/preprints202606.0640.v1

	generation, single-step explanation	and single-process automation	dynamic monitoring, evidence synthesis
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

This Review focuses on studies published over the past 3 years, from January 2023 to March 2026. We searched PubMed, Web of Science, Google Scholar, and the medRxiv and arXiv preprint servers, given their substantial influence on clinical medicine and the rapidly evolving field of medical AI. Eligible studies were required to meet the following criteria: they provided an explicit description of a multi-agent architecture, agentic orchestration, or role-based collaborative mechanism; the study task was directly relevant to clinical practice or healthcare-system applications; and the article reported system design, application scenarios, or comparable performance results. We primarily included English-language full-text articles and peer-reviewed studies, while also considering high-quality preprints that were representative of this fast-moving field. Studies focusing solely on single-agent systems, conventional multi-model ensembles without explicit role-based interaction, general MAS work without direct relevance to healthcare, and purely conceptual proposals lacking concrete system implementation or evaluation were outside the scope of this Review.

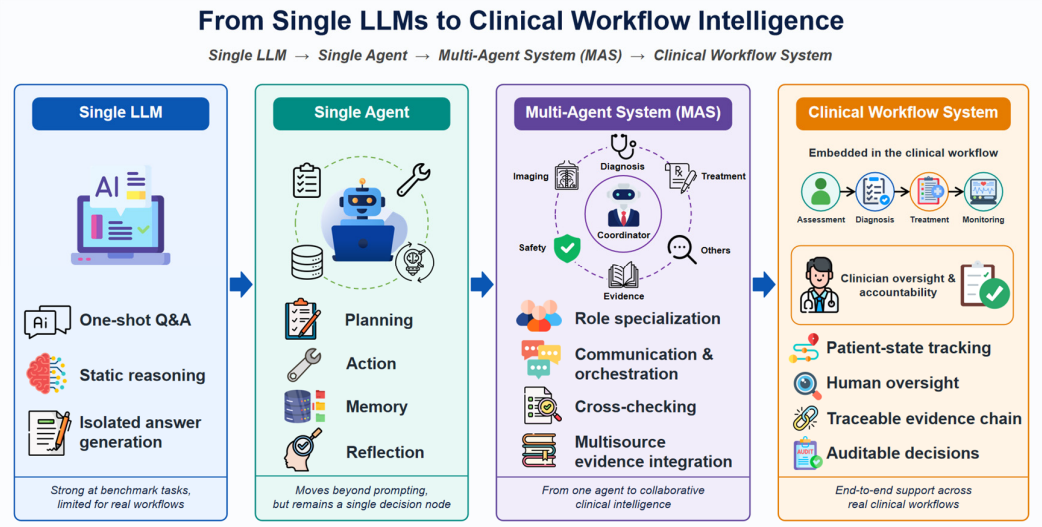

Figure 1. From Single LLMs to Clinical Workflow Intelligence. Medical AI is progressing from isolated one-shot answer generation to agentic systems that can decompose clinical tasks, use tools, update patient context, and iteratively check outputs. Multi-agent systems further extend this transition through 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. Major Application Scenarios of Multi-Agent Systems in Clinical Medicine

2.1. Clinical Applications in Medicine

LLM-based multi-agent systems (MAS) are extending beyond the capability boundaries of single agents, evolving from early medical "chatbots" into systematic applications that can be embedded across the clinical continuum, support multi-role collaboration, and decompose complex tasks. The

 doi:10.20944/preprints202606.0640.v1

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 can usually process tasks only in a serial, single-task manner, MAS coordinate multiple independent agents with dedicated functional roles, simulate the working patterns of clinical teams, and decompose complex medical tasks into subprocesses that closely align with real clinical workflows. In this way, MAS can support full-spectrum 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, MAS can support iterative diagnostic optimization during multi-turn dialogues. This not only aligns with the logic of real clinical practice, but also shows advantages over both human physicians and single agents in diagnostic accuracy and resource-use efficiency[8,9]. In 16 neurological diagnostic cases, the Gregory system achieved 100% diagnostic accuracy, exceeding the average performance of human neurologists (83%) while further reducing diagnostic cost and time[8]. MedAgentSim, a multi-agent system involving patient, doctor, and measurement agents, achieved 79.5% diagnostic accuracy on a real-world clinical dataset, nearly doubling the performance of a single-agent system[9]. For differential diagnosis among diseases with similar symptoms and signs, MAS can configure multiple domain-expert agents for multilayered debate, uncover subtle disease differences, and improve differential diagnostic precision[10].

For rare diseases and complex cases, the objective of MAS is to extend the boundaries of physicians' knowledge and experience by using multi-role collaboration to broaden the coverage of clinical clues and strengthen the diagnostic evidence chain[11]. In rare genetic diseases, multi-agent systems can quantitatively evaluate and rank the pathogenic likelihood of candidate genes, thereby prioritizing the true disease-causing genes for clinical teams and addressing the clinical pain points of inefficient manual screening and high missed-detection rates across large candidate-gene spaces[12,13]. 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 systematically overcomes the knowledge limitations of a single agent and improves the accuracy of pathogenic gene ranking[13]. Rare diseases involve vast candidate spaces and limited clinical experience, and therefore require stronger evidence integration and interpretability. Rare-disease diagnostic MAS can simultaneously process free-text histories, phenotypic terms, and sequencing files to output candidate diseases together with verifiable evidence chains[14]. 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[11,15], 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[16–18], thereby reducing the risk of information omission and cognitive bias in complex or atypical cases and producing more consistent and traceable conclusions.

