LLM-Based Multi-Agent Systems for Clinical Workflows: A Survey of AI Hospitals

Zonghai Yao
University of Massachusetts, Amherst
University of Massachusetts, Lowell
VA Bedford Health Care
zonghaiyao@umass.edu

Hong Yu
University of Massachusetts, Lowell
University of Massachusetts, Amherst
VA Bedford Health Care
Hong_Yu@uml.edu

Abstract

This survey reviews LLM-based multi-agent systems for clinical and healthcare workflows, including diagnosis, triage, consultation, discharge, mental health, and EHR-linked decision support. We define AI hospitals as workflow-level clinical systems in which agents take explicit roles, hand off shared state, use EHR- or guideline-grounded tools, and operate with safety gates and audit-ready logs. We argue that these systems should be compared at the workflow level, rather than only by model components or end-task accuracy, because clinical action, evidence, and accountability are expressed through state transitions and handoffs. We organize the literature through a workflow-level taxonomy covering roles and handoffs, memory and evidence, tools, and reasoning, control, and escalation. We further synthesize major workflow settings and task families, introduce a four-layer evaluation stack spanning safety, process, outcome, and operations, and connect model capabilities to workflow observables relevant to deployment. Finally, we present Integration Readiness Levels (IRL1-IRL6), task-level instrumentation requirements, and recurring workflow failure modes as a practical framework for comparing, evaluating, and deploying clinical LLM agents and AI hospitals. [1]

1 Introduction

Large Language Models (LLMs) are moving from isolated text generation toward agentic work inside clinical workflows. This shift is visible in medicine, where systems are increasingly evaluated on interactive patient tasks, conversational diagnosis, and workflow-grounded virtual-EHR settings rather than only on exam-style or static QA benchmarks (Johri et al., 2025; Tu et al., 2025; Jiang et al., 2025; Bedi et al., 2026). At the same time, recent reviews still find limited prospective evidence, sparse operational reporting, and weak links from benchmark gains to real workflow outcomes (Bedi et al., 2025b; Agrawal et al., 2025). The gap is now less about whether a model can answer a question, and more about whether a system can carry a case across triage, consultation, and discharge without losing state, evidence, or accountability.

We therefore take a workflow-level view. The basic unit of analysis is not an isolated answer but a workflow state transition: one role updates the case state, attaches evidence, and passes the case to the next step under an explicit control policy. This framing makes it possible to study whether handoffs are complete, whether escalation happens for the right reason, whether a later action can be traced back to earlier evidence, and whether operations metrics such as latency or human touches remain acceptable.

Orientation example: triage → consult → discharge

Triage determines urgency and collects the first structured signal. *Consultation* updates the working diagnosis, links evidence to proposed actions, and decides what still needs clarification. *Discharge* turns the current plan into instructions, follow-up, and escalation criteria.

Across the three steps, the same four questions recur: (1) who owns the next action, (2) what state must survive the handoff, (3) what evidence justifies the step, and (4) what gate can block or escalate progression.

In this survey, *AI hospital* refers to a workflow-level multi-agent clinical simulation or deployment with defined roles, shared state across handoffs, evidence-grounded tools, safety gates, and longitudinal traces. These are minimum conditions for hospital-like behavior, not a checklist

[1] To appear in the proceedings of the Main Conference on the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45772–45793

of every hospital function. A system may support consultation, triage, discharge, education, or mental-health follow-up, but it enters this scope only when actions are organized across roles and stages, with persistent context, explicit accountability, and audit-ready logs. The main design axes are workflow span, evidence grounding, control policy, and autonomy scope.

This boundary is intentionally narrow. In scope are multi-agent simulators and clinical systems that coordinate across stages of care, such as triage, consultation, discharge, bed management, care transitions, and training wards linked to EHRs or realistic surrogates. Out of scope are single-agent chatbots, generic role play without persistent workflow state, single-turn QA systems, and tools that lack safety rules, longitudinal memory, or handoff structure. A single-model triage assistant can still be useful, but it is not an AI hospital in the sense used here if it does not preserve shared workflow state, support structured handoffs, and expose audit-ready traces.

This survey is organized around three practical questions. What counts as an AI hospital, how should its design choices be described, and what evidence justifies stronger autonomy? To answer them, we define the minimum boundary of an AI hospital, organize the design space around roles and handoffs, memory, evidence, and tools, and reasoning, control, and escalation, map the main workflow settings and task families, and introduce an evaluation framework with an Integration Readiness Level (IRL) playbook and task-level instrumentation requirements.

This perspective differs from recent surveys of healthcare chatbots, medical multi-agent systems, and LLM agents in medicine (Elkamouchi et al., 2024; Laymouna et al., 2024; Le et al., 2023; Wang et al., 2025b; Bedi et al., 2025b; Tariq, 2024). The emphasis here is on explicit state and handoffs, operations-aware evaluation, and deployment readiness as a staged protocol rather than a loose claim. The rest of the paper follows that logic: Section 2 defines the core design space, Section 3 maps workflow settings, Section 4 turns evaluation and deployment into testable artifacts, and Section 5 ties future progress to concrete workflow failure modes.

2 Core Design Space

Appendix Figure 1 expands the full taxonomy. In the main text, Table 1 anchors the design space in the running example by showing, for each step, the role owner, the state that must survive the handoff, and the gate that can block or escalate progression. These same fields recur throughout the survey and mark the difference between a workflow unit and an isolated assistant.

Step	Primary owner	State and evidence carried forward	Gate before next step
Triage	Intake / triage	Acuity, red flags, first evidence links, and missing-history flags	Block for missing signals; escalate high risk
Consult	Clinician / specialist	Working diagnosis, cited evidence, tests ordered, unresolved uncertainty	Ask for more evidence; abstain or escalate on conflict
Discharge	Coordinator / educator	Instructions, follow-up plan, return precautions, and approvals	Block unsafe counseling; escalate low comprehension or risk

Table 1: Running example that anchors the main-text design space. Across triage, consultation, and discharge, the same workflow questions recur: who owns the step, what must survive the handoff, and which gate can block or escalate progression.

2.1 Roles & Handoffs

The first design decision is whether the workflow truly needs multiple roles. A single agent may be enough for a narrow, low-risk step with strong human review. Multi-agent structure becomes useful when the case crosses care stages, specialties, or approval boundaries, because ownership changes and the next role must inherit explicit state rather than guess it. Appendix Table 6 turns this into a checklist built around workflow span, role diversity, audit needs, escalation needs, and operations coupling. [2]

Patient-facing roles. Patient-facing roles determine what information enters the workflow. Patient agents are used for symptom elicitation, consultation training, education, and interactive evaluation (Yu et al., 2025; Liu et al., 2025c; Johri et al., 2025). Psychological patient agents extend this to mood variation, treatment resistance, and behavior-change dynamics in therapeutic settings (Louie et al., 2024; Wang et al., 2024e, 2025a). Resident or population agents support pathway and public-health simulation when the workflow needs behavior before formal presentation to care (Li et al., 2024b). Methods such as CoT or RAG matter here mainly when they stabilize persona, symptom revelation, and longitudinal consistency rather than only making the dialogue more fluent (Tu et al., 2025; Wu et al., 2026).

[2] Additional details are provided in Appendix A.1 and A.2.

Clinical professional roles. General doctor agents usually own staged history taking and the first working assessment (Johri et al., 2023; Liu et al., 2025c). Specialist agents matter when the case requires domain-specific constraints or structured disagreement, such as rare disease, imaging, or multidisciplinary review (Kim et al., 2024; Chen et al., 2026; Rose et al., 2025). Nurse, therapist, technician, student, and examiner roles extend the workflow to triage, counseling, testing, and training (Bao et al., 2024; Schmidgall et al., 2024). The gain is not simply more experts. It is the ability to localize responsibility, make specialty assumptions visible, and reconcile disagreement before the workflow moves on (Chen et al., 2025c; Liu et al., 2026; Yan et al., 2025a).

Planning and orchestration roles. Planning roles decompose a case into steps, decide what information is still missing, and route the case across reception, triage, consultation, or follow-up (Yu et al., 2025; Yue et al., 2024; Schmidgall et al., 2024; Li et al., 2024b). These roles are useful when the system must coordinate more than one local decision, for example when triage output should determine what the consult agent asks next or which service receives the patient.

Judging, critique, and recording roles. Judge, critic, decision, and recording roles become useful when the workflow needs explicit reconciliation, guideline checking, or auditable summaries rather than one more free-form discussion turn (Johri et al., 2023; Hong et al., 2024; Tang et al., 2024; Ke et al., 2024; Chen et al., 2025a; Zhao et al., 2025; Liu et al., 2025a). They support reason-coded escalation, evidence reconciliation, and replayable handoffs. In practice, multi-agent overhead is justified mainly when these accountability functions are measurable, not simply when the model can generate more branches of reasoning.

