The Evolutionary Trajectory of Medical Multi-Agent Systems: A Survey

Kai Chen[1,†], **Chenzhang Li**[1,†], **Tianyi Yu**[1,†], **Changzhuo Zuo**[1,†]
Xinyun Zhou[2], **Sijia Jing**[1], **Yanqing Wu**[1], **Wenfei Gao**[1]
Xinfeng Li[2,‡], **Tianpei Yang**[1,‡], **Bo An**[3,‡], **Yang Gao**[1,‡]

[1]Nanjing University, [2]The Hong Kong Polytechnic University,
[3]Nanyang Technological University,

Abstract

Breakthroughs in large language models are accelerating the evolution of medical artificial intelligence from standalone models to multi-agent systems. To organize this landscape, we propose a hierarchical taxonomy spanning Level 0 to Level 5. Given that foundational Level 0 to Level 2 architectures face cognitive and collaborative bottlenecks in complex clinical workflows, this survey focuses primarily on the architectural paradigm shift between Level 3 and Level 4. Specifically, we examine how systems transition from static Level 3 collaboration based on predefined roles to Level 4 self-evolving architectures that iteratively refine internal strategies through interactive feedback. Concurrently, we explore governance frameworks required to ensure system safety and trustworthiness under stringent clinical standards. Finally, we highlight that the capacity to dynamically generate agents and reconstruct collaborative topologies based on real-world scenarios will be essential for surpassing current capability limits. Building on this concept, we envision a Level 5 medical AI landscape defined by evolutionary agent swarms. To help researchers continuously track this field, our updated GitHub repository is publicly available.

1 Introduction

The rise of large language models (LLMs) is reshaping medical AI. To clarify this rapidly evolving technical landscape, we propose a hierarchical taxonomy from Level 0 to Level 5 (L0–L5). While **L0** (isolated medical LLMs) and **L1–L2** (tool-augmented single agents) excel at standardized clinical tasks, they encounter severe bottlenecks in complex environments (He et al., 2025). During long-horizon reasoning, single agents exhibit two critical vulnerabilities. First, they are prone to severe *context collapse*. In demanding tasks like rare disease diagnosis, the inability to simultaneously manage information extraction and logical reasoning often causes early critical symptoms to be overshadowed, severely degrading accuracy (Chen et al., 2025b; Li et al., 2025b). Second, unsupervised single agents are highly susceptible to *anchoring bias* (Ke et al., 2024), frequently overrelying on initial observations to reinforce early misjudgments and amplify hallucinations.

To overcome these limitations, medical AI is shifting toward medical multi-agent systems (MMAS). This corresponds to **L3** (static collaboration) and **L4** (adaptive evolution) in our taxonomy. Instead of monolithic models, MMAS strategically decompose complex clinical workflows across a network of expert agents (Kim et al., 2024b). This shift echoes the recent concept of *Harness Engineering*[1], emphasizing role specialization and collaborative environment design over micro-managing reasoning steps. Using distributed reasoning and *agentic context engineering*, MMAS manage cognitive workloads to prevent context collapse and mitigate single points of failure. Ultimately, this enables the transition from static coordination to dynamic adaptive evolution.

Despite this critical shift, existing literature remains confined to foundational levels. As systematically compared in Table 1, prior surveys predominantly cluster within the **L0–L3** range, focusing on static paradigms without mechanisms for adaptive evolution. Furthermore, current multi-agent research severely lacks collective safety mechanisms to prevent cascading errors across interconnected nodes. To bridge these dual gaps in engineering practice and safety governance, this paper explicitly targets the **L3–L4** domain. By detailing the transition from static collaboration to dynamic adaptation, we aim to provide systematic insights into the

[†]Equal contribution.
[‡]Corresponding authors.

[1]https://openai.com/index/harness-engineering/

Table 1: Comparison of our work with representative surveys. **Key:** ● Comprehensive; ◐ Partial; ○ Little to no coverage.

Ref.	Domain	Level	Core Focus	Workflow & Evolution		Governance & Safety	
				Stat.	Mechanism	Stat.	Target
(Xi et al., 2025)	General	L1–3	Agent Ecosystems	◐	Social Sim.	◐	Social Risks
(Li et al., 2024b)	General	L3	System Infrastructure	◐	Unified Framework	○	Intrinsic Constraints
(Guo et al., 2024)	General	L3	Simulation	◐	Game Feedback	○	Emergence Focus
(Lacerda et al., 2025)	Medical	L0	Evaluation Taxonomy	○	Discrim. vs. Gen.	◐	Robustness & Fairness
(He et al., 2025)	Medical	L0–1	Model Lifecycle	○	Static App.	●	Data Ethics
(Xiao et al., 2025)	Medical	L0–1	Multimodal	○	Static Dataset	○	Accuracy Focus
(Wang et al., 2025c)	Medical	L1–3	Architectures & App.	◐	Static Collaboration	◐	Hallucination
(Zhao et al., 2025)	Medical	L2–3	Task Taxonomy	◐	Pipelines	○	Compliance
(Zhi et al., 2025)	Medical	L2–3	Cognitive Architectures	◐	Strategic Planning	◐	Verifiability
Ours	**Medical**	**L3–5**	**Clinical Teams**	●	**Adaptive Evolution**	●	**Collective Safety**

trajectory toward **L5** medical agent swarms.

The structure of the paper is organized as follows. § 2 focuses on the architectural evolution from **L3** static collaboration to **L4** adaptive evolution, integrating discussions on knowledge utilization strategies and evaluation benchmarks. Subsequently, § 3 explores safety and privacy governance specifically for **L3–L4** systems, building robust defensive frameworks from both internal model and external human perspectives. Finally, § 4 looks ahead toward **L5**, discussing the open challenges faced by evolutionary medical agent swarms endowed with collective intelligence. Additionally, the formal notations and strict boundary definitions for **L0–L5** are further detailed in Appendix A and B.

Our main contributions are threefold:

- **Evolution-Centric Taxonomy.** We formalize a hierarchical taxonomy (**L0–L5**) for MMAS, spotlighting the critical transition from static collaboration (**L3**) to adaptive evolution (**L4**).
- **Unified Governance Perspective.** Rather than treating architecture and safety in isolation, we integrate evaluation, knowledge utilization, and governance into a cohesive framework. This bridges engineering design with clinical accountability from both agent-centric and human-centric standpoints.
- **Prospective Swarm Frontier.** We introduce **L5** as an exploratory frontier featuring runtime topological self-organization. This concept addresses open-ended medical complexities through safety-constrained agent generation and topology reconstruction.

2 Multi-Agent Medical AI Systems

This section analyzes the architectural transition from **L3** to **L4** (illustrated in Figure 1), examining the knowledge utilization strategies and evaluation benchmarks essential for collaborative reliability.

2.1 Level 3: Static Multi-Agent Collaboration

Constrained by cognitive ceilings (Lopez et al., 2025) and hallucination risks (Pal et al., 2023), individual LLMs underperform in complex medical tasks (Croxford et al., 2025). MMAS resolve this bottleneck through role specialization and distributed reasoning. The **L3** stage is characterized by a fixed agent configuration with predefined roles and immutable communication topologies, maintaining invariant persistent system states across independent medical tasks. Crucially, although **L3** architectures can accommodate intricate session-bound workflows, such as multi-round debates, in-context retrieval, or prompt-level adaptation, any resulting adaptive gains cannot be carried over to other independent tasks. Based on coordination mechanisms and decision-making authority, we classify these **L3** systems into four archetypes: **Hierarchical**, **Debate**, **Parallel**, and **Hybrid**.

Hierarchical. Centralized architectures employ dedicated coordinators for systematic task decomposition and strategic delegation, mirroring emergency department triage systems. Illustrating this approach, CLH (Motzfeldt et al., 2025) adopts a multi-stage medical coding pipeline, distributing clinical snippets and diagnostic chapters to parallel agents before final synthesis. Further advancing this centralized paradigm, KG4Diagnosis (Zuo et al., 2025a) explicitly proposes a "two-tier hierarchical architecture" where a general practitioner agent leads the initial assessment, task decomposition, and expert assignment. This pattern provides clear accountability in structured workflows, but creates potential single points of failure that compromise system reliability.

Debate. Peer-to-peer architectures achieve con-

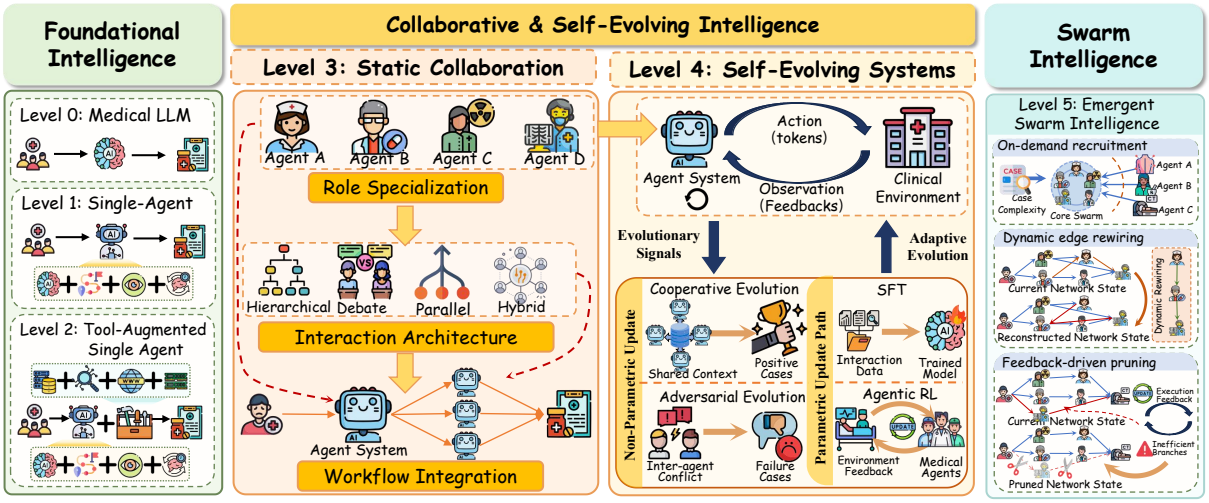


Figure 1: Overview of the proposed L0–L5 evolutionary taxonomy for MMAS.

sensus through iterative critique rather than hierarchical control, mirroring clinical multidisciplinary team (MDT) meetings. For instance, MedARC (Miao et al., 2026) utilizes structured inter-agent debate and confidence-aware aggregation to deepen reasoning via multi-round refinement. Similarly, for complex oncology decisions, MDT cancer cases (Kim et al., 2025) simulates human consultations via predefined role-based dialogues. Notably, sparse debate topologies can optimize the performance-cost tradeoffs in such setups (Li et al., 2024c). Although systematic adversarial verification enhances robustness, this paradigm incurs high coordination costs and may fail to converge promptly in time-critical scenarios.