 doi:10.20944/preprints202606.0640.v1

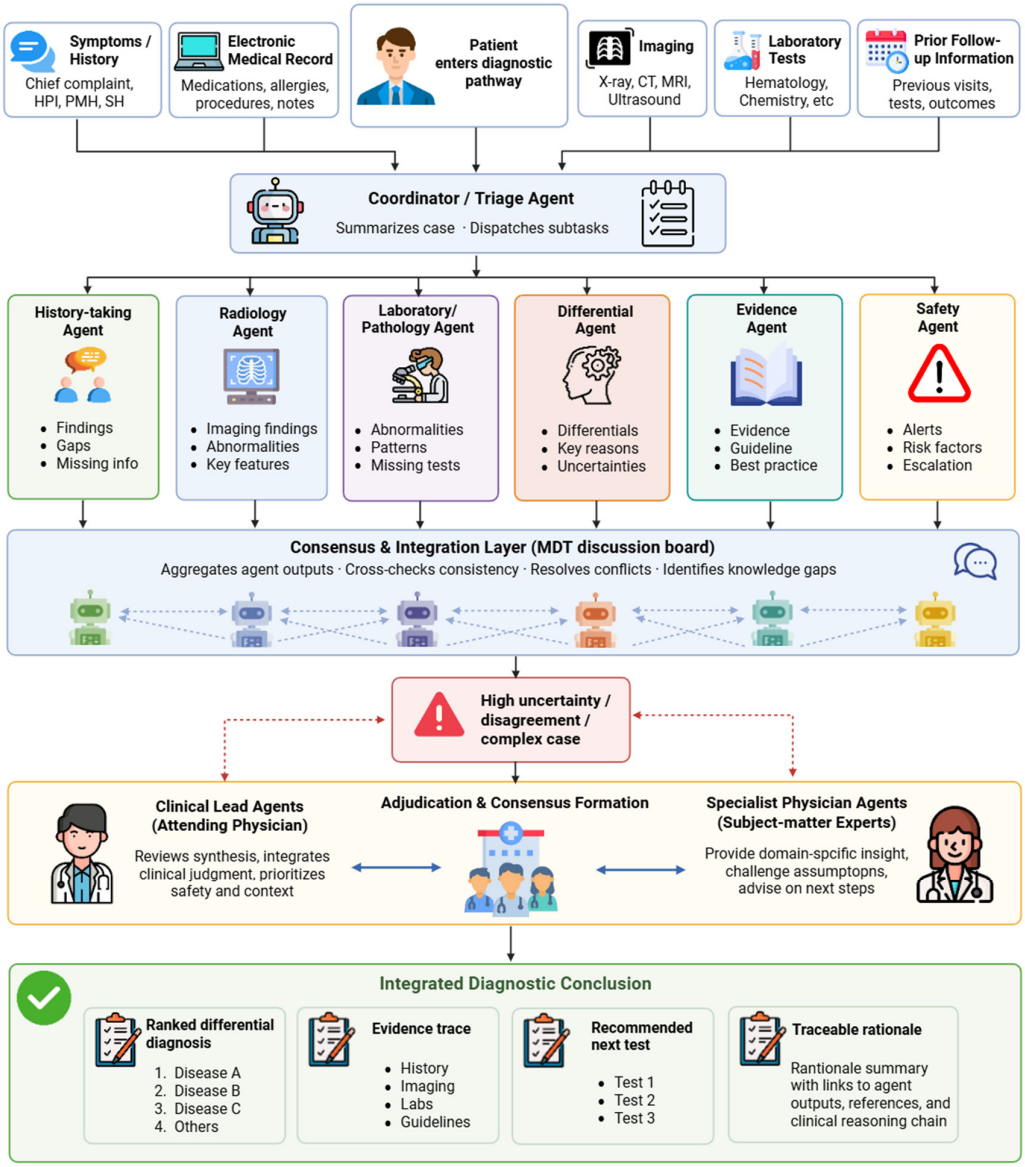

Figure 2. MDT-style multi-agent systems for diagnosis and differential diagnosis. Patient information, including symptoms, medical history, electronic health records, imaging, laboratory tests, and follow-up data, is first summarized by a coordinator/triage agent and distributed to role-specialized agents. 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 gradually becoming an important foundation 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 agents in complex clinical decision-making through multi-role collaboration, planning, and tool use[19]. Through specialized role assignment and collaborative reasoning, MAS can automatically

 doi:10.20944/preprints202606.0640.v1

integrate patients' multimodal clinical information—including electronic health records, imaging data, laboratory tests, and follow-up feedback—and precisely align individual clinical characteristics with authoritative guideline recommendations and the latest evidence-based medical findings. Ultimately, this enables the generation of updated and more appropriate individualized treatment plans for patients in different clinical scenarios[19,20]. In time-sensitive settings such as infectious diseases and critical care, MAS can divide clinical tasks across "antibiotic recommendation-guideline checking-full-process sepsis management." Dedicated agents can separately perform regimen formulation, evidence-based validation, global condition assessment, treatment planning, and dynamic monitoring of clinical deterioration[21]. With retrieval-augmented generation, they can also integrate patient-specific clinical data and guideline evidence in real time, thereby shortening treatment decision time in high-risk settings such as the ICU[21]. A comparative validation study by Chen et al. [22] in ICU critical-care scenarios further confirmed that, compared with a single-agent system, a multi-agent system improved the accuracy of in-hospital mortality prediction ($p = 0.0001$) and ICU length-of-stay prediction ($p < 0.0001$), demonstrating statistically significant advantages for core outcome prediction in critical care. In oncology, the MAS constructed by Ayub et al. [23] used role-based collaboration for risk stratification and treatment recommendation, converting key stratification evidence from unstructured reports into executable treatment-pathway recommendations to improve the efficiency and reliability of treatment selection. In chronic diseases requiring long-term dynamic management[19], MAS can use external tool calling and retrieval augmentation to collect patient clinical data and the latest evidence-based literature in real time, extract and integrate 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[19], 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[24], MAS can link disease-specific clinical guidelines and automatically identify and avoid drug interactions and treatment contraindications arising from conflicting recommendations across guidelines, addressing a central pain point in the management of comorbid conditions. MAS can also optimize medical decision-making by automatically identifying outdated, inappropriate, or missing items in order sets and recommending improvements, thereby helping physicians rapidly issue standardized orders[25]. By distributing treatment decision subtasks to different agents—such as treatment recommendation, guideline checking, and follow-up monitoring—and integrating outputs through collaboration and feedback, MAS can produce clinically actionable therapeutic recommendations that more closely approximate real clinical team workflows and accountability structures.