2.2 Memory, Evidence, & Tools

The second design decision is what the next role must recover after a handoff. After triage, the consult agent should not have to guess the acuity signal, the active problems, the tests already ordered, or the evidence behind the current plan. In AI hospitals, memory and tools are therefore core workflow infrastructure. They make case state recoverable, evidence attributable, and actions executable under audit. [3]

Long-term memory and evidence sources. Internal memory helps fill missing attributes or adapt to clinical data distributions through pretraining, instruction tuning, or fine-tuning (Li et al., 2024d; Wang et al., 2024f). External memory provides stable references such as EHRs, guidelines, triage manuals, knowledge graphs, and curated stores of procedures or prior trajectories (Lu et al., 2024; Wang et al., 2026; Lai et al., 2025). Dynamic updating through RAG or APIs matters when the workflow depends on current records, new evidence, or expert feedback rather than on static knowledge alone (Yang et al., 2024). The design question is not only where knowledge lives, but how each supporting item can be recovered, versioned, and tied to a later action.

Working memory and handoff recovery. Working memory carries the active case state across roles. In practice, that means summaries, intermediate decisions, evidence IDs, pending questions, and escalation markers that let a new agent resume the workflow without reconstructing it from scratch (Liu et al., 2025c; Hong et al., 2024; Tang et al., 2024). Appendix Table 7 compares memory designs by what survives the handoff, where the design fits best, and what failure it tends to introduce. This trade-off matters because memory quality changes handoff completeness, replayability, and escalation quality, especially in long-context or virtual-EHR settings (Wornow et al., 2025; Liu et al., 2025a; Jiang et al., 2025).

Evidence access tools. Some tools primarily help the system find and anchor evidence. Retrieval systems and knowledge graphs connect decisions to records, guidelines, or structured medical relations (Du et al., 2025; Kim et al., 2024; Wu et al., 2025; Xiong et al., 2025; Wang et al., 2024c; Tran et al., 2025, 2026). Curated repositories can also provide reusable procedures or prior trajectories, which is different from retrieving a document because the returned object is an action template or workflow memory rather than a passage (Wang et al., 2026; Lai et al., 2025).

Execution and verification tools. Other tools make regulated checks or downstream actions executable. Decision trees and calculators support explicit guideline checks (Yang et al., 2024; Li

[3] Additional details are provided in Appendix A.3 and A.4.

et al., 2023; Goodell et al., 2025). Virtual-EHR tools expose provenance and stage-level failures in a controlled environment (Jiang et al., 2025). LLM-as-KB styles of retrieval support flexible synthesis, but they still need explicit provenance when used inside a workflow (Yue et al., 2024; Frisoni et al., 2024).

Multimodal and research tools. Multimodal tools tie image or sensor evidence to downstream reasoning, which matters when diagnostic claims must remain linked to what the system actually saw (Li et al., 2024d; Yang et al., 2024; Fallahpour et al., 2025; Wang et al., 2025d; Acosta et al., 2022). Computational reasoning tools support code execution, structured analysis, and research automation (Wang et al., 2024g; Hong et al., 2024). Across these categories, the main point is the same: tools are not optional accuracy boosters. They are the mechanism that turns evidence and actions into auditable workflow objects. Appendix Table 8 compresses these tool classes into workflow function, audit artifact, and failure signal.

2.3 Reasoning, Control, & Escalation

The third design decision is control. The main question is not how sophisticated the reasoning looks, but what policy governs action. In a hospital-like workflow, the system must decide when to continue, when to ask for missing evidence, when to abstain, and when to escalate. The useful observables are therefore omission, disagreement, abstention, evidence coverage, latency, and rollback behavior. Appendix Table 9 condenses common reasoning, memory, and control choices into fit, practical cost, and logging obligations. [4]

Direct reasoning. Single-path reasoning works best when the task is well specified and the guideline surface is stable, especially when intermediate steps can be tied to explicit evidence acquisition or calculators (Li et al., 2024d; Chen et al., 2025a; Rose et al., 2025). CoT-style pipelines may improve transparency, but they can still miss alternatives and propagate early errors if no later gate checks the path (Schmidgall et al., 2024). Their dominant failure mode is silent omission, so they still need citation coverage and tool-call traces.

Multi-path reasoning. Parallel branches, debate, or voting can help when the workflow genuinely contains specialty diversity or unresolved uncertainty (Kim et al., 2024; Chen et al., 2026, 2025c; Liu et al., 2026). They also impose extra token cost, latency, controller burden, and reconciliation work (Zhu et al., 2025; Zhao et al., 2025). The benefit appears only when disagreement is surfaced, logged, and tied to explicit reconciliation or escalation rules rather than hidden inside a final merged answer.

Feedback sources. Feedback can come from clinicians, tools, knowledge bases, or later turns in the interaction. External feedback updates the plan with new evidence or expert correction (Johri et al., 2023; Li et al., 2024c). Self-feedback and reflection can reduce local inconsistencies before commit (Louie et al., 2024). Symbolic controllers and planner agents extend this by coordinating tool calls and alternative paths rather than only editing text (Hong et al., 2024; Liu et al., 2025a, 2024).

Control policies and gates. The key control question is whether feedback is actually enforced by policy, evidence, or uncertainty gates and recorded as an audit-ready trace (Goodell et al., 2025; Wang et al., 2025d). This is what turns reasoning style into admissible autonomy. A useful system does not just reason more. It exposes an autonomy dial that determines when the workflow continues, when it requests missing evidence, when it abstains, and when it escalates.

3 Workflow Settings and Task Families

AI hospitals differ less by raw model family than by the workflow unit they instrument. A triage assistant, a consultation system, a discharge educator, and a mental-health follow-up agent may share components, but they do not fail in the same way and they do not require the same logs. We therefore group the literature by workflow settings and task families, because this is the level where multi-agent structure, handoffs, and deployment claims become comparable. [5]

3.1 Care Workflows with Direct Clinical Impact

The main clinical workflow families differ by where the system sits in the care pathway and by which failure surface dominates. The same architecture can look adequate in one family and brittle

[4] Additional details are provided in Appendix A.5.

[5] Additional details are provided in Appendix B.

in another, because the demands on handoff quality, evidence linking, and escalation policy change with the workflow.

End-to-end clinical workflow simulation and virtual wards. These settings are closest to the running example of triage, consultation, and discharge. The system must acquire missing evidence, update hypotheses across turns, and preserve state through multiple roles or stages (Johri et al., 2025; Tu et al., 2025; Jiang et al., 2025; Yan et al., 2025a; Liu et al., 2025a). Their main value is stage-level visibility. Rather than only scoring the final answer, they show where evidence was missing, where a handoff failed, and where escalation should have happened (Liu et al., 2025c; Johri et al., 2025; Jiang et al., 2025; Schmidgall et al., 2024; Li et al., 2024b). This is why virtual wards and virtual-EHR benchmarks are so useful in this literature: they turn hidden workflow mistakes into observable events.

Consultation, multidisciplinary coordination, and complex decision making. These workflows matter when specialty-specific constraints or persistent uncertainty make a single reasoning path brittle. Representative systems ask follow-up questions, update differentials over time, reconcile structured evidence, or convene multidisciplinary teams for rare and complex cases (Zhao et al., 2026; Rose et al., 2025; Chen et al., 2025a; Liu et al., 2025b; Kim et al., 2024; Chen et al., 2026, 2025c; Liu et al., 2026; Tang et al., 2024). The relevant failure surface is not only diagnostic error. It also includes hidden disagreement, premature convergence, and poorly justified commitments. Multi-agent structure is useful here only when it makes disagreement, reconciliation, and escalation more explicit and more measurable.

Triage, routing, care transitions, and discharge communication. These workflows sit closest to operational constraints. Inputs are noisy, approval boundaries are common, and the quality bar is set by handoff completeness and downstream usability rather than by one diagnosis (Lu et al., 2024; Bao et al., 2024; Tao et al., 2026; Goh et al., 2025; Yao et al., 2025). Patient-facing transition tasks, including discharge communication and follow-up preparation, make this especially clear, because the system must preserve what was decided, what was explained, and what still requires escalation. In these settings, workflow logs are part of the task definition rather than an afterthought.

3.2 Longitudinal, Care-Adjacent, and Research Workflows

Longitudinal patient interaction and mental health. This family emphasizes persistent state, risk tracking, and calibrated escalation rather than one-shot problem solving. The hard part is maintaining coherent patient state across turns, recognizing crisis or relapse signals, and updating the control policy at the right time in counseling or behavior-change workflows (Wu et al., 2026; Wang et al., 2025a; Yao et al., 2026a). Systems in this family often rely on expert-authored behavioral principles or structured therapeutic constraints to reduce unsafe drift over long interactions (Louie et al., 2024; Wang et al., 2024e). Memory, refusal policy, and reason-coded escalation are tightly coupled here.