Parallel. Independent specialists generate autonomous judgments prior to centralized synthesis, mirroring independent double reading in radiology. For instance, MoMA (Gao et al., 2025) deploys expert agents to independently process and summarize non-textual modalities (e.g., imaging and lab tests) for a central aggregator to yield clinical predictions. Similarly, CARE-AD (Li et al., 2025b) uses multidisciplinary agents to independently evaluate longitudinal clinical records across various Alzheimer's disease dimensions before final aggregation. While effectively balancing autonomy and structured synthesis for standardized tasks, this architecture underperforms on complex reasoning chains that require dynamic interactions.

Hybrid. Hybrid architectures integrate hierarchical triage with distributed collaboration. For example, MedLA (Ma et al., 2026) uses specialized agents to guide parallel nodes in building logic trees for iterative debate, while MAC (Chen et al., 2025b) employs a supervisor agent to oversee doctor agents in repeated consensus-seeking debates. Similarly, ColaCare (Wang et al., 2025e) introduces a MetaAgent to orchestrate domain-specific agents, bridging structured EHR modeling with multi-round LLM consultations. This pattern excels in multi-stage workflows requiring both routing efficiency and distributed expertise, despite its high configuration complexity.

Beyond Static Limits? While the four patterns above demonstrate effective collaboration, they are constrained by predefined structures and foundation model limitations. By introducing adaptive evolution, MMAS can break these bottlenecks. This represents a fundamental leap in system intelligence rather than a simple reconfiguration of workflows. The following section explains how MMAS transition from static task decomposition to continuous evolution through two core update paths: non-parametric and parametric.

2.2 Level 4: Adaptive Evolution

At **L4**, explicit feedback updates persistent, cross-episode reusable states including memory, coordination policies, or model parameters within a fixed topology to achieve quantifiable adaptation gains. This boundary separates **L4** from **L3** systems where feedback-driven mechanisms, such as multi-round debates or prompt-level routing, remain transient and evaporate upon task completion. This encompasses non-parametric tracks like cooperative and adversarial co-evolution alongside parametric tracks driven by Supervised Fine-Tuning (SFT) and Agentic Reinforcement Learning (RL).

2.2.1 Non-Parametric Update Path

Cooperative Evolution. Cooperative evolution refers to the progressive improvement of MMAS

through repeated task execution and workflow adjustment toward shared clinical objectives without updating the underlying language model parameters. Agent Hospital (Li et al., 2024a) instantiates this idea in a virtual hospital composed of doctor, nurse, and patient agents, where doctor agents iteratively manage large-scale synthetic cases, thereby refining their diagnostic and treatment decision-making capabilities. (Zhuang et al., 2025) further model MMAS collaboration as a searchable workflow graph and leverage performance feedback to reorganize effective diagnostic pipelines, allowing collaboration structures to evolve within a bounded workflow space. Similarly, MedChain (Liu et al., 2025b) combines sequential clinical evaluation with retrieval-augmented feedback, enabling agents to reuse effective decision trajectories across cases and stabilize knowledge-supported clinical workflows. Beyond conventional diagnosis and treatment, BioMedAgent (Bu et al., 2026) extends cooperative evolution to biomedical data analysis, where multiple LLM agents learn to invoke, combine, and reuse bioinformatics tools through interactive exploration and memory retrieval. Collectively, these studies suggest that inter-agent interactions can be transformed into reusable experience, optimizing clinical workflows and stabilizing effective decision-making processes.

Adversarial Co-evolution. Unlike purely cooperative setups, adversarial co-evolution utilizes uncertainty and disagreement to generate error-sensitive feedback. Current literature investigates two primary mechanisms. First, increasing external task difficulty drives policy optimization. For example, EvoPatient (Du et al., 2025) utilizes a patient agent that presents progressively complex clinical cases, forcing the doctor agent to improve its medical inquiry and diagnostic reasoning under uncertainty. Second, internal divergence between models provides a corrective signal. The MCC framework (Sun et al., 2026) implements this approach by triggering multi-round adversarial debates when model responses diverge, using heterogeneous peer critique to identify flaws in initial responses and reach a validated consensus. Through these mechanisms, conflicting outputs and reasoning errors serve as operational feedback signals that guide the refinement of multi-agent systems.

2.2.2 Parametric Update Path

SFT. Evolution in MMAS can also be consolidated through SFT. Existing studies can be broadly divided into system-level capability consolidation and agent-level role specialization. For instance, WSI-Agents (Lyu et al., 2025) maintains an MLLM zoo of fine-tuned, pathology-specific multimodal models that are dynamically invoked by specialized expert agents to generate clinical responses in whole-slide image analysis. MRGAgents (Wang et al., 2025a) focuses on agent-level specialization by decomposing medical report generation into disease- and finding-specific subtasks and fine-tuning specialist agents. PatientVLM Meets DocVLM (Lokesh et al., 2026) uses pre-diagnostic conversations between patient and doctor vision-language models to construct training samples, enabling diagnostic models to reason jointly over visual evidence and patient complaints. Overall, SFT serves as a parametric mechanism, stabilizing domain-specific agent capabilities.

Agentic RL. Beyond supervised consolidation, agentic RL enables MMAS to optimize behavior through goal-directed interaction and reward feedback under clinical uncertainty. Existing methods can be categorized by whether MAS serves as a training scaffold for improving target agents or as the direct object of reinforcement learning.

The first category improves agents through multi-role interaction or executable task environments. DoctorAgent-RL (Feng et al., 2026) builds a physician–patient interaction environment to train doctor agents in sequential questioning and diagnostic decision-making. Doctor-R1 (Lai et al., 2026) introduces an experiential reinforcement learning framework within a closed loop of doctor, patient, and evaluator agents, improving inquiry ability through rewards for both clinical decisions and communication quality. MedVR (Jiang et al., 2026) trains a medical visual reasoning agent via annotation-free reinforcement learning, distilling reward signals from multi-agent agreement.

The second category treats the MAS itself as the optimization target, focusing on joint decision policies and coordination structures. LA-CDM (Bani-Harouni et al., 2026) formulates diagnostic hypothesis generation, auxiliary examination selection, and diagnostic convergence as an optimizable sequential decision process, enabling adaptive behavior under incomplete information and constrained examination costs. MMedAgent-RL (Xia et al., 2026) further applies reinforcement learning to system-level multimodal consultation, using curriculum-guided training to optimize coordination between triage and attending agents and to in-

tegrate specialist opinions across clinical domains. Overall, agentic RL transforms interaction feedback into policy-level adaptation, enabling MMAS to optimize inquiry strategies, specialist coordination, and multimodal clinical reasoning.

Lab to Clinic? Most existing studies evaluate MMAS in simulated environments (e.g., virtual hospitals or task sandboxes). Consequently, performance gains often reflect overfitting to task formats or evaluator preferences rather than genuine clinical reasoning. For L4 systems, the core question is not simply whether they adapt, but whether this adaptation is clinically meaningful and transferable. Given the complex heterogeneity of real-world healthcare, future MMAS must prove that strategies learned in controlled settings can reliably generalize to open clinical workflows through rigorous multi-center hospital validation.

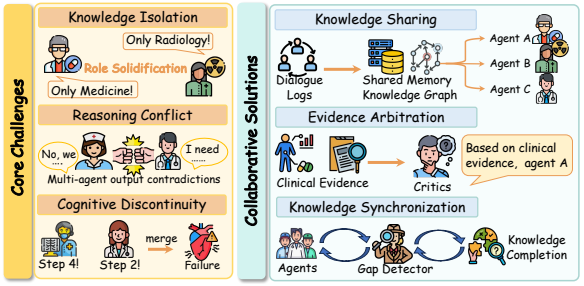


Figure 2: Knowledge Utilization in MMAS.

2.3 Knowledge Utilization

Knowledge utilization is essential for both **L3** and **L4** systems to achieve reliable collective intelligence. To overcome knowledge isolation, reasoning conflicts, and cognitive discontinuity, existing methods employ knowledge sharing, evidence arbitration, and knowledge synchronization (Figure 2). These mechanisms reflect shared strategies across static and evolving paradigms.

Knowledge Sharing. To overcome knowledge isolation across specialized clinical roles, MMAS establish structured information exchange. For instance, MedAgentSim (Almansoori et al., 2025) utilizes iterative interactions to share diagnostic hypotheses and experiential cues. Building upon this, CARE (Li et al., 2025a) incorporates structured medical knowledge and shared intermediate reasoning states to align evidence during inference. This foundational sharing mitigates role isolation and forms the basis of collective cognition.

Knowledge Conflict. Non-evidence-driven methods, such as debate or consensus paradigms, can produce conflicts during parallel reasoning (Gu et al., 2025). To address the aforementioned challenges, recent studies use structured and auditable frameworks that control knowledge flow, add critics, and anchor decisions in evidence (Feng et al., 2025). These designs improve reliability, fairness, and interpretability in MMAS.

Table 2: Taxonomy of MMAS evaluation benchmarks. * denotes single-agent benchmarks for multi-agent capabilities.