2.1.3. Medical Imaging

MAS show substantial potential in medical imaging and have been widely applied to X-ray imaging[26–29], CT[28,29], MRI[18], ultrasonography[30], and pathological slide review[31–33], providing automated and highly precise services that improve diagnostic efficiency and accuracy. In automated radiology report generation for imaging modalities such as chest X-rays, the MAS developed by Yi et al. [34] and Li et al. [35] 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. [26] 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 clinically relevant imaging reports accordingly. This modular multi-role design creates bidirectional anchoring and stable alignment between diagnostic conclusions and visual evidence, reduces generative hallucination through multistep

 doi:10.20944/preprints202606.0640.v1

factual verification and source tracing, and presents the complete reasoning path from imaging features to diagnostic conclusions, thereby substantially enhancing the clinical credibility and interpretability of radiology reports. In ultrasound imaging, the FetalAgents MAS uses collaborative role specialization to automate eight clinical tasks, including ultrasound-plane classification, structural segmentation, and biometric measurement, overcoming the limitation of traditional models that support only static images and enabling ultrasound video-stream analysis and standardized clinical report generation[30]. Automatically generated radiology reports must be evaluated across clinical accuracy, information completeness, and language standardization. MAS can perform quality control throughout the report-generation process[28] and transform judgments of clinical usability into quantifiable scores through comparison with standard reports, evaluation of core clinical dimensions, and continuous monitoring of model performance. In diagnostic efficiency, imaging MAS can rapidly identify key image features and next-step management options around specific clinical questions, reducing repeated communication and delayed decisions[27]. They may also mitigate radiologist shortages and diagnostic pressure in primary care or resource-constrained settings, improve the efficiency of interpreting routine examinations such as chest radiographs, and support triage and decisions regarding further testing[29]. In pathological slide review, MAS collaboratively perform triage, navigation, description, and diagnosis across the full workflow[31,33]. 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[32].

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, thereby identifying acute clinical deterioration earlier and supporting timely intervention[22,36]. MAS can also combine unstructured text from longitudinal electronic health records with collaborative evaluations by specialty agents to capture subtle prodromal signals of disease, enabling earlier risk prediction for chronically progressive conditions such as Alzheimer's disease and providing a new technical pathway for early screening and diagnosis[37]. In post-discharge follow-up and chronic disease management, MAS can support continuous monitoring across the full care cycle by incorporating remote monitoring data, post-treatment response assessments, patient-reported symptoms, and medication behaviors into analysis[38]. When combined with multimodal information from electronic health records, imaging, and laboratory tests, MAS can perform dynamic risk assessment, achieve earlier warnings, enable safer optimization of therapeutic strategies, and construct a complete "monitoring-warning-adjustment-follow-up" loop[19]. For chronic diseases such as type 1 diabetes, in which low willingness to adopt technology and insufficient treatment adherence remain core clinical barriers, the MAS developed by Yao et al. can simulate evidence-based persuasive doctor-patient conversations to deliver personalized patient education, psychological support, and long-term behavioral intervention, thereby addressing key bottlenecks in the real-world implementation of out-of-hospital chronic disease management[39]. 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[38,40]. Through hospital-to-home data integration and dynamic risk assessment, MAS can construct proactive, full-process patient monitoring models, shifting traditional monitoring from passive alerts toward forward-looking clinical decision-making and promoting the transition to intelligent closed-loop care.

 doi:10.20944/preprints202606.0640.v1

2.1.5. Surgical Assistance and Surgical Robotics

The application of MAS in intraoperative assistance and surgical robotics is evolving into a clinical intelligence framework that can be deeply embedded in surgical team collaboration and adapted to the full spectrum of surgical care. Typical application scenarios include intraoperative localization and navigation with coordinated instrument control[41] and surgical plan formulation with perioperative safety management[42]. 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 minimally invasive and complex surgical procedures and provide a clear technical foundation and translational potential for reducing perioperative adverse events and postoperative complications[41,42]. 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[43]. 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 has important application value for high-frequency scenarios in operating rooms and wards, such as instrument and material sorting, delivery, and human-robot collaborative operations[44,45]. The deeper application of MAS can simultaneously improve the operational precision of complex surgery and perioperative safety, while also providing a new technical pathway for upgrading surgical and nursing robots and extending their clinical use cases.

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 can also structurally integrate multisource evidence-based information and clinical data, providing reliable data support for refined hospital management and clinical research. In addition, supporting applications of medical MAS can 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).

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

 doi:10.20944/preprints202606.0640.v1

optimization of healthcare workflows and comprehensive improvement of institutional operational efficiency[46]. In high-load and time-sensitive triage scenarios such as emergency care[47,48], MAS can use role-based collaboration to integrate standardized triage scales, guideline retrieval, and multi-turn interactive decision-making, thereby improving the efficiency and accuracy of triage grading while more consistently reducing triage inconsistency compared with single agents. TRIAGEAGENT[47] improves bias control and grading precision in clinical triage through group classification, confidence aggregation, and consensus decision-making. Han et al.'s[48] multi-agent emergency decision-support system further showed that a collaborative framework simulating the role division of an emergency team not only helps 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 can use modular collaboration to standardize and integrate prehospital information, enabling seamless connection between prehospital and in-hospital care processes[38], lowering barriers to care, and improving continuity of care. The implementation of clinical workflow automation must be grounded in full-process quality and safety control. Built-in evaluation and supervision agents can dynamically monitor system outputs across the entire cycle, reducing the risk that propagated errors or clinically infeasible recommendations enter the workflow[49,50]. MAS provide a new technical model for hospital workflow automation, improving institutional operational quality and efficiency while constructing a full-cycle safety risk-control framework for clinical processes.