Care-adjacent clinical data workflows. Not every relevant AI-hospital task sits at the bedside. Some systems support EHR analytics, tool building, knowledge curation, or clinical-trial matching, and they become part of the AI-hospital landscape when they add provenance, verification, or handoff value to clinical workflows rather than functioning as standalone utilities (Bedi et al., 2026; Wornow et al., 2025; Shi et al., 2024b; Wang et al., 2024g; Yue et al., 2024; Shi et al., 2024a). Their dominant risks are usually provenance gaps, brittle connectors, or evaluation that stops at retrieval quality instead of tracing downstream workflow effects.

Scientific discovery and research workflows. Care-adjacent research workflows extend the same design logic to experiment planning, evidence synthesis, and tool-mediated scientific collaboration (Swanson et al., 2025; Liu et al., 2024). They still rely on explicit roles, external tools, structured traces, and validation under operational constraints, even though the immediate outcome is a research artifact rather than a bedside action. Across all of these families, what changes most is the dominant failure surface, which is why later sections focus on task-specific instrumentation and staged autonomy rather than on one universal benchmark.

4 Evaluation and Deployment Readiness

Evaluation and deployment readiness in AI hospitals require more than benchmark accuracy, because the system must remain safe across steps,

handoffs, tools, and operational constraints before stronger autonomy can be justified. This section starts with what should be measured, then connects model capability to workflow behavior, specifies the minimum logs for common integration tasks, and finally turns readiness into a staged deployment protocol.

4.1 The Evaluation Stack

Evaluation in AI hospitals is moving away from isolated exam-style assessment toward interactive patient tasks, benchmark suites that span clinical task families, virtual-EHR environments, and early prospective studies (Johri et al., 2025; Bedi et al., 2025b, 2026; Zhou et al., 2025; Agrawal et al., 2025). For workflow-level systems, the unit of evaluation is not only the final answer but also behavior across steps and handoffs. That requires instrumented logs, such as tool calls, evidence links, handoff traces, and escalation events, together with structured rubrics that make process quality measurable.

We organize these signals as a four-layer stack: safety, process, outcome, and operations. In the running example of triage → consult → discharge, safety asks whether unsafe advice was blocked, process asks whether each step carried complete state and evidence, outcome asks whether the final disposition or education was correct and useful, and operations asks how much delay, token cost, and human effort the workflow consumed. The same run should therefore support replay, audit, and operational analysis rather than reporting only a final score.

4.2 From Model Capability to Workflow Observables

A capability matters in this setting only when it changes a workflow observable, a logging obligation, or the level of autonomy that can be justified. Better uncertainty calibration matters when high-risk steps are more reliably deferred or escalated (Johri et al., 2025; Tu et al., 2025). Better retrieval matters when consult or discharge outputs carry record-linked evidence rather than uncited claims (Johri et al., 2025; Liu et al., 2025a). Better long-context handling matters when a later role receives a cleaner handoff and the case remains replayable over time (Wornow et al., 2025; Jiang et al., 2025). Better tool reliability matters when regulated actions become executable with provenance and rollback traces (Goodell et al., 2025). Better multimodal grounding matters when image- or sensor-linked evidence stays attached to the decision chain (Fallahpour et al., 2025; Wang et al., 2025d).

This bridge changes how claims should be written. It is not enough to say that a model reasons better, retrieves more, or uses tools more often. The useful claim is what the system can now safely do, what additional evidence must now be logged, and which stronger autonomy claim becomes plausible. Table 2 makes that translation explicit.

4.3 Why Benchmarks Are Not Deployment Readiness

Benchmarks and deployment readiness are not interchangeable (Bean et al., 2026; Machcha et al., 2026; Asadi et al., 2026). Many studies still evaluate simplified settings and do not report the logs needed for replay, escalation, and operational analysis (Bedi et al., 2025b; Agrawal et al., 2025). Workflow-grounded benchmarks narrow part of this gap by exposing state transitions, evidence chains, and staged failures in virtual-EHR or multi-agent settings (Jiang et al., 2025; Liu et al., 2025a; Zhu et al., 2025), but they are still only one part of the evidence needed for deployment claims.

Judge-based and rubric-based scoring are useful for process evaluation and for scaling benchmark coverage, but judge fidelity and reasoning-grading quality remain part of the evaluation problem itself (Zhou et al., 2025; Bedi et al., 2025a; Arora et al., 2025). They help most when paired with logs that make the underlying workflow visible, rather than treated as a substitute for that visibility.

The closer one gets to deployment, the more claims need replay, escalation evidence, and operations evidence in the same package. OSCE-style interactive exams, virtual-EHR benchmarks, broader medical task suites, and early prospective studies each contribute part of that package (Johri et al., 2025; Bedi et al., 2026). The distinction is simple: benchmarks ask whether a system can perform under a designed task, while deployment readiness asks whether autonomy claims remain justified after replay, escalation, and operations are inspected together.

4.4 Instrumentation for Clinical Integration Tasks

A handoff system and a bed-management system do not fail in the same way. The first usually fails by dropping fields, provenance, or evidence links.

Capability gain	Workflow change that should appear	Evidence to report	Stronger claim it can support
Better calibration / abstention	High-risk steps defer or escalate instead of silently committing	Reason-coded abstentions; escalation precision/recall; override outcomes	Approval-required pilot or high-risk gate with monitored coverage
Evidence-grounded retrieval	Consult and discharge outputs carry record-linked evidence rather than uncited claims	Citation or evidence-ID coverage; evidence-gate pass rate; citation-to-record consistency	Replay-ready evidence chains for shadow evaluation
Long-context / longitudinal state	Later roles receive cleaner handoffs and replay remains consistent across visits	Handoff completeness; replay success; cross-visit consistency	Longitudinal shadow replay and multi-visit workflow evaluation
Reliable calculators / tool use	Regulated checks and actions execute with provenance and rollback traces	Tool success/failure; guideline adherence@step; rollback rate	Controlled use for triage, orders, or discharge checks
Multimodal grounding	Image- or sensor-based claims remain linked to observed evidence	Cross-modal consistency; image-to-report links; diagnostic audit pass	Auditable commit in imaging or monitoring workflows

Table 2: Capability-to-workflow bridge. A model capability becomes deployment-relevant only when it changes a visible workflow behavior, creates a new logging obligation, or supports a stronger autonomy claim.

Integration task	What usually fails first	Minimum artifacts to keep	Release signals to report	Weak claim to avoid
Clinic scheduling / bed management	Queue state ignored; unsafe speedup under load	Queue snapshots; bed-board state; transport timestamps; overload escalation events	Overload escalation precision/recall; severity-weighted errors; TTD; throughput; P95 latency	Mean latency or throughput only
Order routing / prior authorization	Silent fallback; missing approval state; API delay	Request payloads; approval-state transitions; tool/API versions; retries; rollback events	Unsafe-order block rate; guideline adherence@step; order delay; error severity	Benchmark gain without version or rollback logs
Shift handoffs (ED → ward)	Fluent but lossy summary; missing provenance	Structured handoff fields; evidence IDs; retrieved record fields; cross-shift snapshots	Handoff completeness; citation coverage; cross-shift consistency; handoff time	Final-answer judging only
Discharge education / follow-up	One-shot counseling; no retention check	Pre/post checks; multilingual template version; follow-up callbacks; low-comprehension escalations	Unsafe-counseling block rate; education gain; 7/30-day retention proxy; latency	Post-test only, no follow-up
Live EHR integration (FHIR)	Permission drift; schema drift; PHI leakage	Access audits; consent IDs; queried resources; de-identification audits; chaos-test logs	PHI leakage rate; citation-to-record consistency; chaos-test pass rate; API latency / availability	Retrieval score only, no access or schema audit

Table 3: Instrumentation map for common integration tasks. Each row starts with the dominant early failure surface, then specifies the minimum artifacts and release signals needed before stronger autonomy can be claimed. TTD denotes time-to-disposition, PHI protected health information, and FHIR Fast Healthcare Interoperability Resources.

The second fails by ignoring queue state, transport delay, or latency tails. Table 3 therefore starts from the first failure surface in each integration task, then lists the minimal artifacts and release signals that make that failure visible.

4.5 IRL-Based Gating from Sandbox to Deployment

Integration Readiness Levels (IRL1–IRL6) turn readiness into a staged deployment protocol rather than a loose label. Table 4 organizes the ladder into simulation, replay/pilot, and rollout phases. For each level it states the highest autonomy the system may take and the evidence required before promotion, so readiness depends on progressively stronger replay, escalation, and monitoring evidence rather than on one strong benchmark number.