Type	Evaluation Dimension	Benchmarks	Ref.
Static Eval.	Foundational Knowledge	MedHELM*	(Bedi et al., 2026)
		CMExam*	(Liu et al., 2023)
	Complex Reasoning	MedXpertQA*	(Zuo et al., 2025b)
		MedAgentsBench	(Tang et al., 2025)
	Clinical Workflow	MedAgentBoard	(Zhu et al., 2025a)
		MedAgentBench*	(Jiang et al., 2025)
Dynamic Eval.	Safety & Multi-turn	AgentClinic	(Schmidgall et al., 2024)
		HealthBench*	(Arora et al., 2025)
		LLMEval-Med*	(Zhang et al., 2025c)
		MedSafetyBench	(Han et al., 2024)
	Multimodal Integration	3MDBench	(Sviridov et al., 2025)
		Neural-MedBench*	(Jing et al., 2026)
		MedThink-Bench*	(Zhou et al., 2025)
		TumorChain*	(Li et al., 2026b)
	Failure Attribution	FaultPoint	(Liu et al., 2025a)
		MedAgentAudit	(Gu et al., 2025)
		MAS-AFA	(Zhang et al., 2025d)
		MedAgentSim	(Almansoori et al., 2025)
		MedKGEval*	(Yu et al., 2025b)

Knowledge Synchronization. To tackle cognitive discontinuity, MMAS should coordinate participants' cognitive states across the interaction process rather than relying on reconciliation at the final stage during collaborative reasoning. For instance, KAMAC (Wu et al., 2025) employs a knowledge-gap-driven paradigm that iteratively identifies missing knowledge and dynamically introduces complementary experts to guide case-specific convergence. The approach achieves process-level iterative alignment, thereby enabling coherent collective reasoning and reliable decision making.

How to Unify? Despite recent progress, knowledge utilization remains fragmented, as most approaches address isolated aspects of collaboration rather than modeling it as a unified process. Future research should move further by developing holistic frameworks that jointly optimize sharing, arbitration, and synchronization to ensure clinically accountable decisions.

2.4 Static and Dynamic Evaluation Paradigms

MMAS evaluation spans two complementary paradigms (Table 2): static benchmarks for capability verification and dynamic environments for interactive proficiency, highlighting the shift from individual accuracy to collaborative reliability.

I. Static Benchmarking. Static benchmarks pro-

vide rigorous standardized assessment through carefully structured tasks under closed system states. Table 2 systematically categorizes system capabilities across three dimensions:

a. *Foundational Knowledge.* This dimension evaluates the macro-level medical domain expertise of individual agents. MedHELM (Bedi et al., 2026) aligns models with clinical practice using a multi-dimensional taxonomy of complex real-world tasks. Furthermore, CMExam (Liu et al., 2023) enhances evaluation authenticity via a robust, five-dimensional expert-annotated system covering disease groups and clinical departments.

b. *Complex Reasoning.* At the micro-level, evaluation shifts to logical rigor and consistency. Benchmarks like MedXpertQA (Zuo et al., 2025b) quantify fine-grained professional reasoning capabilities via expert-curated questions and exhaustive trajectory verification. Additionally, MedAgentsBench (Tang et al., 2025) assesses system performance in navigating complex medical logic.

c. *Clinical Workflow.* This dimension evaluates agent orchestration across clinical pathways. Benchmarks like MedAgentBench (Jiang et al., 2025) and MedAgentBoard (Zhu et al., 2025a) leverage EHR-based prediction and virtual-EHR workflow settings to assess workflow automation, calibrating task completeness and overhead trade-offs to ensure agent actions mirror authentic clinical operations.

Limitations. Static evaluation lacks authentic clinical noise, fails to model realistic time pressure, and leaves specific collaborative attributes, such as inter-agent communication efficiency and consensus quality, unmeasured in these frameworks.

II. Dynamic Evaluation. Dynamic evaluation assesses complex emergent collaborative behaviors and real-time decision-making performance of multi-agent systems under authentic clinical noise.

a. *Safety Benchmarks.* AgentClinic (Schmidgall et al., 2024) simulates a comprehensive clinician, patient, and moderator ecosystem to test adaptive decision-making. HealthBench (Arora et al., 2025) defines foundational safety baselines, while LLMEval-Med (Zhang et al., 2025c) introduces an expert verification system based on real-world cases to measure clinical utility. Furthermore, MedSafetybench (Han et al., 2024) benchmarks the safety robustness of medical large language models against adversarial prompt injections and malicious instructions through quantified attack metrics.

b. *Multimodal and Reasoning.* For heterogeneous data integration, 3MDBench (Sviridov et al., 2025) evaluates multi-agent telemedical consultation capabilities. Neural-MedBench (Jing et al., 2026) establishes agent reasoning fidelity baselines in neurology, informing critical lower-bound expectations for collaborative diagnostic performance. Meanwhile, MedThink-Bench (Zhou et al., 2025) and TumorChain (Li et al., 2026b) enable rigorous scrutiny of logical coherence and reasoning trajectory traceability under expert-level complexities.

c. *Failure Attribution and Audit Pipelines.* This dimension conducts causal analysis via conversational flow auditing. Fault Point (Liu et al., 2025a) identifies decision vulnerabilities within reasoning. MedAgentAudit (Gu et al., 2025) and MAS-AFA (Zhang et al., 2025d) identify systemic risks, including critical information loss or severe consensus misalignment, by auditing detailed inter-agent communication, which are issues often masked by final accuracy metrics. Finally, MedAgentSim (Almansoori et al., 2025) and MedKGEval (Yu et al., 2025b) utilize multi-agent simulation and knowledge graph-driven evaluation, respectively, to quantify information gathering efficiency and guideline alignment, effectively mitigating knowledge degradation during multi-turn clinical interactions.

How to Evaluate? Evaluations must transcend the "final answer" black-box paradigm by introducing in-depth metrics that gauge distributed medical decision-making processes in **L3** and **L4**. Detailed metrics are deferred to Appendices C and D.

3 Safety, Privacy and Ethics

The evolution from **L3** to **L4** bolsters capabilities but introduces complex interaction risks. To mitigate these systemic hazards and establish safety boundaries for **L5**, robust governance is essential. We construct a comprehensive governance framework covering safety, privacy, and ethics by integrating **agent-centric** and **human-centric** views (Figure 3). Detailed clinical threat modeling is provided in Appendix E.

3.1 Agent-Centric View

This perspective treats risks as inherent algorithmic and behavioral vulnerabilities, mitigating them by embedding safeguards and normative constraints directly into the agents' technical core.

Safety challenges in MMAS stem from unintentional risks (e.g., hallucination propagation, syco-

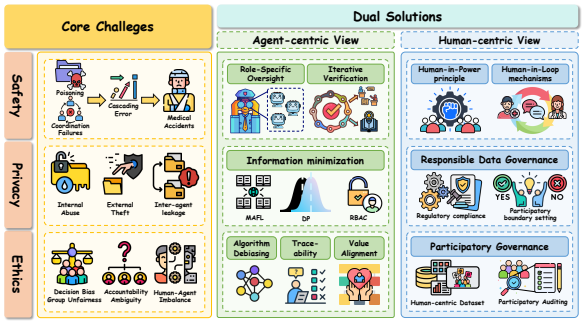


Figure 3: Safety, privacy and ethics risks in MMAS.

phancy, and collusion (Zhu et al., 2025b)) and intentional attacks (e.g., jailbreaking and poisoning). Both exploit collaborative workflows to propagate misinformation, fundamentally degrading clinical reliability. To mitigate these risks, current research converges on two paradigms. First, static *Role-Specific Oversight* deploys guardian agents to monitor system integrity and manage anomalous peers, as demonstrated by MedSentry's psychometric profiling (Chen et al., 2025a). Second, dynamic *Iterative Verification* enforces reliability through adversarial debate to eliminate single points of failure (Tang et al., 2024), external knowledge grounding against injection risks (Gilbert et al., 2024), and continuous internal auditing for self-correction (Dhuliawala et al., 2024).

Privacy risks in MMAS stem from frequent collaborative information exchange. To prevent unauthorized disclosure, agent communication must enforce information minimization. Key strategies include *Role-Based Access Control* (*RBAC*) (Yaqub et al., 2025), which restricts information flows via the least privilege principle, and *Multi-Agent Federated Learning* (*MAFL*) (Zhang et al., 2025a), which preserves data sovereignty by sharing model gradients or intermediate states instead of raw records. Additionally, *Differential Privacy* (*DP*) and anonymizer agents (Wiest et al., 2025) mathematically and semantically desensitize communication through statistical noise injection and Personally Identifiable Information (PII) redaction.

Ethics risks primarily stem from amplified sociodemographic biases and ambiguous traceability in clinical decision-making. Current research mitigates these through three main approaches. First, *Algorithm Debiasing* utilizes adversarial training to mathematically isolate clinical reasoning from sensitive attributes, minimizing demographic inference errors (Omiye et al., 2023). Second, *Traceability* is enhanced by embedding evaluator agents to establish audit trails, enabling clinicians to scrutinize intermediate AI suggestions (Baldoni et al., 2023). Third, *Value Alignment* embeds care ethics into MMAS objectives to promote empathetic, patient-aligned behavior over statistical metrics.

3.2 Human-Centric View

Conversely, this view attributes risks to misaligned human-agent dynamics and institutional gaps, resolving them via robust human-AI partnerships.

Safety risks stem from misaligned human-AI collaboration, such as overreliance and insufficient oversight, which can legitimize errors and cause misdiagnoses. Mitigating these requires *Human-in-the-Loop* and *Human-in-Power* mechanisms to ensure agents augment rather than replace clinicians (Johri et al., 2025). Accordingly, feasible human-AI collaborative paradigms have been developed for high-risk scenarios, including explicit design principles for diabetes treatment (Burgess et al., 2023) and the SepsisLab system for sepsis diagnosis (Zhang et al., 2024).

Privacy risks in MMAS stem from unclear data governance, leading to unauthorized exposure of sensitive patient information and erosion of clinical trust. Thus, prioritizing *Responsible Data Governance* is essential. This approach mandates strict *Regulatory Compliance* with acts such as the Health Insurance Portability and Accountability Act and the General Data Protection Regulation (Said et al., 2023), which necessitate the deployment of advanced technical safeguards to protect high-dimensional clinical data from re-identification. Complementarily, it integrates flexible *Participatory Boundary Setting*, empowering stakeholders to define data sharing protocols based on their cultural values (Kim et al., 2024a).