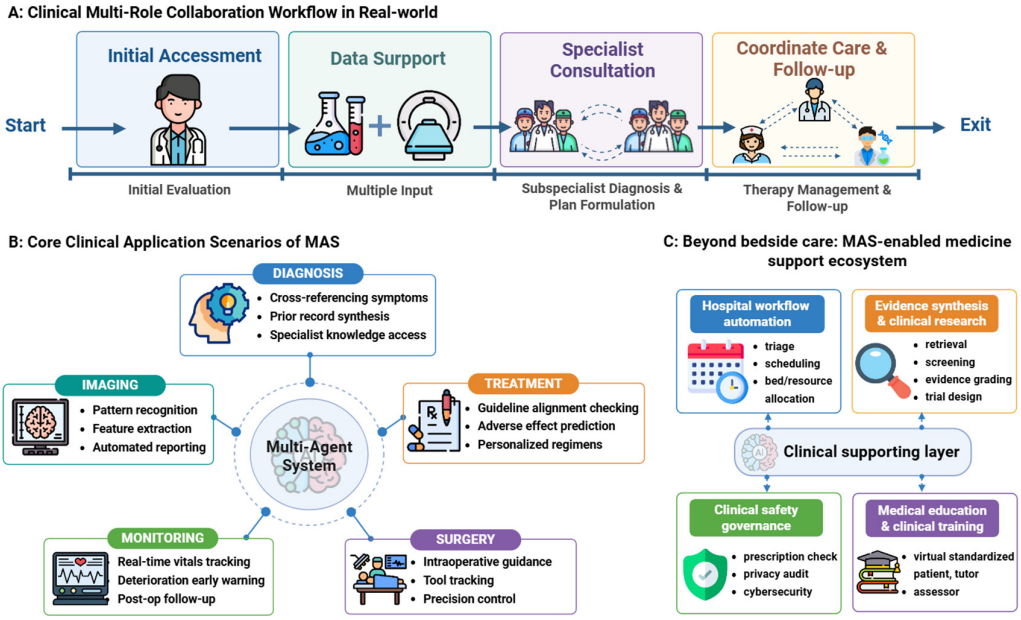

Figure 3. Major application scenarios of medical multi-agent systems. (A) MAS simulate real-world clinical teamwork, spanning initial assessment, multimodal data support, specialist consultation, and coordinated treatment follow-up. **(B)** In core clinical care, MAS support diagnosis, treatment, imaging, 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.

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-

 doi:10.20944/preprints202606.0640.v1

quality grading, as well as the difficulty of producing standardized evidence outputs that can be directly reused by clinicians and researchers[51,52]. Through role-based division of labor among agents, MAS can automate the entire workflow of evidence synthesis: dedicated agents separately perform literature retrieval and evidence collection, data screening and information summarization, and citation/source verification. Multi-round consistency checks and result validation improve output stability and methodological rigor, ultimately generating standardized evidence-based text for systematic reviews, guideline development, and clinical question answering[53,54]. In research contexts requiring high levels of evidence, MAS can include independent quality assessment and supervision agents[55], strictly follow evidence-grading standards throughout the process, and use repeated checks to ensure citation traceability and high-quality sources, thereby avoiding core problems such as fabricated citations. In key research scenarios such as clinical trial design and analysis, MAS have begun to show clear application potential. TrialGenie[56] 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, thereby improving the efficiency and iteration capacity of trial design. ClinicalAgent[57] 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 outperforms direct prompt-based LLM methods in trial outcome prediction and provides automated technical support for evidence-based decision-making across the clinical trial life cycle. In clinical genomics and precision medicine, MAS can automatically integrate multisource evidence such as genetic variants and pharmacogenomics, perform evidence scoring and interpretation based on standardized rules, and support individualized diagnosis and treatment[58,59]. In time-sensitive emergency and critical-care scenarios such as sepsis, MAS can precisely align patient clinical data with the latest treatment guidelines and research literature to rapidly generate guideline-concordant clinical decision recommendations[21]. At the same time, automated systems based on multi-agent architecture can complete clinical RCT literature retrieval, PICO element extraction, and network meta-analysis statistical modeling, compressing a traditional 3- to 4-month clinical evidence-synthesis process to within one week and providing efficient evidence-based technical support for emergency/critical care and clinical research[60].

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[61]. Such systems can simulate real clinical training environments and materially improve learners' history-taking and communication skills, multidisciplinary team (MDT) collaboration, and integrated diagnostic competence. In medical education, MAS can transform static knowledge-point question banks into dynamic virtual clinical conversations[62], allowing learners to complete more clinically realistic diagnostic training and standardized assessment through real-time interaction with virtual standardized patients. This approach can be adapted to specialty residency training and stage-based competency assessment. In radiology education, MAS have shown stronger teaching-assistance effects than single agents. The system developed by Awasthi et al. can systematically compare radiology trainees' eye-tracking patterns, report text, and expert diagnostic standards, precisely identify cognitive blind spots and diagnostic biases during image interpretation, and generate personalized teaching guidance. In diagnostic error classification, it achieved

 doi:10.20944/preprints202606.0640.v1

substantial improvements in accuracy and F1 score compared with a single agent[63]. Beyond core scenario-based clinical training, MAS can also support teaching-management tasks such as quality evaluation of question banks and assessment of teaching-feedback effectiveness[64,65]. 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[66], 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 can reduce the risk of clinical medication errors through a standardized closed loop of recommendation generation-indication verification-screening for drug contraindications and interactions-output of a clinical review summary[67]. They can also proceduralize core steps such as order-compliance review, evidence-based medical knowledge retrieval, and drug-indication verification to reduce the entry of outdated or incorrect orders into clinical pathways[25]. By preferentially retaining high-quality recommendations and introducing a second-confirmation mechanism, they can mitigate context drift, omission of key information, and error accumulation in long and complex clinical interactions, thereby reducing prescription deviations caused by these problems[68]. 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[38]. For cybersecurity protection of medical devices, MAS can automatically 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 exploitation pathways, precisely identify medical-device security vulnerabilities, and output targeted defense optimization recommendations[69], thereby protecting the cybersecurity and stable operation of in-hospital devices.

3. Multi-Agent System Architectures for Clinical Workflows

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[8]. Inpatient pathway support has been organized as a continuous collaboration chain[70]. SOAP-note clinical problem detection uses a collaborative format in which a manager organizes expert discussion[71]. Order-set optimization further assigns content critique, dynamic retrieval, knowledge retrieval, and medication checking to different nodes[25]. 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.

 doi:10.20944/preprints202606.0640.v1

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[71].