This is where evaluation and governance meet. IRL1 and IRL2 stay in simulation. IRL3 and IRL4 require replayability, evidence chains, and monitored escalation before limited pilot use. IRL5 and IRL6 add drift control, incident response, canary strategy, and multi-site stability. The thresholds shown here are protocol defaults, not universal constants, and should be calibrated to task risk, site policy, baseline comparator, and acceptable failure budget.

4.6 Synthetic Data as Evidence Generation, Not Deployment Evidence

In this survey, synthetic data is useful for training and stress testing, not as deployment evidence. It can generate dialogues, workflow state transitions, missing-information cases, and interaction patterns that are hard to collect at scale (Tu et al., 2025; Johri et al., 2025). In healthcare this is appealing because synthetic data can expand coverage while reducing direct privacy exposure, al-

Phase	IRL	Typical setting	Highest autonomy allowed	Evidence needed before promotion
Simulation	IRL1	Static or scripted sandbox	Manual review before any commit	Versioned configuration; baseline task success; unsafe-output rate below site threshold
	IRL2	Noisy, missing-info, or adversarial simulation	Manual review plus safety gateway	Seeded high-risk coverage; jailbreak resistance; citation coverage; block-time logs
Replay / pilot	IRL3	Shadow replay on real data, no live impact	Model proposes, humans decide	Replayable evidence chains; handoff completeness; versioned prompts/models; PHI leakage equal to zero in audit
	IRL4	Limited human-in-the-loop pilot	Auto-suggest for bounded steps, mandatory approval for high-stakes steps	Escalation recall on labeled high-risk cases; zero severe misses; override postmortems; stable P95 latency
Rollout	IRL5	Limited rollout with end-to-end monitoring	Auto-default with exception review	Drift dashboard; incident response logs; no regression in TTD or throughput; budget stability
	IRL6	Scaled multi-site or multi-language deployment	Supervised autonomy with periodic audits	Cross-site parity targets; stable drift control; positive ROI; no major incidents across review windows

Table 4: IRL ladder for staged autonomy. The rows turn readiness into a promotion protocol by pairing each level with a setting, an autonomy ceiling, and the evidence required before moving upward. PHI denotes protected health information, P95 the 95th-percentile latency, TTD time-to-disposition, and ROI return on investment.

though these benefits depend on grounded provenance and validation (Giuffré and Shung, 2023; Bedi et al., 2025b; Agrawal et al., 2025). Synthetic data can also be combined with expert-curated data to support downstream BioNLP and clinical NLP tasks (Mishra et al., 2024; Yao et al., 2023, 2026c).

Its most useful roles are coverage expansion, counterfactual stress testing, and workflow perturbation. Good pipelines rely on grounded evidence, versioned prompts, adjudication loops, and provenance logs, and representative approaches often combine self-play or multi-agent generation with structured generation loops to reduce annotation cost while keeping clinical constraints explicit (Wang et al., 2024d; Tu et al., 2025; Johri et al., 2025). But synthetic workflows can distort prevalence, coordination patterns, and error surfaces. Stronger autonomy claims still require real-data replay or prospective evidence (Bedi et al., 2025b; Agrawal et al., 2025).

Taken together, the evaluation stack, capability bridge, task-level instrumentation, and IRL ladder convert abstract model progress into workflow-level evidence about what a system can safely do. They also make failure modes easier to state precisely, because the dominant errors appear as failures of state tracking, escalation, provenance, or operational control rather than as generic model weakness.

5 Failure Modes and Roadmap

AI-hospital failures rarely sit in a single module. They appear as workflow-level breakdowns across roles, memory, tools, and control. Table 5 turns these failures into release blockers by showing where each one first appears, the smallest convincing test, the minimum log bundle, and the design response needed before progression.

Longitudinal drift. This occurs when the system mixes past history with current state, loses cross-visit consistency, or cannot reconstruct why a later action followed from an earlier one. Minimal checks include event-graph reconstruction, follow-up agreement, and relapse detection. The design response is workflow-aware memory with replayable longitudinal traces (Wornow et al., 2025; Liu et al., 2025a).

Capacity-blind planning. A clinically plausible plan can still fail if it ignores beds, staffing, transport delay, or queue load. These systems need stress-tested shadow replay, queue-aware simulation, and non-regression checks on time-to-disposition, throughput, and latency tails before broader rollout.

Ungated deviation and weak escalation. This failure appears when the system departs from guidelines without explicit reasons or fails to abstain on high-risk cases. Guideline versioning, reason-coded deviation logs, citation coverage, and es-

Failure mechanism	Where it first appears	Minimal release test	Minimum logs	Block stage and design response
Longitudinal drift	Follow-up, chronic care, mental-health interaction	Event-graph reconstruction; follow-up agreement; relapse sensitivity	Event IDs; timestamps; cross-handoff snapshots; replay traces; evidence links	Block IRL3+; use event-indexed memory and replayable handoffs
Capacity-blind planning	Bed management, routing, ward transfer	Overload replay; change in TTD / throughput; tail latency under load	Queue snapshots; bed-board state; transport timestamps; latency tails; human interventions	Block IRL3–IRL5; use queue-aware planning and ops non-regression gates
Ungated deviation / weak escalation	Triage, discharge, prior authorization, dosing	Adherence@step with reason audit; escalation precision/recall on high-risk cases	Guideline version; citation or evidence-ID coverage; deviation reason codes; escalation logs	Block IRL2–IRL4; calibrate uncertainty thresholds and escalation policy
Untraceable action chains	EHR-linked actions, orders, handoffs	Replay attribution; citation-to-record consistency; FHIR chaos tests	Signed role-tagged actions; tool versions; approval states; rollback states; access audits	Block IRL3–IRL4; use provenance-first connectors and signed handoffs
Drift, jailbreak, and cost shock	Any deployed workflow with external users or tools	Periodic red teaming; jailbreak pass rate; recovery time; tokens/case; P95 latency	Gateway logs; red-team seed registry; drift alerts; human touches; budget traces	Block IRL2–IRL5; use canary rollout, rollback playbook, and budget guardrails

Table 5: Workflow failure roadmap. Each row connects a recurring failure to the first workflow setting where it should block release, the smallest convincing test, the minimum log bundle, and the design response it demands.

calation precision/recall become central readiness criteria in this regime (Goodell et al., 2025; Tao et al., 2026).

Untraceable action chains. Errors spread quickly when handoffs, orders, or EHR calls lose provenance, approval state, or replayability. Even a locally correct step becomes hard to audit if later reviewers cannot reconstruct who acted, with what evidence, under which version or permission state.

Drift, jailbreak, and cost shock. A workflow can also fail after deployment because guard behavior drifts, prompts jailbreak, or token and latency costs rise until the system becomes unsafe or unusable. These are not peripheral operations issues. They mark the boundary between benchmark-strong systems and systems that can sustain staged autonomy claims (Agrawal et al., 2025; Bedi et al., 2025b, 2026).

Roadmap for robust and auditable AI hospitals. The roadmap follows directly from these mechanisms. Near-term progress depends less on adding more discussion turns and more on building workflow-aware memory, capacity-coupled planning, calibrated escalation policies, audit-ready provenance, and deployment playbooks with explicit promotion gates. The field needs tighter control of state, evidence, and operations, not only stronger generation.

A shared reporting protocol would help most. Workflow observables, minimum logs, escalation coverage, replay evidence, and operational costs should be reported together, so systems can be compared at the level where deployment claims are actually made. Over a longer horizon, AI hospitals can be seen as workflow-level approximations to medical world models, because they maintain state, act through tools and clinical interfaces, observe downstream consequences, and re-plan under partial observability. That view becomes credible only when deployment evidence is treated as a staged protocol rather than as a by-product of benchmark gains.

6 Conclusion

AI hospitals are best understood as workflow-level LLM-based multi-agent clinical systems, not as isolated question-answering agents. The main design questions are who owns each step, what state and evidence survive each handoff, and what control policy governs continuation, abstention, and escalation. From that view, evaluation should focus on workflow observables across safety, process, outcome, and operations, and deployment should proceed through explicit instrumentation, audit-ready logs, and staged IRL gates. The most reusable output of this survey is therefore not only a taxonomy of systems, but a reporting and gating framework for stateful, evidence-grounded, tool-mediated clinical workflows under partial observability.

Limitations and Ethical Considerations

This survey is constrained by space, so we summarize systems at the level of roles, memory, tools, and control rather than providing full implementation details for every method; readers may still need to consult original papers and code repositories. Our coverage prioritizes major NLP and ML venues (ACL, NeurIPS, ICLR, ICML, AAAI) and selected medical journals and recent preprints (arXiv, medRxiv, bioRxiv), so relevant work outside these channels may be missing. The taxonomy and comparisons may therefore require updates as deployment practices, regulations, and clinical integration standards continue to evolve. IRL thresholds in this survey should also be read as governance templates or protocol defaults rather than universal constants; they must be calibrated to task risk, site policy, baseline comparator, and acceptable failure budget.