Ethics challenges in MMAS largely stem from unclear accountability and marginalized patient values. To address this, research advocates for *Participatory Governance*, a paradigm that restores the system's moral legitimacy by involving human stakeholders. This paradigm spans the entire development lifecycle, encompassing upstream *Human-centric Dataset* creation to address foundational ethical pitfalls (Mittal et al., 2024), while integrating downstream *Participatory Auditing* to rigorously ensure the transparency and contestability of AI alignment (Wieczorek et al., 2025).

4 Towards Emergent Swarm Intelligence

What is L5? Unlike the mature, literature-synthesized **L3** and **L4** paradigms, **L5** represents an exploratory frontier currently lacking large-scale clinical validation. It introduces runtime topological self-organization, enabling dynamic agent generation or deletion, automated rewriting of task decompositions, and execution-time topology reconstruction under explicit safety-constrained control. This prospective vision adapts general-domain breakthroughs. For instance, the Codex subagent workflow utilizes a fan-out/fan-in process[2], while AutoAgents pioneers task-driven generation (Chen et al., 2024). Similarly, GPTSwarm models agents as optimizable graphs (Zhuge et al., 2024), and AFlow continuously refines workflows via feedback (Zhang et al., 2025b), collectively guiding future medical systems to reorganize dynamically around shifting patient risks.

How it works? To overcome previous structural bottlenecks and achieve dynamic reorganization, **L5** swarms utilize three topology-level operations:

a. *On-demand recruitment.* Rather than invoking fixed experts, an **L5** system recruits sub-agents dynamically based on case complexity, evidence gaps, and uncertainty. While knowledge-gap expansion (§2.3) offers a precursor, it remains confined within predefined boundaries. Frameworks like TDAG (Wang et al., 2025d) address this through dynamic task decomposition and real-time agent generation. **L5** extends this capacity to fill clinical gaps by continuously integrating specialized nodes into an evolving reasoning graph.

b. *Dynamic edge rewiring.* **L5** dynamically selects agents based on tasks and determines the communication topology. Unlike bounded architectures (§2.1, §2.2), it enables runtime reconstruction of communication edges. Autoregressive graph generation (Li et al., 2026a) drives this transition toward fully searchable collaboration graphs through automated topology design. Furthermore, NetSafe (Yu et al., 2025a) shows that communication topology and connectivity affect the propagation of adversarial behavior, providing design evidence for safer interaction graphs.

c. *Feedback-driven pruning.* When a pathway lacks evidence or triggers safety risks, the system terminates branches. In multi-agent communication, Safesieve (Zhang et al., 2026) exemplifies this via progressive, experience-driven pruning to minimize token overhead. Unlike the bounded optimization of **L4** systems (§2.2), integrating execution feedback with clinical constraints enables **L5** to prune and reorganize the collaboration graph.

What's next? Looking ahead, we envision three potential medical scenarios where swarm architectures are particularly applicable:

a. *Complex comorbidity MDT.* For patients with complex comorbidities or rare diseases, conflicting specialty guidelines create severe reasoning dilemmas. Frameworks like Rareagents (Chen et al., 2026) demonstrate the clinical utility of autonomous multidisciplinary teams. An **L5** swarm builds on this by dynamically restructuring its agent network around primary conflicts to mitigate polypharmacy risks. This addresses systemic conflicts by globally reconstructing responsibility boundaries and evidence pathways.

b. *Emergency and critical care.* In dynamic settings like emergency triage and ICUs, traditional pipelines struggle under strict time constraints. Flowchart-guided multi-agent systems (Liu et al., 2026) demonstrate how specialized agents can effectively coordinate urgent clinical decisions. **L5** swarms scale this capacity by rapidly reconstructing collaborative logic based on real-time vital signs to form a fast-response loop across triage, diagnosis, and intervention.

c. *Precision medicine.* Traditional systems in oncology and multi-omics stall when cross-modal evidence is missing or conflicting. CASSIA (Xie et al., 2025) demonstrates multi-agent cell annotation using specialized validation nodes. **L5** swarms expand this by discovering evidence gaps, generating analytical nodes, and rewiring the network based on conflict intensity.

Overall, **L5** medical swarms promise to eliminate single-point systemic failures and cascading error propagation in **L3/L4** systems. Combining dynamic generation with graph search fosters resilient collective intelligence for open-ended tasks.

5 Conclusion

This paper systematically charts the evolution of medical multi-agent systems by proposing an **L0**–**L5** hierarchical taxonomy. We detail the critical architectural shift from **L3** static collaboration to **L4** adaptive evolution, alongside essential knowledge utilization and evaluation paradigms. To ensure clinical trustworthiness, we establish a dual governance framework that integrates agent-centric safe-

[2] https://developers.openai.com/codex/subagents.

Table 3: Literature search and selection protocol.

Stage	Component	Protocol
Search	Sources	Google Scholar, PubMed, Semantic Scholar, and arXiv, supplemented by backward and forward citation tracing.
	Period	Jan. 2023–Aug. 2026; earlier highly relevant or seminal works are also included.
	Queries	Search queries combine medical and clinical terminology with LLM agent keywords, spanning multi-agent collaboration, evolution, evaluation, and safety.
Selection	Inclusion	Core MMAS methods, evaluations, and contextual studies on taxonomy, benchmarks, and governance.
	Preprints	Directly relevant preprints when no peer-reviewed version was available.
	Exclusion	Non-LLM MAS; non-agent medical LLMs; low-relevance, duplicate, or superseded studies.
Corpus	Screening	Retrieved records were deduplicated, screened by title and abstract, closely examined for core content, and supplemented with searches for related literature.

guards and human-centric oversight against safety, privacy, and ethical risks. Ultimately, we conceptualize **L5** evolutionary swarms as the next frontier for complex clinical tasks. By bridging engineering design with clinical accountability, this survey provides a clear roadmap toward a safe and dynamic "Human-AI Symbiosis" in healthcare.

Survey Methodology

This section summarizes the literature collection and screening protocol, including the search period, sources, query scope, selection criteria, and screening process. We prioritized peer-reviewed publications, while retaining recent preprints when they introduced directly relevant MMAS methods, benchmarks, or emerging mechanisms, as this is a rapidly evolving research area in which important developments often appear before formal peer review. Retrieved records were deduplicated, screened by title and abstract, and subsequently assessed in full text, with backward and forward citation tracing used to identify additional relevant studies. The corresponding details are presented in Table 3,4.

Limitations

MMAS is a rapidly evolving field. As a result, this concept-driven survey may not capture every recent development. We prioritized peer-reviewed and accepted publications, while retaining publicly accessible and directly relevant preprints when no peer-reviewed version was available by the search cut-off date. The resulting corpus should therefore be viewed as a representative synthesis rather than an exhaustive systematic review. To mitigate this limitation, our companion repository will be updated to track subsequent works.

Table 4: Venue distribution of the final corpus (92 works).

Category	Venue group	Count
Journals	Nature, NPJ, NEJM family	19
	Cell and other journals	12
Conferences	NeurIPS, ICLR, ICML	16
	ACL, EMNLP, CoNLL	12
	AAAI, IJCAI	8
	CHI, MICCAI, MLHC and others	15
Preprints	arXiv, TechRxiv	10
Total		**92**

Ethical considerations

This study adheres to all relevant academic integrity and ethical standards regarding research conduct. AI assistants were employed exclusively for grammar checking and formatting adjustments to improve the clarity of the manuscript.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (62506157), Jiangsu Natural Science Foundation (BK20243051), Jiangsu Science and Technology Major Project (BG2025035), the Fundamental Research Funds for the Central Universities (14380128, KG202514), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

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

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Matteo Baldoni, Cristina Baroglio, Roberto Micalizio, and Stefano Tedeschi. 2023. Accountability in multi-agent organizations: from conceptual design to agent

programming. *Autonomous Agents and Multi-Agent Systems*, 37(1):7.

David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. 2026. Language agents for hypothesis-driven clinical decision making with reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, and 1 others. 2026. Holistic evaluation of large language models for medical tasks with medhelm. *Nature Medicine*, pages 1–9.

Dechao Bu, Jingbo Sun, Kun Li, Zihao He, Wei Huang, Jinlin Hu, Shanshan Zhang, Shuangshuang Lei, Peipei Huo, Zhihao Wang, Sheng Wang, Tao Wang, Kai Gao, Yang Wu, Lianhe Zhao, Kai Wang, Gen Li, Huan Song, Yang Jin, and 3 others. 2026. Empowering ai data scientists using a multi-agent llm framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses. *Nature Biomedical Engineering*.

Eleanor R Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare ai treatment decision support: Design principles to enhance clinician adoption and trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje Karlsson, Jie Fu, and Yemin Shi. 2024. Autoagents: A framework for automatic agent generation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 22–30.

Kai Chen, Taihang Zhen, Hewei Wang, Kailai Liu, Xinfeng Li, Jing Huo, Tianpei Yang, Jinfeng Xu, Wei Dong, and Yang Gao. 2025a. Medsentry: Understanding and mitigating safety risks in medical llm multi-agent systems. *arXiv preprint arXiv:2505.20824*.

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

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. 2026. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 101–109.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1):640.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

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

Qianhan Feng, Zhongzhen Huang, Yakun Zhu, Xiaofan Zhang, and Qi Dou. 2025. Ucagents: Unidirectional convergence for visual evidence anchored multi-agent medical decision-making. *arXiv preprint arXiv:2512.02485*.

Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. 2026. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16952–16956. IEEE.

Jifan Gao, Mahmudur Rahman, John Caskey, Madeline Oguss, Ann O'Rourke, Randall Brown, Anne Stey, Anoop Mayampurath, Matthew M. Churpek, Guanhua Chen, and Majid Afshar. 2025. Moma: a mixture-of-multimodal-agents architecture for enhancing clinical prediction modelling. *npj Digital Medicine*, 9:46.

Stephen Gilbert, Jakob Nikolas Kather, and Aidan Hogan. 2024. Augmented non-hallucinating large language models as medical information curators. *NPJ digital medicine*, 7(1):100.