Furthermore, different tasks require different depths of collaboration. For rule-based, lower-risk subtasks, lightweight routing and specialized execution nodes are often sufficient[72]. Pharmacy consultation scenarios show that multi-role peer review can improve self-assessment and safety without substantially expanding the workflow[68]. By contrast, complex cases with high uncertainty and high consequences are better suited to multi-role checking and escalation mechanisms[73]. 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[18]. 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[18]. In-hospital rules and local knowledge bases determine whether recommendations are executable[25]. Pharmacy consultation scenarios demonstrate that drug labels, contraindications, and dialogue norms can also be independently handled by dedicated nodes[68]. The combination of structured EHRs and medical knowledge retrieval allows patient state and general medical evidence to be used distinctly within the same workflow[36]. Longitudinal clinical notes further add patient-specific information along the temporal dimension[37]. 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[18]. 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[36]. 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[25]. 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[37]. Memory modules in treatment planning similarly indicate that if a system cannot continuously preserve previous strategies and intermediate results, each iteration may

 doi:10.20944/preprints202606.0640.v1

regress to repeatedly restarting the same case[42]. 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[36]. 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 translation of visual results into clinical language[74]. MAS for radiotherapy planning and focused ultrasound treatment planning organize task decomposition, parameter optimization, outcome assessment, and plan revision into continuous feedback chains[42,75]. 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

 doi:10.20944/preprints202606.0640.v1

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[70]. Order-set optimization must be embedded in clinical decision-support environments[25]. Radiotherapy treatment planning depends on specialized software interfaces such as treatment planning systems (TPS) [75]. FUAS-Agents are likewise built on specific toolchains[42]. 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[36]. 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[76]. 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[25,77]. More importantly, malicious nodes, communication contamination, and structural vulnerabilities in multi-agent topologies may expand local errors into systemic risks[77]. 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 termination conditions. Existing studies generally follow two paths. One uses language-based reflection or expert feedback to correct local behaviors within established workflows[56,78]. 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[79,80]. 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[15,81]. 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[79]. 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[80]. Third, existing studies rely heavily on synthetic patients and proxy evaluators[79]. Architecture search is still mainly validated on controlled benchmarks[81]. Improvements in self-evolving consultation frameworks are also largely based on limited case

 doi:10.20944/preprints202606.0640.v1

accumulation[15]. 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 will not be determined by the best score on a single benchmark, but by whether they can simultaneously satisfy more fundamental requirements: workflow alignment, evidence traceability, risk governability, and accountable explanations (Figure 4).

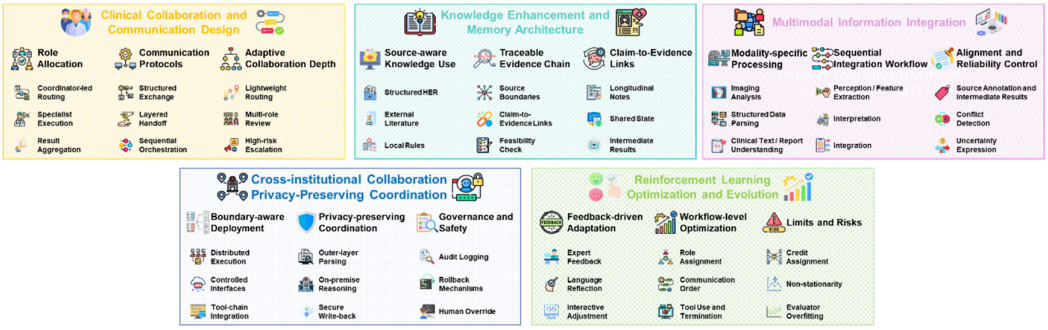

Figure 4. Overview of multi-agent system architecture for clinical workflows. The framework organizes clinical multi-agent systems into five interconnected dimensions: collaboration and communication design, knowledge-enhanced evidence and memory architecture, multimodal information integration, cross-institutional 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 therefore should not depend solely on the final answer or on whether a particular node produces a more fluent response, but on whether the entire collaboration chain can still deliver more stable judgments, clearer evidence chains, and a more governable risk structure after introducing additional 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[72] (Figure 5).

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

 doi:10.20944/preprints202606.0640.v1

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[72]. At the same time, evaluation standards are moving closer to the dimensions used in real hospital review. Order-set 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[25]. 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[70]. 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."[82] Thus, medical MAS are better evaluated through a framework that links outcome-level, process-level, and deployment-level metrics.

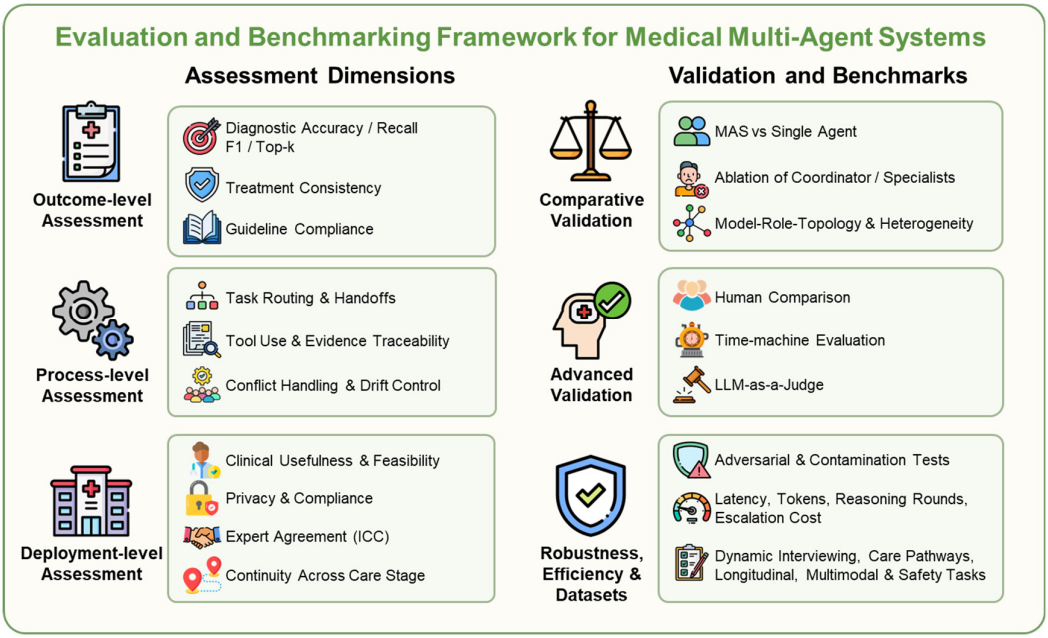

Figure 5. Overview of the evaluation and benchmarking framework for medical multi-agent systems. The left side summarizes outcome-, process-, and deployment-level assessment dimensions. The right side highlights comparative and advanced validation methods together with robustness, efficiency, and representative benchmark datasets and tasks.