While this survey introduces no direct system-level societal impact, the framework is ultimately motivated by healthcare impact and therefore inherits healthcare risks if misused. In particular, emphasizing autonomy without rigorous instrumentation and gating could encourage premature deployment, potentially increasing harm by leading to missed escalations, opaque tool failures, or inequitable performance across populations. Conversely, a workflow-first framing can support safer translation by making hidden failure modes measurable (e.g., auditability, PHI leakage, and handoff loss) and by encouraging reporting that reflects real clinical constraints (staffing, queues, and follow-up). From a social impact perspective, an important open direction is to connect the framework to equity and governance: stratified reporting across demographic and language groups, documentation of who benefits and who bears added burden (patients, caregivers, clinicians), and alignment with institutional accountability processes (incident response, audit trails, and oversight).

References

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

Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. 2022. Multimodal biomedical ai. *Nature Medicine*, 28:1773 – 1784.

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

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

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. 2026. Mirage the illusion of visual understanding. *arXiv preprint arXiv:2603.21687*.

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

Andrew M Bean, Rebecca Elizabeth Payne, Guy Parsons, Hannah Rose Kirk, Juan Ciro, Rafael Mosquera-Gómez, Sara Hincapié M, Aruna S Ekanayaka, Lionel Tarassenko, Luc Rocher, et al. 2026. Reliability of llms as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*, pages 1–7.

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

Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah. 2025a. Fidelity of medical reasoning in large language models. *JAMA Network Open*, 8(8):e2526021.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. 2025b. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 333(4):319–328.

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

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

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

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

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

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. 2025d. 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):101–109.

Zhuoyun Du, LujieZheng LujieZheng, 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. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17278–17306.

Rahma Elkamouchi, Abdelaziz Daaif, and Kamal Elguemmat. 2024. Multi-agents system in healthcare: A systematic literature review. In *International Conference on Smart Applications and Data Analysis*, pages 200–214. Springer.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray. In *International Conference on Machine Learning*, pages 15661–15676. PMLR.

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

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9878–9919.

Mauro Giuffré and Dennis L. Shung. 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digital Medicine*, 6.

Ethan Goh, Robert J Gallo, Eric Strong, Yingjie Weng, Hannah Kerman, Jason A Freed, Joséphine A Cool, Zahir Kanjee, Kathleen P Lane, Andrew S Parsons, et al. 2025. Gpt-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nature Medicine*, 31(4):1233–1238.

Alex J Goodell, Simon N Chu, Dara Rouholiman, and Larry F Chu. 2025. Large language model agents can use tools to perform clinical calculations. *npj Digital Medicine*, 8(1):163.

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

Won Seok Jang, Hieu Tran, Manav Mistry, SaiKiran Gandluri, Yifan Zhang, Sharmin Sultana, Sunjae Kwon, Yuan Zhang, Zonghai Yao, and Hong Yu. 2025. Chatbot to help patients understand their health. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, pages 6598–6627.

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.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2024. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075.

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, et al. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2023. Guidelines for rigorous evaluation of clinical llms for conversational reasoning. *medRxiv*, pages 2023–09.

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.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024.

Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Sunjae Kwon, Zonghai Yao, Harmon S Jordan, David A Levy, Brian Corner, and Hong Yu. 2022. Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 11733.

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

Moustafa Laymouna, Yuanchao Ma, David Lessard, Tibor Schuster, Kim Engler, and Bertrand Lebouché. 2024. Roles, users, benefits, and limitations of chatbots in health care: rapid review. *Journal of medical Internet research*, 26:e56930.

Tyler Alise Le, Arpi Jivalagian, Tasneem Hiba, Joshua Franz, Shahab Ahmadzadeh, Treniece Eubanks, Leisa Oglesby, Sahar Shekoohi, Elyse M Cornett, and Alan D Kaye. 2023. Multi-agent systems and cancer pain management. *Current Pain and Headache Reports*, 27(9):379–386.

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. 2024a. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760.

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

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. 2024c. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. 2024d. Leveraging large language model as simulated patients for clinical education. *arXiv preprint arXiv:2404.13066*.

Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. Automatic interactive evaluation for large language models with state aware patient simulator. *arXiv preprint arXiv:2403.08495*.

Jung Hoon Lim, Sunjae Kwon, Zonghai Yao, John P Lalor, and Hong Yu. 2024. Large language model-based role-playing for personalized medical jargon extraction. *arXiv preprint arXiv:2408.05555*.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, SU Yihang, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael Lyu, and Wenting Chen. 2025a. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Qicai Liu, Zhichao Hu, Tao Huang, Yupeng Niu, Xinche Zhang, Shanwu Ma, Chutong Lin, Goh Kim Huat, Hyeokkoo Eric Kwon, Feng Gao, et al. 2026. Evomdt: a self-evolving multi-agent system for structured clinical decision-making in multi-cancer. *npj Digital Medicine*.

Sizhe Liu, Yizhou Lu, Siyu Chen, Xiyang Hu, Jieyu Zhao, Tianfan Fu, and Yue Zhao. 2024. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. In *2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle*.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. 2025b. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 31(3):932–942.

Zhaocheng Liu, Quan Tu, Wen Ye, Yu Xiao, Zhishou Zhang, Hengfu Cui, Yalun Zhu, Qiang Ju, Shizheng Li, and Jian Xie. 2025c. Exploring the inquiry-diagnosis relationship with advanced patient simulators. *arXiv preprint arXiv:2501.09484*.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764.

Sravanthi Machcha, Sushrita Yerra, Sahil Gupta, Aishwarya Sahoo, Sharmin Sultana, Hong Yu, and Zonghai Yao. 2026. Knowing when to abstain: Medical LLMs under clinical uncertainty. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6153–6182, Rabat, Morocco. Association for Computational Linguistics.

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

Prakamya Mishra, Zonghai Yao, Parth Vashisht, Feiyun Ouyang, Beining Wang, Vidhi Dhaval Mody, and Hong Yu. 2024. Synfac-edit: Synthetic imitation edit feedback for factual alignment in clinical summarization. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20061–20083.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.

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

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*.

Hanwen Shi, Jin Zhang, and Kunpeng Zhang. 2024a. Enhancing clinical trial patient matching through knowledge augmentation with multi-agents. *arXiv preprint arXiv:2411.14637*.

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

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. In *International Conference on Machine Learning*, pages 45883–45905. PMLR.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2025. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723.

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. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Xinge Tao, Shuya Zhou, Kai Ding, Sairan Li, Yanzeng Li, Boyou Wu, Qirui Huang, Wangyue Chen, Muzi Shen, En Meng, et al. 2026. An llm chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. *Nature Medicine*, pages 1–9.

Muhammad Usman Tariq. 2024. Multi-agent models in healthcare system design. In *Bioethics of Cognitive Ergonomics and Digital Transition*, pages 143–170. IGI Global.

Hieu Tran, Zonghai Yao, Nguyen Luong Tran, Zhichao Yang, Feiyun Ouyang, Shuo Han, Razieh Rahimi, and Hong Yu. 2026. Prime: Planning and retrieval-integrated memory for enhanced reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33268–33276.

Hieu Tran, Zonghai Yao, Zhichao Yang, Junda Wang, Yifan Zhang, Shuo Han, Feiyun Ouyang, and Hong Yu. 2025. Rare: Retrieval-augmented reasoning enhancement for large language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18305–18330.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450.

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

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024b. Towards a client-centered assessment of llm therapists by client simulation. *arXiv preprint arXiv:2406.12266*.

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

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024c. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Junda Wang, Zonghai Yao, Lingxi Li, Junhui Qian, Zhichao Yang, and Hong Yu. 2025a. Chatthero: An llm-supported chatbot for behavior change and therapeutic support in addiction recovery. *arXiv preprint arXiv:2508.20996*.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024d. Notechat: a dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201.

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev

Jones, Kate Hardy, Hong Shen, et al. 2024e. Patient-psi: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797.

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

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

Zifeng Wang, Benjamin Danek, Ziwei Yang, Zheng Chen, and Jimeng Sun. 2024g. Can large language models replace data scientists in biomedical research? *arXiv preprint arXiv:2410.21591*.

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

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025d. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *The Thirteenth International Conference on Learning Representations*.

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

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Re, Sanmi Koyejo, and Nigam Shah. 2025. Context clues: Evaluating long context models for clinical prediction tasks on ehr data. In *The Thirteenth International Conference on Learning Representations*.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467.