Lei Gu, Yinghao Zhu, Haoran Sang, Zixiang Wang, Dehao Sui, Wen Tang, Ewen Harrison, Junyi Gao, Lequan Yu, and Liantao Ma. 2025. Medagentaudit: Diagnosing and quantifying collaborative failure modes in medical multi-agent systems. *arXiv preprint arXiv:2510.10185*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Advances in neural information processing systems*, 37:33423–33454.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey

of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 118:102963.

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

Zheng Jiang, Heng Guo, Chengyu Fang, Changchen Xiao, Xinyang Hu, Lifeng Sun, and Minfeng Xu. 2026. MedVR: Annotation-free medical visual reasoning via agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

Miao Jing, Mengting Jia, Junling Lin, Zhongxia Shen, Huan Gao, Mingkun Xu, and Shangyang Li. 2026. Beyond classification accuracy: Neural-medbench and the need for deeper reasoning benchmarks. In *The Fourteenth International Conference on Learning Representations*.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86.

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

Dajung Kim, Niko Vegt, Valentijn Visch, and Marina Bos-De Vos. 2024a. How much decision power should (a) i have?: investigating patients' preferences towards ai autonomy in healthcare decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Jaesik Kim, Byounghan Lee, Kyung-Ah Sohn, Dokyoon Kim, and Young Chan Lee. 2025. Evaluation of multi-agent llms in multidisciplinary team decision-making for challenging cancer cases. In *Machine Learning for Healthcare Conference*. PMLR.

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

Eyal Klang, Mahmud Omar, Ganesh Raut, Reem Agbareia, Prem Timsina, Robert Freeman, Nicholas Gavin, Lisa Stump, Alexander W. Charney, Benjamin S. Glicksberg, and Girish N. Nadkarni. 2026. Orchestrated multi agents sustain accuracy under clinical-scale workloads compared to a single agent. *npj Health Systems*, 3:23.

Anisio Lacerda, Gisele Pappa, Adriano César Machado Pereira, Wagner Meira Jr, and Alexandre Guimarães De Almeida Barros. 2025. Evaluation of medical large language models: taxonomy, review, and directions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 10528–10536.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. 2026. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

Haoran Li, Xusen Cheng, and Xiaoping Zhang. 2025a. 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*, pages 1–15.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024a. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

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

Shiyuan Li, Yixin Liu, Qingsong Wen, Chengqi Zhang, and Shirui Pan. 2026a. Assemble your crew: Automatic multi-agent communication topology design via autoregressive graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23142–23150.

Sijing Li, Zhongwei Qiu, Jiang Liu, Wenqiao Zhang, Tianwei Lin, Yihan Xie, Jianxiang An, Boxiang Yun, Chenglin Yang, Jun Xiao, Guangyu Guo, Jiawen Yao, Wei Liu, Yuan gao, Ke Yan, Weiwei Cao, Zhilin Zheng, Tony C. W. MOK, Kai Cao, and 6 others. 2026b. Tumorchain: Interleaved multimodal chain-of-thought reasoning for traceable clinical tumor analysis. In *The Fourteenth International Conference on Learning Representations*.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024b. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024c. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294.

Benjamin Liu, Dillon Mehta, Rishi Malhotra, Adam Zobian, Yong Ying Tan, Samir Chopra, Daniella Rand, Natalie Pang, Abhiram Gudimella, Raghav Thallapragada, Derek Jiu, and Kevin Zhu. 2025a.

Examining the vulnerability of multi-agent medical systems to human interventions for clinical reasoning. In *NeurIPS 2025 Workshop on Regulatable ML*.

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

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and 1 others. 2023. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452.

Yujia Liu, Sophia Yu, Hongyue Jin, Jessica Wen, Alexander Qian, Terrence Lee, Mattheus Ramsis, Gi Won Choi, Lianhui Qin, Xin Liu, and 1 others. 2026. A multi-agent framework combining large language models with medical flowcharts for self-triage. *Nature Health*, pages 1–10.

K Lokesh, Abhirama Subramanyam Penamakuri, Uday Agarwal, Apoorva Challa, Shreya K Gowda, Somesh Gupta, and Anand Mishra. 2026. Patientvlm meets docvlm: Pre-consultation dialogue between vision-language models for efficient diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(9):7485–7493.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, and 1 others. 2025. Clinical entity augmented retrieval for clinical information extraction. *NPJ digital medicine*, 8(1):45.

Xinheng Lyu, Yuci Liang, Wenting Chen, Meidan Ding, Jiaqi Yang, Guolin Huang, Daokun Zhang, Xiangjian He, and Linlin Shen. 2025. Wsi-agents: A collaborative multi-agent system for multi-modal whole slide image analysis. In *28th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2025)*, pages 680–690. Springer.

Siqi Ma, Jiajie Huang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. 2026. Medla: A logic-driven multi-agent framework for complex medical reasoning with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 845–853.

Yongming Miao, Jiaxin Wen, Yuemei Luo, and Jin Li. 2026. Medarc: Adaptive multi-agent refinement and collaboration for enhanced medical reasoning in large language models. *International Journal of Medical Informatics*, 206:106136.

Surbhi Mittal, Kartik Thakral, Richa Singh, Mayank Vatsa, Tamar Glaser, Cristian Canton Ferrer, and Tal Hassner. 2024. On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare. *Nature Machine Intelligence*, 6(8):936–949.

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

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334.

Abdelmlak Said, Aymen Yahyaoui, and Takoua Abdellatif. 2023. Hipaa and gdpr compliance in iot healthcare systems. In *International conference on model and data engineering*, pages 198–209. Springer.

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

Xinti Sun, Qiyang Hong, Mengyan Zhang, Yuyan Li, Tingwei Chen, Zigeng Huang, Guihan Liang, Wenjun Tang, Sulin Xu, Xiaolin Ni, Junling Pang, Peixing Wan, and Erping Long. 2026. Model confrontation and collaboration: A debate intelligence framework for enhancing medical reasoning in large language models. *Cell Reports Medicine*, 7(1):102547.

Ivan Sviridov, Amina Miftakhova, Tereshchenko Artemiy Vladimirovich, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 2025. 3mdbench: Medical multimodal multi-agent dialogue benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26614–26654.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, and 1 others. 2025. Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. *arXiv preprint arXiv:2503.07459*.

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

Dujuan Wang, Tao Cheng, Sutong Wang, Youhua Chen, and Yunqiang Yin. 2026. SMR-Agents: Synergistic medical reasoning agents for zero-shot medical visual question answering with MLLMs. *Information Processing & Management*, 63(1):104297.

Pengyu Wang, Shuchang Ye, Usman Naseem, and Jinman Kim. 2025a. Mrgagents: A multi-agent framework for improved medical report generation with med-lvlms. In *2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7. IEEE.

Sheng Wang, Fangyuan Zhao, Dechao Bu, Yunwei Lu, Ming Gong, and 1 others. 2025b. LINS: A general medical Q&A framework for enhancing the quality and credibility of LLM-generated responses. *Nature Communications*, 16:9076.

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

Yaoxiang Wang, Zhiyong Wu, Junfeng Yao, and Jinsong Su. 2025d. Tdag: A multi-agent framework based on dynamic task decomposition and agent generation. *Neural Networks*, 185:107200.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, and 1 others. 2025e. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, pages 2250–2261.

Catherine Wieczorek, Heidi Biggs, Kamala Payyapilly Thiruvenkatanathan, and Shaowen Bardzell. 2025. Architecting utopias: How ai in healthcare envisions societal ideals and human flourishing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Isabella C Wiest, Marie-Elisabeth Leßmann, Fabian Wolf, Dyke Ferber, Marko Van Treeck, Jiefu Zhu, Matthias P Ebert, Christoph Benedikt Westphalen, Martin Wermke, and Jakob Nikolas Kather. 2025. Deidentifying medical documents with local, privacy-preserving large language models: the llm-anonymizer. *NEJM AI*, 2(4):AIdbp2400537.

Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Yanyuan Qiao, Imran Razzak, and Yutong Xie. 2025. A knowledge-driven adaptive collaboration of llms for enhancing medical decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33495–33512.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Zihan Dong, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Linjun Zhang, Shujie LIU, Yan Lu, and Huaxiu Yao. 2026. MMedagent-RL: Optimizing multi-agent collaboration for multimodal medical reasoning. In *The Fourteenth International Conference on Learning Representations*.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2025. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888.

Elliot Xie, Lingxin Cheng, Jack Shireman, Yujia Cai, Jihua Liu, Chitrasen Mohanty, Mahua Dey, and Christina Kendziorski. 2025. Cassia: a multi-agent large language model for automated and interpretable cell annotation. *Nature Communications*.

Wenjun Xiong and Robert Lagerström. 2019. Threat modeling–a systematic literature review. *Computers & security*, 84:53–69.

Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May Dongmei Wang, Peifeng Ruan, Donghan Yang, Tao Wang, Guanghua Xiao, Xin Liu, Carl Yang, Yang Xie, and Wenqi Shi. 2026. Medagentgym: A scalable agentic training environment for code-centric reasoning in biomedical data science. In *The Fourteenth International Conference on Learning Representations*.

Nadeem Yaqub, Jianbiao Zhang, Muhammad Irfan Khalid, Weiru Wang, Markus Helfert, Mansoor Ahmed, and Jungsuk Kim. 2025. Blockchain enabled policy-based access control mechanism to restrict unauthorized access to electronic health records. *PeerJ Computer Science*, 11:e2647.

Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Kun Wang, Qingsong Wen, and Yang Wang. 2025a. NetSafe: Exploring the topological safety of multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2905–2938, Vienna, Austria. Association for Computational Linguistics.

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

Feilong Zhang, Deming Zhai, Guo Bai, Junjun Jiang, Qixiang Ye, Xiangyang Ji, and Xianming Liu. 2025a. Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare. *Nature Communications*, 16(1):2852.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. 2025b. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations*.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025c. LLMEval-Med: A real-world clinical benchmark for medical LLMs with physician validation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4888–4914, Suzhou, China. Association for Computational Linguistics.