4.2. 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 show that when retrieval, extraction, and calculation are assigned to different workers, the system can maintain greater stability under high mixed-task pressure. This improvement primarily comes from task isolation and orchestration structure, rather than simply from a larger model backbone[72]. Inpatient pathway studies also suggest that the advantage of MAS

 doi:10.20944/preprints202606.0640.v1

lies not only in endpoint scores, but also in tighter alignment with clinical workflows across triage, examination, diagnosis, and treatment recommendations[70].

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. TrialGenie 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[56]. 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[80]. 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, TrialGenie 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[8,56]. 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[64,82].

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-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[37]. 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[70]. 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[51]. 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[81].

 doi:10.20944/preprints202606.0640.v1

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[8,62]. 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[25]. In addition, LLM-as-a-Judge has shown practical value in evaluating the quality of medical text generation. It can automatically score factual correctness, completeness, and clinical utility, and use metrics such as ICC to quantify consistency between system ratings and expert ratings, providing a low-cost and quantifiable supplement for large-scale predeployment review[82].

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 evidence were sufficient to support early decision-making[37]. 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[64]. 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[8,64]. 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.” [37] Traceable rare-disease diagnosis studies further require that candidate conclusions and evidence chains be stored together, enabling evaluators to verify whether final conclusions truly arise from intermediate evidence[14].

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[77]. 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[77]. 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[18].

 doi:10.20944/preprints202606.0640.v1

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[72]. 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[8,64]. 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[83].

In terms of dataset and benchmark composition, current evaluations of medical MAS are broadly organized around four categories of tasks. Dynamic inquiry and progressive information-disclosure tasks test information acquisition and interaction organization[64]. Inpatient pathway tasks assess stability across multistage role transitions[70]. Longitudinal disease-course tasks evaluate shared states and early prediction ability[37]. Safety evaluation frameworks test vulnerability and risk diffusion across different topologies[77]. Cost-sensitive diagnostic tasks emphasize trade-offs between operational cost and escalation strategies[83]. Multimodal reasoning tasks further assess cross-modal integration[80]. 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 more ideal future evaluation framework should simultaneously preserve outcomes, routing logs, tool invocations, evidence chains, disagreement states, and resource consumption, so that why MAS are better can be directly compared rather than inferred indirectly from final scores alone (Table 2).

Table 2. Core evaluation datasets and benchmarks: task types, evaluation focuses, and representative references 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	[64]
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	[56,70,72,83]

 doi:10.20944/preprints202606.0640.v1

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	[82]
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	[37]
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	[25]
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	[72]

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 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.

 doi:10.20944/preprints202606.0640.v1

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[73]. 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[25]. Safety evaluations further show that once upstream nodes are contaminated or misled, local errors may diffuse along the collaborative topology into systemic risks[77].

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[72]. 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[51]. 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[80]. 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[81].

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[37]. 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[76]. 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[14].

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 into systemic issues running through the entire collaboration chain[84–87]. 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

 doi:10.20944/preprints202606.0640.v1

invocations. Governance therefore no longer targets only a single model, but the entire orchestration process and its life cycle[85,87–89].

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[85,86,88]. AgentLeak shows that the overall privacy exposure of MAS is approximately 1.6 times that of single-agent systems, indicating that risk has extended from output-level leakage to continuous exposure within internal data flows[88]. 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[88–92].

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[85,86]. An inappropriate clinical recommendation may simultaneously involve task allocation, evidence retrieval, specialty reasoning, result verification, and workflow design[85,89,93]. This structure more closely resembles real clinical collaboration, but it also makes transparency, interpretability, and accountability boundaries more ambiguous[85,87]. 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, and evidence-alignment mechanisms that improve error localization and result explanation, as well as approaches inspired by medical algorithmic auditing to continuously evaluate error patterns, vulnerable links, and post-deployment drift[89,93–95].

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[85,87,93]. 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[87]. 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 [86,93–95] (Figure 6).

 doi:10.20944/preprints202606.0640.v1

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," but "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 three directions: 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 [16]. 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 [76]. Thus, the emphasis of future "federated self-evolving collaboration" lies not only in using distributed updating and parameter-sharing mechanisms to alleviate cross-institutional data silos [96], but also in combining differential

 doi:10.20944/preprints202606.0640.v1

privacy and secure aggregation to reduce raw patient-data exposure while maintaining system performance and collaboration efficiency as much as possible [97]. 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 [76].

Second, 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 [98]. Traceable rare-disease diagnostic systems further demonstrate that MAS become genuinely auditable only when conclusions and intermediate evidence are preserved together [14]. 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 [25]. The human-in-the-loop (HITL) mechanism embodied by TrialGenie 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 [56]. 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 [99–106]. For medical MAS, this trend means that their application focus will gradually expand into systematic support for cross-institutional, cross-scale, and cross-modal data collaboration.

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 [99]. Remote cognitive assessment based on smartphones and smartwatches has been validated as feasible in large populations, suggesting that early screening and diagnosis can be performed in daily-life settings [107]. 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 [100]. Population-level studies have also extended applications to infection control, immunization planning, long-term health modeling, and resource optimization [101]. If medical MAS can connect data from wearable devices, home monitoring, internet hospitals, and community health centers, their value will increasingly 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 [102]. Studies have shown that the generative AI-discovered TNIK inhibitor rentosertib has entered human studies in patients with idiopathic pulmonary fibrosis [103]. Digital twins are being defined as individualized, dynamically updated, and predictive patient models, but only a minority of current studies truly meet this standard [104]. 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%[105], indicating that digital twins are beginning to move from conceptual models into the clinical frontier of intervention optimization and trial evaluation [104,105]. At the clinical-trial level, recent studies have proposed frameworks for dynamic deployment and continuous clinical validation [106], and AI is

 doi:10.20944/preprints202606.0640.v1

increasingly being applied to the assessment of trial safety, efficacy, and operational risks [108]. These advances suggest that future MAS may not be limited to assisting decision-making, but may be more deeply applied to medical discovery and clinical translation.