Yuqi Wu, Guangya Wan, Jingjing Li, Shengming Zhao, Lingfeng Ma, Tianyi Ye, Mike Zhang, Ion Pop, Yanbo Zhang, and Jie Chen. 2026. Wisemind: a knowledge-guided multi-agent framework for accurate and empathetic psychiatric diagnosis. *npj Digital Medicine*.

Yihang Xiao, Jinyi Liu, YAN ZHENG, Shaoqing Jiao, Jianye HAO, Xiaohan Xie, Ruitao Wang, Fei Ni, Yuxiao Li, Zhen Wang, et al. 2025. Cellagent: Llm-driven multi-agent framework for natural language-based single-cell analysis. In *The Fourteenth International Conference on Learning Representations*.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, et al. 2025. Rag-gym: Optimizing reasoning and search agents with process supervision. *arXiv preprint arXiv:2502.13957*.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, and Jiayi Wang. 2025a. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, and Jiayi Wang. 2025b. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29.

Dingkang Yang, Jinjie Wei, Mingcheng Li, Jiyao Liu, Lihao Liu, Ming Hu, Junjun He, Yakun Ju, Wei Zhou, Yang Liu, et al. 2025. Medaide: information fusion and anatomy of medical intents via llm-based agent collaboration. *Information Fusion*, page 103743.

Zonghai Yao, Talha Chafekar, Junda Wang, Shuo Han, Feiyun Ouyang, Junhui Qian, Lingxi Li, and Hong Yu. 2026a. Chatclids: Simulating persuasive ai dialogues to promote closed-loop insulin adoption in type 1 diabetes care. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39539–39547.

Zonghai Yao, Nandyala Siddharth Kantu, Guanghao Wei, Hieu Tran, Zhangqi Duan, Sunjae Kwon, Zhichao Yang, and Hong Yu. 2024. Readme: Bridging medical jargon and lay understanding for patient education through data-centric nlp. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12609–12629.

Zonghai Yao, Benjamin Schloss, and Sai Selvaraj. 2023. Improving summarization with human edits. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2604–2620.

Zonghai Yao, Michael Sun, Won Seok Jang, Sunjae Kwon, Soie Kwon, and Hong Yu. 2025. Dischargesim: A simulation benchmark for educational doctor–patient communication at discharge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10783–10809.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and Hong Yu. 2026b. MedQA-CS: Objective structured clinical examination (OSCE)-style benchmark for evaluating LLM clinical skills. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6183–6257, Rabat, Morocco. Association for Computational Linguistics.

Zonghai Yao, Youxia Zhao, Avijit Mitra, David A Levy, Emily Druhl, Jack Tsai, and Hong Yu. 2026c. Synthehr-eviction: enhancing eviction sdoh detection with llm-augmented synthetic ehr data. *npj Digital Medicine*.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Jie Sun, Xiang Li, Jingxian He, Wenyue Hua, et al. 2025. Simulated patient systems powered by large language model-based ai agents offer potential for transforming medical education. *Communications Medicine*.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2025. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25796–25804.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, et al. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, pages 1–10.

Yutian Zhao, Huimin Wang, Yefeng Zheng, and Xian Wu. 2025. A layered debating multi-agent system for similar disease diagnosis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 539–549.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, et al. 2025. Automating expert-level medical reasoning evaluation of large language models. *npj Digital Medicine*.

Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Zixiang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2025. 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*.

Checklist question	Signal for single-agent choice	Signal for multi-agent choice	Why it matters
How wide is the workflow span?	One bounded step or one care stage	Multiple stages with explicit handoffs	Workflow width determines whether state must be transferred or can stay local
How different are the roles?	One role with one policy regime	Different clinical or operational roles with different constraints	Role heterogeneity determines whether responsibility can stay implicit
How much audit structure is needed?	Final-output review is usually enough	Per-role accountability and replayable handoffs are required	Audit burden is where extra roles start paying for themselves
How explicit must escalation be?	Manual review is enough in practice	Explicit triggers and coverage checks are needed	Escalation needs often force a controller or judge role
How tightly is the task coupled to operations?	Weak queue or capacity coupling	Queue, transport, or approval state changes the plan	Operational coupling often requires separate planner or routing roles
What cost overhead is acceptable?	Lower tokens and lower coordination latency are preferred	Extra turns are acceptable in exchange for clearer ownership and repair	Many multi-agent gains are only worth it when the added cost is tolerable

Table 6: Single-agent versus multi-agent as a workflow checklist. The rows identify the decision questions that determine when extra roles add measurable workflow value rather than only adding more reasoning turns.

Memory design	What survives the handoff	Best fit	Main risk	Minimum persisted artifacts
Sliding window	Recent turns and current task state	Short, single-visit workflows	Context drop; contraindication loss	Recent turns; current task state
Retrieval-linked memory	Query-linked record or guideline snippets with citations	Evidence lookup at decision time	Bad retrieval; uncited claim; stale reference	Query; evidence IDs; citations
Event buffer	Timestamped state updates and milestone events	Multi-step visit or shift handoff	Missing event; wrong temporal order	Structured events; timestamps; state snapshots
Temporal knowledge graph	Entities, relations, timeline links, and provenance	Longitudinal care or rare-disease reasoning	Patient mix-up; entity drift; privacy over-link	Entities; timeline links; provenance

Table 7: Memory designs in workflow terms. The comparison centers on what a later role can actually recover after a handoff, where each design fits best, and which failure it most often introduces.

Tool family	Workflow function	Best fit	What must be audited	Typical failure signal
Simulated EHR	Safe workflow rehearsal with synthetic or controlled records	Sandbox evaluation and early stress testing	Scenario seed; state transitions; tool-call traces	Good-looking workflow with weak realism
Live EHR (FHIR-backed)	Record-grounded decision support and handoff recovery	Shadow replay or monitored deployment	Queries; retrieved fields; consent or audit IDs; access events	Permission drift; schema drift; PHI leakage
Guideline or calculator tools	Explicit checks for triage, dose, or contraindications	Regulated steps with clear rules	Inputs; tool version; guideline version; exceptions	Wrong input; stale guideline; missed check
Action connectors (orders, booking, prior auth)	Execute or route high-stakes downstream actions	Approval-bound workflows	Request payloads; approval state; overrides; rollback events	Silent fallback; missing approval trace
External clinical APIs	Pull current facts beyond local stores	Coverage gaps in local evidence	Query; returned IDs; source version or date; failures	Source drift; rate-limit failure; uncited fact
Multimodal connectors	Tie text decisions to image or sensor evidence	Imaging, monitoring, or mixed-modality tasks	Input modality IDs; evidence links; preprocessing versions	Claim not anchored to observed modality
Research automation tools	Run code, analysis, or modeling for biomedical workflows	Study design and scientific support	Data provenance; code or tool version; parameters; runtime logs	Non-reproducible result or hidden execution error

Table 8: Tool stack as workflow connectors. The rows group tools by what they do inside the workflow, what must be audited when they are used, and which failure signal usually appears first.

Choice	Best fit	Main cost or failure	Minimum logs
Reasoning choices			
Single-path CoT	Short, well-specified tasks	Omission errors compound; weak branch coverage	Tokens or latency; citation rate; tool-call trace; error tags
Multi-path debate or self-consistency	Open differential diagnosis or persistent uncertainty	High controller cost; hard-to-reconcile disagreement	Number of branches; vote margin; disagreement rate; escalation triggers
Memory choices			
Static or retrieval-backed memory	Stable guidelines; bounded history recovery	Staleness; weak personalization; retrieval miss	Last-updated time; source provenance; evidence IDs; citation coverage
Temporal or structured longitudinal memory	Multi-visit care and rare-disease tracking	Higher privacy and maintenance burden; wrong merge or split across time	Node or event updates; timestamps; provenance links; conflict flags
Control choices			
Tool-first control	Regulated steps with explicit checks	Rigid behavior when tools fail or edge cases appear	Plan steps; tool schema or version; retries; constraint checks
LLM-first control	Open-ended, personalized interaction	Higher verification burden; unstable provenance	Safety flags; refusal rate; tool provenance; post-hoc check status

Table 9: Design trade-offs in operational form. The rows pair common reasoning, memory, and control choices with the setting where they fit, the practical cost they introduce, and the minimum logs needed to keep the workflow inspectable.

A Expanded Design Taxonomy

This appendix keeps the same design questions as the main text and adds finer-grained categories, representative systems, and local trade-offs for roles, interaction patterns, tools, memory, and reasoning.

A.1 Agent Roles

A.1.1 Patient-Centered Agents

Patient-Centered Agents simulate patient state, communication, and help-seeking behavior across training, intake, education, and follow-up.

Patient Agent models symptom reporting, history disclosure, and patient-side communication in consultation, intake, and education workflows (Bao et al., 2024; Wang et al., 2024d). The main design challenge is role stability: the agent should reveal information, use lay terminology or support jargon translation, and remain behaviorally consistent across turns rather than drifting into generic assistant behavior (Du et al., 2025; Li et al., 2024d; Yu et al., 2025; Liu et al., 2025c; Kwon et al., 2022; Lim et al., 2024).