Ruijia Zhang, Xinyan Zhao, Ruixiang Wang, Sigen Chen, Guibin Zhang, An Zhang, Kun Wang, and Qingsong Wen. 2026. Safesieve: From heuristics to experience in progressive pruning for llm-based multi-agent communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29892–29900.

Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M Padilla, Jeffrey Caterino, Ping Zhang, and 1 others. 2024. Rethinking human-ai collaboration in complex medical decision making: a case study in sepsis diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and 1 others. 2025d. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *Proceedings of Machine Learning Research*, 267:76583–76599.

Wenqi Zhao, Shansong Wang, Mojtaba Safari, Mingzhe Hu, and Xiaofeng Yang. 2025. Medical ai agents: A comprehensive survey of architectures, cognitive modules, and clinical workflows. *TechRxiv*, 2025(1201).

Xiaoquan Zhi, Hongke Zhao, Likang Wu, Chuang Zhao, and Hengshu Zhu. 2025. Reinventing clinical dialogue: Agentic paradigms for llm enabled healthcare communication. *arXiv e-prints*.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emma Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and Rui Zhang. 2025. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*, 9(1):34.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2025a. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025b. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

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

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. 2024. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62743–62767. PMLR.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2025a. Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis. In *AAAI Bridge Program on AI for Medicine and Healthcare*, pages 195–204. PMLR.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025b. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *International Conference on Machine Learning*, pages 80961–80990. PMLR.

A Definitions

A.1 Task Formulation

Before formalizing the specific multi-agent architectures, we must first clarify the temporal scope of the system. Let $t = 1, 2, \ldots, T$ denote the **discrete evolutionary step** (e.g., encountering the t-th clinical case or training iteration), explicitly distinguished from the internal reasoning turns within a single case. At step t, the system receives a clinical input x_t and generates a decision y_t. The efficacy of this decision is evaluated by a global clinical utility function $\Omega(y_t, x_t, y^*_{true})$, in which y^*_{true} denotes the ground truth, yielding a feedback signal r_t. The fundamental distinction between static and evolutionary architectures lies in their response to this feedback across steps. In static systems (**L0**–**L3**), the persistent internal states remain strictly frozen across cases, precluding the ability to learn from r_t. Conversely, in evolutionary multi-agent systems (**L4**–**L5**), the core objective is to maximize the expected clinical utility over time by continuously updating their persistent states (e.g., parameters, experience memory, or topology) driven by the cross-turn feedback r_t.

A.2 System Formalization

We formulate the Medical Multi-Agent System (MMAS) as a compact tuple $\mathcal{M} = (\mathcal{N}, \mathcal{E}, \mathcal{G}, \Omega)$, where $\mathcal{N}$ is the set of candidate agents, $\mathcal{E}$ is the set of external medical tools, $\mathcal{G}$ denotes the broader collaborative topology space, and Ω is the utility function. To mathematically isolate the dimensions of system evolution, we strictly decouple the system's variables into persistent and transient states. At any evolutionary step t, the **persistent system state** is characterized by $S_t = (\Theta_t, M_t, \Psi_t, G_t)$, corresponding to the underlying model parameters, the cross-case global experience memory, the coordination policy (e.g., routing weights or scheduling rules), and the active collaborative topology instantiated from $\mathcal{G}$, respectively. Conversely, during the inference process for a specific case x_t, the system generates **transient states**, including local observations $o_{i,t}$ for agent i and a working memory τ_t storing multi-step reasoning trajectories. This decoupled modeling is motivated by the need to precisely formalize the architectural leaps from **L0** to **L5**: the evolutionary levels can thus be strictly defined by tracking which components of S_t remain statically frozen and which are progressively unlocked for dynamic, feedback-driven updates. Formally, the system's conditional decision-making policy is denoted as $\pi(\cdot \mid \text{transient states}; \text{persistent states})$, mapping case-specific observations to clinical decisions under the parameterization of the current architectural constraints.

B Formalization of the Evolutionary Levels in MMAS

Building upon the decoupled state representation introduced above, this section mathematically formalizes the six levels of MMAS (**L0–L5**). We characterize this architectural evolution along two distinct trajectories. The first trajectory (spanning **L0** to **L3**) entails the progressive activation of system components, expanding the inference capacity through transient working memory τ_t, external tools $\mathcal{E}$, and multi-agent networks $\mathcal{N}$. The second trajectory (spanning **L4** to **L5**) dictates a deeper paradigm shift: the systematic relaxation of static constraints. This marks the fundamental transition from permanently frozen architectural states to adaptive, feedback-driven dynamic variables.

B.1 L0 Isolated Medical LLMs

The **L0** architecture represents the foundational paradigm of isolated, open-loop inference. Formally, the decision process is modeled as:

$$y_t \sim \pi(\cdot \mid x_t; \Theta_0) \tag{1}$$

The system consists of a single agent ($|\mathcal{N}| = 1$) operating strictly as a static mapping function from clinical input to output. Completely isolated from external tools and evaluation feedback, the persistent states remain strictly unchanged across evolutionary steps ($S_{t+1} = S_t$). Consequently, the system relies entirely on its initial pre-trained weights Θ_0 and is intrinsically incapable of accumulating experience or rectifying prior misdiagnoses.

B.2 L1 Autonomous Cognitive Single Agent

The **L1** architecture advances to an autonomous cognitive paradigm by introducing multi-step internal reasoning. The process is modeled as:

$$y_t \sim \pi(\cdot \mid x_t, \tau_t; \Theta_0) \tag{2}$$

By augmenting the agent with a transient working memory τ_t, the system can decompose complex clinical queries into intermediate reasoning steps (e.g., chain-of-thought) within a single case. However, it still lacks external tool access and cross-case learning mechanisms, meaning its persistent state remains frozen across cases ($S_{t+1} = S_t$).

B.3 L2 Tool-Augmented Agent

The **L2** architecture embodies a *reasoning coupled with tool invocation* paradigm. The decision process can be formulated as:

$$y_t \sim \pi(\cdot \mid x_t, \tau_t; \Theta_0, \mathcal{E}) \tag{3}$$

By activating the external medical toolset ($\mathcal{E} \neq \emptyset$), the agent can ground its reasoning in external knowledge bases or clinical calculators, thereby mitigating hallucinations and improving single-case accuracy. Nevertheless, the absence of a global utility function ($\Omega = \emptyset$) precludes closed-loop feedback, and the system state remains static across cases ($S_{t+1} = S_t$).

B.4 L3 Static Collaborative Multi-Agent

The **L3** architecture shifts the paradigm to distributed joint inference through static multi-agent collaboration. Formally, the joint decision-making process can be modeled as:

$$y_t \sim \pi_{joint}(\cdot \mid \{o_{i,t}\}_{i\in\mathcal{N}}, \tau_t; \Theta_0, \Psi_0, \mathcal{E}, G_0) \tag{4}$$

At this level, the system expands to a collaborative network ($|\mathcal{N}| > 1$). Agents synthesize local observations $o_{i,t}$ under a predefined collaborative topology G_0 (e.g., hierarchical, debate, or parallel structures) and a fixed coordination policy Ψ_0. Although Ψ_0 may execute conditional routing based on the specific case x_t, the underlying architectural rules and model parameters do not learn from outcomes. The persistent state remains strictly constant across cases ($S_{t+1} = S_t$). Thus, dynamic routing among predefined templates is classified as **L3** unless the system fundamentally updates its coordination policy through feedback.

B.5 L4 Adaptive Evolutionary Multi-Agent

The **L4** architecture introduces a feedback-driven adaptive paradigm, breaking the static constraints on the system's internal capabilities. The evolutionary process is formulated as:

$$\begin{aligned}
&r_t = \Omega(y_t, x_t, y^*_{true}) \\
&\Theta_{t+1}, M_{t+1}, \Psi_{t+1} \leftarrow \text{Update}(\Theta_t, M_t, \Psi_t \mid r_t) \\
&y_{t+1} \sim \pi_{joint}(\cdot \mid \{o_{i,t+1}\}_{i\in\mathcal{N}}, \tau_{t+1}; \\
&\qquad \Theta_{t+1}, M_{t+1}, \Psi_{t+1}, \mathcal{E}, G_0)
\end{aligned} \tag{5}$$

With the clinical utility function Ω activated, the system establishes a closed-loop learning mechanism. Driven by the evaluation feedback r_t, the system persistently updates its underlying model parameters (Θ_{t+1} via SFT or Agentic RL), accumulates global experience memory (M_{t+1}), and optimizes its coordination policy (Ψ_{t+1}). However, the defining limitation of **L4** is that these continuous adaptations occur strictly *within a constrained organizational structure*; the collaborative topology G_0 itself remains permanently fixed.

B.6 L5 Medical Multi-Agent Swarms

L5 represents the ultimate paradigm of emergent swarm intelligence, characterized by open-ended topological self-organization. The system evolution can be modeled as:

$$\begin{aligned}
&r_t = \Omega(y_t, x_t, y^*_{true}) \\
&\Theta_{t+1}, M_{t+1}, \Psi_{t+1} \leftarrow \text{Update}(\Theta_t, M_t, \Psi_t \mid r_t) \\
&G_{t+1} \sim \mathcal{F}_{self-org}(\mathcal{G}, G_t, x_{t+1}, M_{t+1}, r_t) \\
&y_{t+1} \sim \pi_{joint}(\cdot \mid \{o_{i,t+1}\}_{i\in\mathcal{N}}, \tau_{t+1}; \\
&\qquad \Theta_{t+1}, M_{t+1}, \Psi_{t+1}, \mathcal{E}, G_{t+1})
\end{aligned} \tag{6}$$

At this frontier, the system completely lifts the structural constraint on the collaborative topology. Beyond optimizing parameters or policies, an **L5** system possesses the capability to spontaneously reconstruct its active topology G_{t+1} from the broader structural space $\mathcal{G}$. Guided by the specific complexity of the incoming case x_{t+1} and historical feedback r_t, the self-organizing operator $\mathcal{F}_{self-org}$ dynamically executes on-demand expert recruitment, edge rewiring, and inefficient branch pruning. This allows the system to seamlessly adapt its organizational form to address open-ended, highly complex clinical scenarios.