7. Conclusions

Medical MAS are moving medical AI from model applications oriented toward single tasks toward systems-level collaboration embedded in clinical workflows. This comprehensive 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 demonstrated stronger task adaptability and expansion 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 will further 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, 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 real healthcare systems and may profoundly reshape medical practice, medical research, and healthcare service delivery in the years ahead.

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).

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.

Declaration of interests: The authors declare no competing interests.

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. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *The new england journal of medicine*. Published online 2023.
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
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

 doi:10.20944/preprints202606.0640.v1

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. Sorka M, Gorenshtein A, Abramovitch H, Soontrapa P, Shelly S, Aran D. AI vs Human Performance in Conversational Hospital-Based Neurological Diagnosis.
9. Almansoori M, Kumar K, Cholakkal H. Self-Evolving Multi-Agent Simulations for Realistic Clinical Interactions. Published online 2025. doi:10.48550/ARXIV.2503.22678
10. Zhao Y, Wang H, Zheng Y, Wu X. A Layered Debating Multi-Agent System for Similar Disease Diagnosis.
11. 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
12. Neeley MB, Mao D. Survey and Improvement Strategies for Gene Prioritization with Large Language Models.
13. Zhou X, Ren Y, Zhao Q, et al. An LLM-Driven Multi-Agent Debate System for Mendelian Diseases. Published online April 11, 2025. doi:10.48550/arXiv.2504.07881
14. Zhao W, Wu C, Fan Y, et al. An Agentic System for Rare Disease Diagnosis with Traceable Reasoning. Published online February 16, 2026. doi:10.48550/arXiv.2506.20430
15. 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. Published online March 18, 2025. doi:10.48550/arXiv.2503.13856
16. Esteitieh Y, Mandal S, Laliotis G. Towards Metacognitive Clinical Reasoning: Benchmarking MD-PIE Against State-of-the-Art LLMs in Medical Decision-Making. Published online January 29, 2025. doi:10.1101/2025.01.28.25321282
17. 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
18. Peng Q, Cui J, Xie J, Cai Y, Li Q. Tree-of-Reasoning: Towards Complex Medical Diagnosis via Multi-Agent Reasoning with Evidence Tree. Published online August 5, 2025. doi:10.48550/arXiv.2508.03038
19. 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
20. 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
21. 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
22. Chen YJ, Albarqawi A, Chen CS. Enhancing Clinical Decision-Making: Integrating Multi-Agent Systems with Ethical AI Governance. Published online September 22, 2025. doi:10.48550/arXiv.2504.03699
23. 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).
24. 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
25. 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
26. 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
27. Yi Z, Liu J, Xiao T, Albert MV. A Multi-Agent System for Complex Reasoning in Radiology Visual Question Answering. Published online August 4, 2025. doi:10.48550/arXiv.2508.02841
28. Zhang Z, Lee K, Jing P, et al. GEMA-Score: Granular Explainable Multi-Agent Scoring Framework for Radiology Report Evaluation. Published online 2025. doi:10.48550/ARXIV.2503.05347

 doi:10.20944/preprints202606.0640.v1

29. 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
30. Hu X, Huang J, Liu M, et al. FetalAgents: A Multi-Agent System for Fetal Ultrasound Image and Video Analysis. Published online March 10, 2026. doi:10.48550/arXiv.2603.09733
31. Ghezloo F, Seyfioglu MS, Soraki R, et al. PathFinder: A Multi-Modal Multi-Agent System for Medical Diagnostic Decision-Making Applied to Histopathology. Published online February 13, 2025. doi:10.48550/arXiv.2502.08916
32. Chen C, Weishaupt LL, Williamson DFK, et al. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology.
33. Seyfioglu MS. Towards Autonomous Histopathological Diagnosis: An End-to-End Multi-Agent AI Framework for Diagnostic Decision-Making and Interpretation.
34. Yi Z, Xiao T, Albert MV. A Multimodal Multi-Agent Framework for Radiology Report Generation. Published online May 14, 2025. doi:10.48550/arXiv.2505.09787
35. Li J, Zhou T, Zhou Z, et al. Experience-guided multi-agent interpretable framework for radiology report summarization. *Computer Methods and Programs in Biomedicine*. 2026;273:109078. doi:10.1016/j.cmpb.2025.109078
36. 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
37. 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
38. Gawade S, Akhouri S, Kulkarni C, et al. Multi Agent based Medical Assistant for Edge Devices.
39. Yao Z, Chafekar T, Wang J, et al. ChatCLIDS: Simulating Persuasive AI Dialogues to Promote Closed-Loop Insulin Adoption in Type 1 Diabetes Care. Published online 2025. doi:10.48550/ARXIV.2509.00891
40. Sun B, Die, Hu. CTG-Insight: A Multi-Agent Interpretable LLM Framework for Cardiotocography Analysis and Classification. Published online 2025. doi:10.48550/ARXIV.2507.22205
41. Yao T, Xu Y, Wang H, Qiu X, Althoefer K, Qi P. Multi-Agent Fuzzy Reinforcement Learning with LLM for Cooperative Navigation of Endovascular Robotics.
42. Zhao L, Bai J, Bian Z, et al. Autonomous Multi-Modal LLM Agents for Treatment Planning in Focused Ultrasound Ablation Surgery. Published online July 15, 2025. doi:10.48550/arXiv.2505.21418
43. 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
44. 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
45. 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
46. 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
47. 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
48. 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.
49. Kim Y, Jeong H, Park C, et al. Tiered Agentic Oversight: A Hierarchical Multi-Agent System for Healthcare Safety. Published online September 28, 2025. doi:10.48550/arXiv.2506.12482