Psychological Patient Agent simulates counseling and behavior-change interactions where mood, resistance, and crisis sensitivity affect what the next therapeutic step should be (Wang et al., 2024b; Wei et al., 2024; Wang et al., 2025a; Yao et al., 2026a). The main difficulty is not only realism but stable behavioral response under long interaction, including cognitive distortions, treatment resistance, and adaptive response to counseling prompts (Louie et al., 2024; Wang et al., 2024e; Chen et al., 2023).

Resident Agents simulate general populations before and during care seeking, so they are useful for public-health scenarios, intake demand, and transitions from healthy residents to active patients (Li et al., 2024b; Williams et al., 2023). Their value comes from modeling help-seeking behavior, disease progression, and policy response at population scale rather than from detailed bedside interaction.

A.1.2 Medical Professional Agents

Medical Professional Agents own clinical or support steps that update the case state, from first assessment to testing, coordination, counseling, and training.

General Doctor Agent performs first-pass assessment, follow-up questioning, and coordination

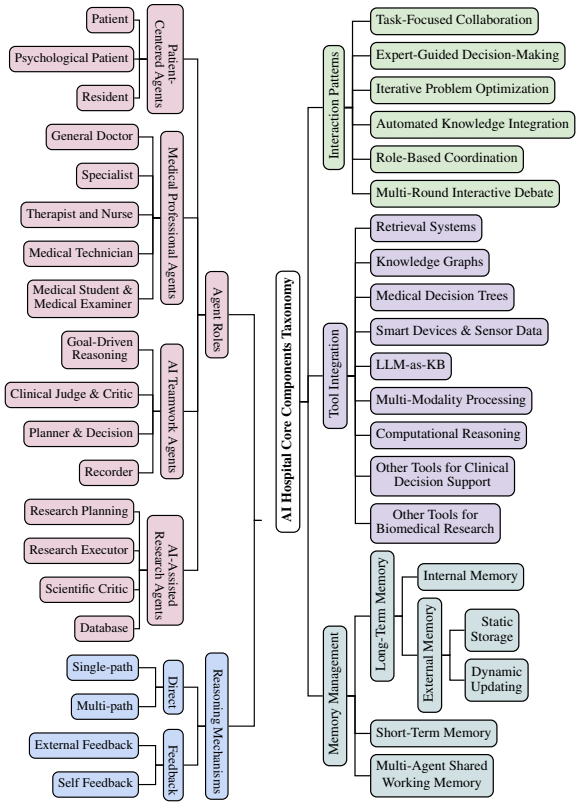

Figure 1: Expanded taxonomy of AI-hospital core components. The figure unpacks the finer-grained role, interaction, tool, memory, and reasoning categories that are compressed into the main-text design space.

of the diagnostic path, so it often acts as the local owner of intake or consultation workflows (Liu et al., 2025c; Du et al., 2025; Johri et al., 2023; Kim et al., 2024; Wang et al., 2024a; Fan et al., 2025). Its main requirement is to keep uncertainty explicit while deciding whether to continue, request more evidence, or pass the case to a specialist.

Specialist Agent contributes domain-specific constraints for complex cases, especially when rare conditions, imaging, or discipline-specific evidence make a single general path brittle (Chen et al., 2026; Kim et al., 2024). The practical gain is not merely more expertise, but more explicit disagreement and reconciliation across specialties before a plan is committed.

Therapist Agent provides counseling, emotional support, and psychotherapy under structured behavioral principles (Wang et al., 2024b; Qiu and Lan, 2024; Chen et al., 2023).

Nurse Agent supports triage, routine care coordination, and patient-side communication across intake and follow-up workflows (Bao et al., 2024;

Li et al., 2024b).

Medical Technician Agents carry diagnostic or testing steps that supply reliable measurements to later clinical decisions (Schmidgall et al., 2024).

Medical Student and Examiner Agents support training workflows by simulating history taking, communication, and diagnostic assessment under structured evaluation (Li et al., 2024d; Yao et al., 2026b).

A.1.3 Medical AI Teamwork Agents

Medical AI Teamwork Agents exist when one long reasoning trace is not enough and the workflow needs explicit routing, critique, reconciliation, or record keeping. Systems such as MedAIDE show the same idea at a broad assistant level, but the key point is workflow separation rather than role count alone (Yang et al., 2025).

Goal-Driven Reasoning Agent organizes multi-step reasoning so that the case progresses through explicit subgoals, tool calls, or symbolic steps instead of one opaque answer (Yu et al., 2025; Hong et al., 2024; Shi et al., 2024b).

Clinical Judge Agent checks whether a diagnosis or plan is ready to pass a gate, typically by testing factual consistency, guideline adherence, or decision readiness (Johri et al., 2023; Yue et al., 2024).

Critic Agent challenges local reasoning and surfaces errors, bias, or missing evidence before commitment (Ke et al., 2024; Hong et al., 2024).

Planning Agent decomposes the workflow into executable steps and decides what should happen next under time, evidence, or routing constraints (Yue et al., 2024; Shi et al., 2024a).

Decision Agent reconciles conflicting assessments and converts them into one bounded commitment or escalation choice (Tang et al., 2024; Wang et al., 2025c).

Recording Agent writes the structured handoff state, so later roles and auditors can reconstruct what was decided, why, and under which evidence (Ke et al., 2024; Yu et al., 2025).

A.1.4 AI-Assisted Research Agents

AI-assisted research agents extend the same workflow logic to scientific planning, execution, critique, and information management.

Research Planning Agent structures research goals into tractable subproblems, choosing sequences of analysis or experiment that make complex scientific workflows executable (Swanson et al., 2025; Xiao et al., 2025).

Scientific Critic Agent checks whether generated hypotheses, analyses, or claims are methodologically sound enough to survive review (Xiao et al., 2025).

Scientific Critic Agent checks whether generated hypotheses, analyses, or claims are methodologically sound enough to survive review (Xiao et al., 2025).

Database Agent retrieves and organizes external biomedical information so later research agents can ground their decisions in accessible evidence stores (Shi et al., 2024b).

A.2 Interaction Patterns

This subsection describes how role separation and control policies become visible at the interaction level.

Task-Focused Collaboration splits a case into ordered subtasks, which fits workflows with clear stage boundaries or specialized micro-decisions (Li et al., 2024d; Yu et al., 2025; Yue et al., 2024; Shi et al., 2024b). Its value comes from making state transitions explicit rather than leaving the whole workflow inside one long prompt.

Expert-Guided Decision-Making inserts specialty constraints or expert review into the decision path when clinical validity cannot be trusted to one generic reasoning trace (Du et al., 2025; Chen et al., 2026; Kim et al., 2024; Tang et al., 2024). It is most useful when the workflow needs consensus, guideline-sensitive review, or domain-specific correction.

Iterative Problem Optimization (IPO) revises the current plan after new feedback, tool output, or interaction evidence arrives (Yu et al., 2025; Du et al., 2025; Bao et al., 2024; Tang et al., 2024; Shi et al., 2024b). It fits workflows where local repair is cheaper and safer than restarting the entire case.

Automated Knowledge Integration (AKI) merges patient state with retrieved knowledge, structured memory, or multimodal evidence so later steps remain context-aware (Shi et al., 2024a; Liao et al., 2024; Du et al., 2025; Yang et al., 2024;

Wang et al., 2024a; Lu et al., 2024; Hong et al., 2024). Its main value is to keep external evidence attached to the evolving case rather than appended as an afterthought.

Role-based Coordination assigns different agents explicit ownership over intake, diagnosis, counseling, or follow-up steps (Du et al., 2025; Wang et al., 2024b; Qiu and Lan, 2024; Wang et al., 2025c; Chen et al., 2026; Schmidgall et al., 2024; Li et al., 2024b). It becomes useful when the workflow crosses responsibility boundaries and each handoff must preserve state and accountability.

Multi-Round Interactive Debate delays commitment so that disagreement can be surfaced, compared, and resolved before the workflow moves on (Fan et al., 2025; Chen et al., 2025d; Kim et al., 2024; Tang et al., 2024; Smit et al., 2024; Lu et al., 2024; Swanson et al., 2025). It is helpful under persistent uncertainty, but it raises controller cost and only pays off when the final reconciliation rule is explicit.

A.3 Tool Integration

This subsection groups tools by workflow function, where they fit best, what must be audited, and which failure signal they expose first.

Retrieval systems pull record- or guideline-linked evidence into the current step, so consult and discharge outputs can cite what they rely on (Du et al., 2025; Kim et al., 2024). Their first failure is usually retrieval miss, stale evidence, or uncited use of retrieved content.