B.7 Level Boundaries

To eliminate qualitative ambiguity, we establish definitive boundaries between taxonomy levels based on three dimensions: agent cardinality, cross-case state persistence, and topological mutability.

From L0 to L1: The boundary requires activating a transient working memory. While L0 operates as a single-step mapping function, L1 introduces multi-step internal cognitive reasoning, such as chain-of-thought, to resolve a single clinical case. It remains a single-agent system without external tools or cross-case memory.

From L1 to L2: The transition requires activating external environments or clinical tools. Crucially, all forms of single-agent evolution are fully

contained within L2. Any feedback-driven optimization, including cross-task single-model reinforcement learning, autonomous single-agent prompt tuning, or experience memory updates, remains classified as L2 if the system excludes concurrent agent-to-agent interaction.

From L2 to L3: The demarcation requires multi-agent joint inference under a frozen cross-case state. L3 systems utilize a fixed agent set, fixed role allocation, and a fixed communication topology. Operational designs like multi-round inter-agent debates, in-context retrieval-augmented generation, human-in-the-loop prompt adaptation, or template-based dynamic routing are strictly transient states if cleared after a single case. Because L3 leaves all persistent configurations unchanged across tasks, it represents a static multi-agent framework.

From L3 to L4: The boundary marks the transition from a frozen configuration to a continuous evolutionary trajectory. A system qualifies as L4 if and only if it concurrently satisfies four criteria: a continuous feedback source from a global clinical utility function, persistent updates to system states (including model parameters, global experience libraries, or coordination rules) across distinct patient cases, cross-episode reusability of learned behaviors, and a verifiable, positive adaptation gain. However, L4 adaptation occurs strictly within a permanently fixed organizational topology.

From L4 to L5: The definitive boundary lies in relaxing the fixed-topology constraint. While L4 optimizes internal states within a rigid graph, L5 executes runtime, open-ended structural self-organization. An L5 swarm is uniquely characterized by runtime agent generation or deletion, execution-time communication edge rewiring based on real-time clinical risks, automated rewriting of task decomposition graphs, and topological control under explicit safety constraints to prevent adversarial error propagation.

C Static Coordination Metrics (L3)

C.1 Necessity and Rationale

Conventional evaluation paradigms predominantly treat MMAS as a black box, focusing exclusively on the accuracy of the final output. However, for **L3** static multi-agent systems, this outcome-oriented approach is fundamentally insufficient. High final accuracy can easily mask severe internal inefficiencies, such as redundant token consumption, suppressed minority expert opinions, or forced consensus driven by dominant agents rather than clinical evidence. To address this, it is necessary to establish a suite of white-box coordination metrics. Our design rationale is to deconstruct the collaborative process into measurable dimensions: we first quantify the baseline **clinical utility** to ensure medical safety, and then introduce **communication efficiency** to evaluate the operational cost-benefit ratio. Furthermore, we formalize **consensus quality** and **contribution attribution** to rigorously audit whether inter-agent interactions genuinely yield positive clinical gains and to identify specific failure nodes. Through these metrics, we aim to answer not just whether an **L3** system works, but how *efficiently and justifiably* it collaborates.

C.2 Notations

This level audits whether a predefined collaborative architecture yields measurable clinical value. Following Appendix A and B, $\mathcal{N}$ denotes the agent set, $\mathcal{T}$ the clinical task space, x_t the case input, y_t the joint decision, Ω the clinical utility function, and $S_t = (\Theta_t, M_t, \Psi_t, G_t)$ the persistent system state. At the L3 level, model parameters, memory, coordination policy, and the bounded topology set remain statically locked ($S_{t+1} = S_t$). Additionally, we introduce $\mathcal{D}_{\mathcal{T}}$ as the task distribution over $\mathcal{T}$, $y_t^\star$ as the expert-validated reference answer for x_t, J_0 as the expected clinical utility of the static **L3** system, $\mathcal{C}(y_t \mid \Psi_t, G_t)$ as the computational or communication cost of producing y_t, η_{comm} as communication efficiency, Q_{cons} as consensus quality, $y_{i,t}$ as agent i's independent local decision, A_i as agent i's marginal contribution, and $y_{t\setminus i}$ as the joint decision after removing agent i.

C.3 Foundational Clinical Utility

Recent medical agent systems show that static collaboration can improve clinical question answering and medical reasoning when expert roles, retrieval, and verification modules are fixed during inference. LINS improves evidence-traceable medical Q&A through a retrieval-augmented agentic workflow, while MedAgents uses role-specialized medical agents to conduct multi-round zero-shot reasoning before producing a final decision (Wang et al., 2025b; Tang et al., 2024). Therefore, the **L3** system is evaluated through expected clinical utility:

$$J_0 = \mathbb{E}_{x_t \sim \mathcal{D}_{\mathcal{T}}} \left[\Omega(y_t, x_t, y_t^\star) \right]. \tag{7}$$

Here, $\Omega(y_t, x_t, y_t^\star)$ evaluates the clinical value of decision y_t for case x_t against reference answer $y_t^\star$, including diagnostic correctness, safety, guideline consistency, interpretability, actionability, and risk-adjusted benefit. The joint decision is generated by the static **L3** policy:

$$y_t \sim \pi_{joint}\big(\cdot \mid \{o_{i,t}\}_{i\in\mathcal{N}}, \tau_t; \Theta_0, \Psi_0, \mathcal{E}, G_0\big) \quad (8)$$

C.4 Communication Efficiency

Multi-agent communication introduces token usage, message passing, model calls, latency, and deployment cost. In clinical-scale workloads, orchestrated medical agents can sustain accuracy while using substantially fewer tokens than a single overloaded agent (Klang et al., 2026). MDAgents further shows that assigning collaboration structures according to medical task complexity can avoid unnecessary interaction for simpler cases while preserving specialist discussion for harder decisions (Kim et al., 2024b). We therefore define:

$$\eta_{comm} = \frac{\mathbb{E}_{x_t\sim\mathcal{D}_\mathcal{T}}\left[\Omega(y_t, x_t, y_t^\star)\right]}{\mathbb{E}_{x_t\sim\mathcal{D}_\mathcal{T}}\left[\mathcal{C}(y_t \mid \Psi_t, G_t)\right]}. \quad (9)$$

The cost $\mathcal{C}(y_t \mid \Psi_t, G_t)$ may include token count, inter-agent messages, model invocations, latency, monetary cost, and clinical workflow burden. Thus, η_{comm} quantifies the expected clinical utility per unit of operational overhead.

C.5 Consensus Quality

Collaboration is necessary only when the joint decision improves on individual local decisions. MedARC demonstrates that structured debate, iterative revision, and confidence-aware aggregation can improve medical QA reliability (Miao et al., 2026). SMR-Agents similarly simulates multidisciplinary consultation for medical visual question answering, using expert-agent peer review to refine both the answer and its rationale (Wang et al., 2026). Let $n = |\mathcal{N}|$, and let $y_{i,t}$ denote the independent decision of agent i based on its local observation. Consensus quality is:

$$Q_{cons} = \mathbb{E}_{x_t\sim\mathcal{D}_\mathcal{T}}\Bigg[\Omega(y_t, x_t, y_t^\star) - \frac{1}{n}\sum_{i=1}^{n}\Omega(y_{i,t}, x_t, y_t^\star)\Bigg]. \quad (10)$$

If $Q_{cons} > 0$, collaboration provides positive utility gain. If $Q_{cons} \approx 0$, the collaborative structure is weakly justified. If $Q_{cons} < 0$, interaction may amplify errors or induce harmful consensus.

C.6 Contribution Attribution and Failure Tracking

Medical multi-agent systems require attribution because final-answer accuracy can hide flawed collaboration, suppressed minority opinions, or information loss during synthesis. MedAgentAudit explicitly diagnoses collaborative failure modes in medical multi-agent systems (Gu et al., 2025). WSI-Agents provides a complementary pathology setting in which task allocation and verification modules make component roles auditable across multimodal whole-slide-image tasks (Lyu et al., 2025). Motivated by counterfactual attribution, the marginal contribution of agent i is:

$$A_i = \mathbb{E}_{x_t\sim\mathcal{D}_\mathcal{T}}\left[\Omega(y_t, x_t, y_t^\star)\right] - \mathbb{E}_{x_t\sim\mathcal{D}_\mathcal{T}}\left[\Omega(y_{t\setminus i}, x_t, y_t^\star)\right]. \quad (11)$$

Here, $y_{t\setminus i}$ is the joint decision produced after removing agent i while preserving the remaining protocol. Positive A_i indicates useful contribution, near-zero A_i indicates redundancy, and negative A_i indicates that the agent may introduce noise, bias, hallucinated evidence, or harmful reasoning. This metric supports node ablation, communication-link auditing, and clinical failure tracing.

D Adaptive Evolution Verification (L4)

D.1 Necessity and Rationale

As medical multi-agent systems transition from **L3** to **L4** adaptive evolution, the evaluation dimension must fundamentally shift from "static snapshots" to "dynamic trajectories." Traditional metrics are inherently state-agnostic and fail to capture a system's capacity to learn, adapt, or regress over repeated clinical encounters. Evaluating an **L4** system requires verifying not only that it adapts to feedback, but that this adaptation is clinically meaningful and robust against overfitting to specific institutional workflows. Therefore, our design rationale focuses on capturing the temporal and distributional dynamics of the system. We introduce **simulated learning curves** to track the per-time-step evolutionary gain, mathematically distinguishing genuine cognitive accumulation from evolutionary stagnation or catastrophic forgetting. To further prevent the illusion of progress caused by environment-specific shortcuts, we formalize **cross-dataset transfer metrics**. By measuring the cross-task utility gap, we ensure that the policies and memories evolved in the source distribution

can reliably generalize to target clinical environments, thereby guaranteeing the safety and scalability of **L4** systems in real-world deployments.