 doi:10.20944/preprints202606.0640.v1

50. 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
51. Akinseloyin O, Jiang X, Palade V. An LLM-based Multi-Agent Collaborative Approach for Abstract Screening towards Automated Systematic Reviews.
52. Wu H, Zhu Y, Wang Z, et al. EHRFlow: A Large Language Model-Driven Iterative Multi-Agent Electronic Health Record Data Analysis Workflow.
53. Angulo J, Yeste V. Notebook for the BioASQ Task 13b Lab at CLEF 2025.
54. 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
55. 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
56. Li H, Pan W, Rajendran S, Zang C, Wang F. TrialGenie: Empowering Clinical Trial Design with Agentic Intelligence and Real World Data.
57. 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
58. Moran J. EAGLE-AI: A large language model workflow for automated extraction and scoring of literature evidence linking genes to autism spectrum disorder.
59. Wysocki O, Wysocka M, Jacobo M, Unsworth H, Freitas A. Biomedical reasoning in action: Multi-agent System for Auditable Biomedical Evidence Synthesis. Published online 2025. doi:10.48550/ARXIV.2510.05335
60. Livieratos A. MetaMind: A Multi-Agent Transformer-Driven Framework for Automated Network Meta-Analyses.
61. Wei H, Qiu J, Yu H, Yuan W. MEDCO: Medical Education Copilots Based on A Multi-Agent Framework.
62. Sangwon KL. A Multi-AI Agent Framework for Interactive Neurosurgical Education and Evaluation: From Vignettes to Virtual Conversations.
63. Awasthi A, Chang BV, Vu AM, et al. MAARTA:Multi-Agentic Adaptive Radiology Teaching Assistant.
64. Sangwon KL. Evaluating Large Language Model Diagnostic Performance on JAMA Clinical Challenges via a Multi-Agent Conversational Framework.
65. 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
66. Lim E, He YV, Joselowitz J, et al. MATRIX: Multi-Agent simulaTion fRamework for safe Interactions and conteXtual clinical conversational evaluation. Published online 2025. doi:10.48550/ARXIV.2508.19163
67. 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
68. Sabel J, Wingren M, Lundell A, Andersson S. Medication counseling with large language models: improving self-evaluation through multi-agent systems.
69. 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
70. Chen Z, Peng Z, Liang X, et al. MAP: Evaluation and Multi-Agent Enhancement of Large Language Models for Inpatient Pathways. Published online March 17, 2025. doi:10.48550/arXiv.2503.13205
71. Lee Y, Wang X, Yang CC. Automated Clinical Problem Detection from SOAP Notes using a Collaborative Multi-Agent LLM Architecture. Published online August 29, 2025. doi:10.48550/arXiv.2508.21803
72. 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

 doi:10.20944/preprints202606.0640.v1

73. 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
74. 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
75. 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
76. 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
77. Chen K, Zhen T, Wang H, et al. MedSentry: Understanding and Mitigating Safety Risks in Medical LLM Multi-Agent Systems. Published online May 27, 2025. doi:10.48550/arXiv.2505.20824
78. Chan TK, Dinh ND. ENTAgents: AI Agents for Complex Knowledge Otolaryngology. Published online January 7, 2025. doi:10.1101/2025.01.01.25319863
79. Feng Y, Wang J, Zhou L, Lei Z, Li Y. DoctorAgent-RL: A Multi-Agent Collaborative Reinforcement Learning System for Multi-Turn Clinical Dialogue. Published online October 14, 2025. doi:10.48550/arXiv.2505.19630
80. Xia P, Wang J, Peng Y, et al. MMedAgent-RL: Optimizing Multi-Agent Collaboration for Multimodal Medical Reasoning. Published online January 26, 2026. doi:10.48550/arXiv.2506.00555
81. Zhuang Y, Jiang W, Zhang J, Yang Z, Zhou JT, Zhang C. Learning to Be A Doctor: Searching for Effective Medical Agent Architectures. Published online August 15, 2025. doi:10.48550/arXiv.2504.11301
82. 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.
83. Zhao H, Zhu Y, Wang Z, Wang Y, Gao J, Ma L. ConfAgents: A Conformal-Guided Multi-Agent Framework for Cost-Efficient Medical Diagnosis. Published online August 6, 2025. doi:10.48550/arXiv.2508.04915
84. Wang W, Ma Z, Wang Z, et al. A Survey of LLM-based Agents in Medicine: How far are we from Baymax?
85. 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
86. 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
87. *Ethics and Governance of Artificial Intelligence for Health: Large Multi-Modal Models. WHO Guidance*. 1st ed. World Health Organization; 2024.
88. Yagoubi FE, Mallah RA, Badu-Marfo G. AgentLeak: A Full-Stack Benchmark for Privacy Leakage in Multi-Agent LLM Systems. Published online February 12, 2026. doi:10.48550/arXiv.2602.11510
89. Prakash C, Lind M, Sisodia A. Agentic AI Governance and Lifecycle Management in Healthcare. Published online January 22, 2026. doi:10.48550/arXiv.2601.15630
90. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020;2(6):305-311. doi:10.1038/s42256-020-0186-1
91. 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
92. Juneja G, Pasupulati JNS, Albalak A, Hua W, Wang WY. MAGPIE: A benchmark for Multi-AGent contextual PrIvacy Evaluation. Published online October 16, 2025. doi:10.48550/arXiv.2510.15186
93. Chen YJ, Albarqawi A, Chen CS. Reinforcing Clinical Decision Support through Multi-Agent Systems and Ethical AI Governance. Published online September 22, 2025. doi:10.48550/arXiv.2504.03699
94. 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
95. 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
96. 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

 doi:10.20944/preprints202606.0640.v1

97. 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
98. 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
99. 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
100. 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
101. 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
102. 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
103. 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
104. 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
105. 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
106. 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
107. 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
108. Teodoro D, Naderi N, Yazdani A, Zhang B, Bornet A. A Scoping Review of Artificial Intelligence Applications in Clinical Trial Risk Assessment. Published online January 22, 2025. doi:10.1101/2025.01.21.25320310