Knowledge graphs preserve entities, relations, and temporal links that are hard to keep stable in free text alone (Li et al., 2024d; Yu et al., 2025; Chen et al., 2026). They fit longitudinal or relation-heavy workflows, but can fail through entity drift or incorrect linkage.

Medical decision trees turn regulated checks into explicit pathways for triage, contraindication review, or structured diagnosis (Yang et al., 2024; Li et al., 2023). Their main value is predictable control, and their main risk is brittleness when the input state is incomplete or off-distribution.

LLM-as-KB uses model-internal knowledge as a flexible synthesis layer when no single external database is enough (Yue et al., 2024; Frisoni et al., 2024). It is useful for broad reasoning but should not replace provenance-bearing evidence stores.

Smart device and sensor data bring streaming physiologic or behavioral signals into the workflow, which is valuable for monitoring and personalized follow-up (Yang et al., 2024; Abbasian et al., 2025). The first failure is usually weak alignment between the observed signal and the text decision it is supposed to support.

Multi-modality processing tools connect text reasoning to images, reports, and sensor streams so decisions remain anchored to observed modality-specific evidence (Li et al., 2024d; Yang et al., 2024; Li et al., 2024a). They matter when visual or multimodal evidence cannot be safely collapsed into text alone.

Computational reasoning tools add executable analysis, code, or formal inference when the workflow depends on calculations or reproducible research steps (Wang et al., 2024g; Hong et al., 2024). Their advantage is precise execution, but only when code version, parameters, and outputs are logged.

Other clinical decision support tools connect the workflow to external APIs, predictive models, or structured reporting services that fill gaps left by local tools (Wang et al., 2024a; Li et al., 2024a). They are best treated as specialized connectors whose outputs still need provenance and compatibility checks.

Other biomedical research tools support tasks such as drug discovery, genomics, and molecular analysis when the workflow extends into scientific experimentation (Swanson et al., 2025; Jin et al., 2024; Liu et al., 2024). Their main requirement is reproducible execution rather than fluent explanation alone.

A.4 Memory Management

This subsection focuses on the state that must survive a handoff, where different memory designs fit, and what failure each design tends to introduce.

A.4.1 Long-Term Memory (LTM)

Long-term memory (LTM) keeps knowledge that should remain available across visits, sessions, or repeated workflows.

External Memory keeps stable references outside the model, including records, guidelines, knowledge graphs, trial registries, prior trajectories, and domain repositories (Wang et al., 2024a,e; Lu et al., 2024; Yang et al., 2024; Shi et al., 2024a;

Yue et al., 2025; Chen et al., 2026; Yue et al., 2024; Liu et al., 2024; Wang et al., 2026; Lai et al., 2025). It fits workflows that need auditable retrieval and step-specific grounding, but it can fail through staleness, retrieval miss, or weak linkage between stored evidence and the active case state.

External Memory keeps stable references outside the model, including records, guidelines, knowledge graphs, trial registries, prior trajectories, and domain repositories (Wang et al., 2024a,e; Lu et al., 2024; Yang et al., 2024; Shi et al., 2024a; Yue et al., 2025; Chen et al., 2026; Yue et al., 2024; Liu et al., 2024; Wang et al., 2026; Lai et al., 2025). It fits workflows that need auditable retrieval and step-specific grounding, but it can fail through staleness, retrieval miss, or weak linkage between stored evidence and the active case state.

Dynamic Updating refreshes the current case with newly retrieved evidence, updated guidelines, expert feedback, or prior cases that become relevant at the present step (Louie et al., 2024; Yang et al., 2024; Wang et al., 2024g; Shi et al., 2024b; Schmidgall et al., 2024; Bao et al., 2024; Wang et al., 2026; Lai et al., 2025). It matters when the workflow would otherwise keep reasoning on stale state.

A.4.2 Short-Term Memory (STM) and Multi-Agent Shared Working Memory (WM)

STM and WM support different parts of the workflow. Short-term memory (STM) keeps the current interaction coherent inside one local step, for example by retaining dialogue history, extracted entities, or a temporary summary (Liu et al., 2025c). Working memory (WM) is shared across roles and stores the active case state, intermediate decisions, feedback, and execution traces that later agents need in order to resume the workflow without reconstruction (Lu et al., 2024; Hong et al., 2024; Kim et al., 2024; Xiao et al., 2025; Tang et al., 2024; Wang et al., 2025c; Swanson et al., 2025; Wang et al., 2026; Lai et al., 2025).

A.5 Reasoning Mechanisms

A.5.1 Direct Reasoning

This subsection expands the main-text discussion by asking when a reasoning pattern fits, what it costs in practice, and what must be logged if the workflow depends on it.

Single-path Reasoning follows one explicit chain from evidence to answer, which fits short and well-specified tasks where the main risk is omission rather than unresolved disagreement (Li et al., 2024d; Wang et al., 2024e,b; Kim et al., 2024; Yan et al., 2025b; Schmidgall et al., 2024; Bao et al., 2024; Wang et al., 2024g; Yue et al., 2024; Chen et al., 2025b). Its main weakness is that local errors propagate forward unless a later gate interrupts the chain.

Multi-path Reasoning keeps several candidate paths alive through parallel specialists, debate, self-consistency, or planner-generated alternatives (Du et al., 2025; Chen et al., 2026; Kim et al., 2024; Tang et al., 2024; Wang et al., 2025c; Li et al., 2024c; Wang et al., 2024a; Hong et al., 2024; Liu et al., 2024; Swanson et al., 2025). It is useful under persistent uncertainty, but it increases controller cost and requires explicit reconciliation or escalation rules.

A.5.2 Feedback-Based Reasoning

Feedback-based reasoning updates the current plan rather than treating the first answer as final.

External Feedback Reasoning revises the workflow after patient interaction, expert correction, or tool output adds new evidence (Chen et al., 2025b; Johri et al., 2023; Li et al., 2024c). It fits settings where the case state changes during execution and the system must adapt rather than restate the old plan.

Self Feedback Reasoning revises reasoning internally when external supervision is delayed or unavailable, for example through reflection, self-play, or structured error checking (Louie et al., 2024; Yu et al., 2025; Schmidgall et al., 2024; Wang et al., 2024g). It can reduce local inconsistency, but it still needs a stopping rule and later verification before high-stakes commitment.

B Expanded Workflow Settings and Task Examples

This appendix section extends the workflow families from the main text with representative systems and sharper task-specific distinctions.

B.1 Clinical Workflow Simulation

Clinical workflow simulation covers consultation, diagnosis, discharge, and follow-up under explicit

stage structure. Some systems focus on consultation itself, using patient, doctor, and evaluator agents to expose where questioning, evidence use, or abstention breaks down across stages (Liu et al., 2025c; Johri et al., 2023; Li et al., 2024c; Fan et al., 2025; Schmidgall et al., 2024). Others extend the workflow to intake, reception, or the full patient journey, which makes routing, handoff completeness, and state persistence central rather than secondary (Bao et al., 2024; Li et al., 2024b). A closely related patient-facing line focuses on discharge and post-visit support, where the key outcomes are explanation quality, comprehension, and follow-up readiness rather than diagnosis alone (Cai et al., 2023; Yao et al., 2024, 2025; Jang et al., 2025).

Given the communication-centric nature of mental health care, this line can be read as a longitudinal variant of the same workflow logic. These systems care most about stable patient state, behavior change, and timely escalation over repeated turns, often using expert-authored behavioral rules or structured counseling principles to reduce unsafe drift (Louie et al., 2024; Wang et al., 2024e; Chen et al., 2023; Wang et al., 2024b; Wei et al., 2024).

B.2 Multi-Disciplinary Medical Team Simulation

This family studies cases where explicit specialty disagreement is part of the task. Systems such as RareAgents, MDAgents, MEDAGENTS, and CoLaCare use coordinated specialists, moderators, or meta-agents to compare views, integrate evidence, and convert disagreement into a final recommendation rather than hiding it inside one trace (Chen et al., 2026; Kim et al., 2024; Tang et al., 2024; Wang et al., 2025c).

B.3 Simulated Patients for Medical Education

This family uses patient and clinician roles to train communication, history taking, and diagnostic reasoning under controlled evaluation. The main design goal is not bedside autonomy but fidelity, structured feedback, and repeatable assessment of learner performance (Du et al., 2025; Wei et al., 2024; Mehandru et al., 2024; Yao et al., 2026b).

B.4 Other Medical Process Optimization and Cross-Disciplinary Simulation

This family extends the same multi-agent logic to process optimization, biomedical research, and public-health simulation. The shared pattern is explicit role separation, tool-mediated work, and structured traces over complex workflows rather than direct bedside interaction (Swanson et al., 2025; Williams et al., 2023).