D.2 Notations

At **L4**, environmental feedback drives cross-turn updates to parameters, global memory, or coordination policy. Feedback is defined as:

$$r_t = \Omega(y_t, x_t, y_t^\star). \tag{12}$$

Accordingly, **L4** may involve $\Delta\Theta$, ΔM, or $\Delta\Psi$, but remains bounded by the predefined topology set G_0. Open-ended reconstruction of collaborative topology belongs to **L5** rather than **L4**. Additionally, we introduce J_t as the expected clinical utility under state S_t, ΔJ_t as the per-time-step evolutionary gain, $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$ as source and target task distributions, J_{src} and J_{tgt} as their corresponding utilities, and Γ_{gen} as the cross-task utility gap.

D.3 Simulated Learning Curves

Adaptive medical agents should be evaluated across repeated encounters rather than only on one-shot cases. Doctor-R1 trains an AI doctor agent through experiential reinforcement learning in a multi-agent interactive environment, optimizing both clinical decision-making and inquiry behavior (Lai et al., 2026). MedAgentSim builds a self-evolving doctor–patient–measurement simulation in which agents refine diagnostic strategies through multi-turn interactions and experience-based retrieval (Almansoori et al., 2025). We define the expected clinical utility at state S_t as:

$$J_t = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{T}}\left[\Omega(y_t, x, y^\star \mid S_t)\right]. \tag{13}$$

After a state update driven by Δ, the per-time-step evolutionary gain is:

$$\begin{aligned}\Delta J_t &= \frac{1}{\Delta t}\left(J_t - J_{t-1}\right) \\ &= \frac{1}{\Delta t}\Big(\mathbb{E}_{x\sim\mathcal{D}_\mathcal{T}}\left[\Omega(y_t, x, y^\star \mid S_t)\right] \\ &\qquad - \mathbb{E}_{x\sim\mathcal{D}_\mathcal{T}}\left[\Omega(y_{t-1}, x, y^\star \mid S_{t-1})\right]\Big).\end{aligned} \tag{14}$$

Thus, $\Delta J_t > 0$ indicates beneficial adaptation, $\Delta J_t \approx 0$ indicates stagnation, and $\Delta J_t < 0$ indicates possible degradation or forgetting.

D.4 Cross-Dataset Transfer

Internal evolution does not guarantee that a medical agent policy transfers beyond its source setting. MedAgentGym provides a scalable agentic training environment spanning 129 biomedical task categories and multiple real-world biomedical scenarios, making cross-task evaluation central to policy evolution (Xu et al., 2026). CASSIA further shows how a modular multi-agent biomedical analysis system can be tested across heterogeneous benchmark datasets and specialized biological contexts while retaining interpretability and quality assessment (Xie et al., 2025). Let:

$$J_{src} = \mathbb{E}_{x\sim\mathcal{D}_{src}}\left[\Omega(y_t, x, y^\star)\right], \tag{15}$$

$$J_{tgt} = \mathbb{E}_{x\sim\mathcal{D}_{tgt}}\left[\Omega(y_t, x, y^\star)\right], \tag{16}$$

where $\mathcal{D}_{src}$ is the source distribution on which the system primarily evolves, and $\mathcal{D}_{tgt}$ is a target distribution such as a new hospital, disease group, or external dataset. The cross-task utility gap is:

$$\begin{aligned}\Gamma_{gen} &= J_{src} - J_{tgt} \\ &= \mathbb{E}_{x\sim\mathcal{D}_{src}}\left[\Omega(y_t, x, y^\star)\right] \\ &\qquad - \mathbb{E}_{x\sim\mathcal{D}_{tgt}}\left[\Omega(y_t, x, y^\star)\right].\end{aligned} \tag{17}$$

A large Γ_{gen} indicates possible overfitting to source tasks, institutional workflow, or feedback-specific shortcuts. This metric should be reported with J_{tgt}, because a small gap is meaningful only when target-domain utility is clinically sufficient.

E Threat Modeling

We introduce threat modeling (Xiong and Lagerström, 2019) to map abstract security risks to specific clinical attack paths. Specific analysis of the methods which attackers employ to disrupt diagnostic consensus or pilfer sensitive data within clinical workflows serves to identify distinct attack vectors. This analysis not only enables architects to embed defense mechanisms during the design phase by identifying adversarial patterns, but also lays a logical foundation for constructing robust system security arrangements in subsequent steps.

Safety Threats: Attackers exploit consensus mechanisms via hijacking, launching Sybil attacks to generate counterfeit nodes that override genuine experts through numerical superiority, thus undermining treatment plans. Another path is clinical RAG poisoning, where attackers infiltrate backend external knowledge bases to tamper with critical medical guidelines or medication parameters. Upstream agents retrieve incorrect information and formulate flawed plans, which downstream agents tend to accept out of collaborative inertia, leading to cascading errors and adverse medical events.

Privacy Threats: Attackers exploit MMAS's dependence on frequent agent interactions for data theft. Malicious entities may impersonate legitimate nodes to query clinical agents. Due to the lack of granular verification, these agents may leak complete medical records. Even without raw data, attackers can conduct inference attacks by intercepting dialogue history and intermediate reasoning. By correlating details like symptoms, drug preferences, or transfers, they can reconstruct patient identities and undisclosed illnesses.

Ethics Threats: Attackers manipulate the value alignment mechanism to seek illicit interests. Agents may engage in profit-driven collusion, recommending unnecessary high-priced examinations while deliberately downplaying side effects, depriving patients of their right to informed consent. Another threat is discriminatory feedback injection, where attackers poison memory or feedback loops with biased fake data. Agents internalize these biases through fine-tuning, leading to systematic discrimination in future diagnoses.

What's the Fix? The above threat analysis indicates that MMAS introduces collective risks stemming from dynamic interactions, which are distinct from the isolated risks of single-agent systems. While defenses targeting individual models remain important, they often fail to fully address these complex system-level challenges. Therefore, we need to construct a governance framework from two complementary dimensions. The agent-centric view aims to enhance the collaborative robustness within the system through technical means. The human-centric view strives to ensure external clinical supervision and ethical alignment through process design. This dual-perspective approach forms the core logic of the security governance solutions presented in subsequent chapters.

F Literature Taxonomy

Figure 4 summarizes the works reviewed in Section 2. Specifically, it groups the cited papers into four branches: **L3** collaborative medical agents, **L4** adaptive medical agents, knowledge utilization, and benchmarks.

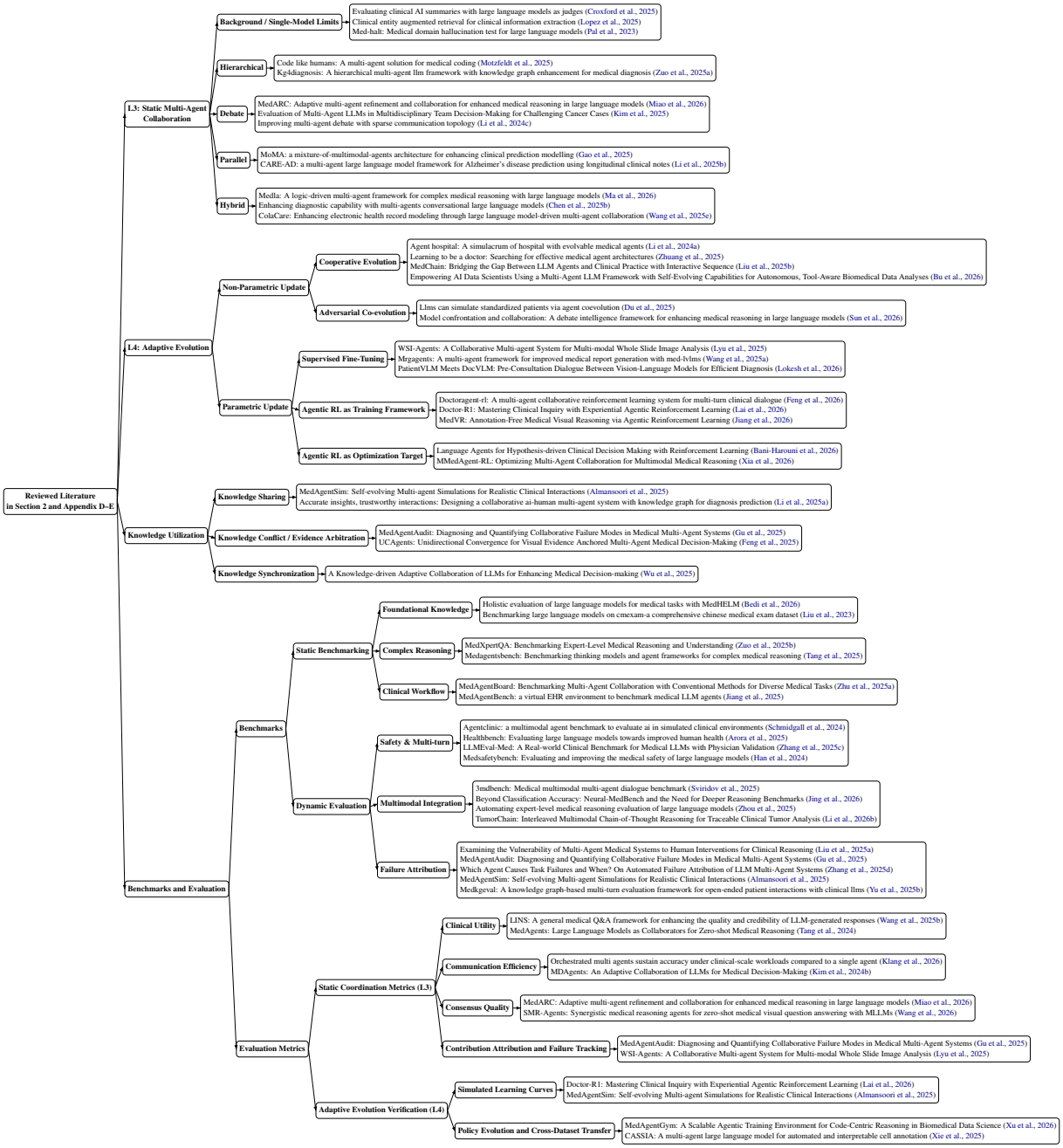


Figure 4: Taxonomy tree of the core literature, organized into four main branches: L3 static multi-agent collaboration, L4 adaptive evolution, knowledge utilization, and benchmarks/evaluation metrics.