A Survey on LLM-based Multi-Agent AI Hospital

Anonymous ACL submission

Abstract

AI hospitals are workflow-level multi-agent systems built on large language models that run inside clinical processes. Agents take explicit roles, maintain shared state through handoffs, use EHR- and guideline-grounded tools, and operate under safety gateways with audit logs. Prior work is rich but fragmented across tasks and settings. This survey defines the scope and boundaries of AI hospitals and compiles designs into a compact taxonomy with head-to-head trade-off matrices. We introduce a layered evaluation stack that measures safety, clinical processes, outcomes, and operations (e.g., time-to-disposition, throughput, and token/latency costs), and we use Integration Readiness Levels (IRL1–IRL6) to gate autonomy from sandbox to deployment, with required logs and pass criteria. To make deployment claims testable, we map key integration tasks to minimal instrumentation and formulate several challenges as workflow-failure mechanisms with concrete tests and IRL gates. We close with a practical roadmap on workflow-aware memory, queue-aware planning, escalation learning, traceability, and playbook adoption.

1 Introduction

Large Language Models (LLMs) are moving from text generation to agentic work in complex settings. In this survey, we use "AI hospital" to refer to a workflow-level multi-agent system constrained by safety rules, handoffs, and operational limits. This setting captures this shift especially well. These systems span decision support, patient education, and mental health support. Across settings, the key difficulty is coordination under safety, time, and resource limits. AI hospitals are organised around three interlocking modules: Roles & Interaction, Memory & Tools, and Reasoning & Control. This decomposition is not de novo: it mirrors how hospitals separate team roles and handoffs, clinical information systems (records and tools), and safety-critical control (gates, escalation, and audit).

What is an AI Hospital? An AI hospital is not a chatbot. Our definition follows how real hospitals run care: roles, handoffs, tools, and audit are the minimum units for verifiable safe workflows. We define an AI hospital as a workflow-level multi-agent simulation or deployment for clinical care, with the following elements:

- **Defined roles:** clinicians, patients, staff, operations managers, and researchers with clear duties.
- **Workflow state and handoffs:** shared context across triage, consultation, and discharge, with structured transfers.
- **Tools and data:** access to EHRs, medical knowledge bases, or realistic surrogates to ground actions.
- **Safety gates and audit logs:** guard models and logging for ethics and traceability.
- **Longitudinal memory:** persistent patient history and agent traces to support follow-up care.

We treat these as minimum requirements for hospital-like workflows, rather than a full checklist of all clinical functions.

Scope and Boundaries We distinguish AI hospital systems from related work. **In scope** are multi-agent simulators for triage, consultation, discharge, bed management, and care coordination, and also training wards that link to EHRs or realistic data. These systems model workflows across multiple roles and time points, allowing for the evaluation of coordination, reasoning, and safety under realistic constraints. This scope lets us evaluate not only clinical correctness, but also handoff quality, escalation behavior, and operational cost in a unified way. **Out of scope** are single-agent chatbots, generic role play without workflow state, single-turn QA, and tools without longitudinal memory or safety rules. For example, a strong single-model triage assistant can be useful, but it is out of scope

if it has no workflow state, no handoffs, and no audit-ready logs.

Contributions. 1. We give a workflow-level definition and boundary for AI hospitals. 2. We compress system designs into pattern taxonomies and head-to-head matrices. 3. We introduce an evaluation stack that covers safety, process, outcomes, and operations. 4. We propose IRL levels as workflow gates for staged deployment. 5. We formulate several challenges as testable hypotheses with failure modes and measurements, then derive a roadmap. **How this differs from prior surveys.** We align with and compare to recent reviews in healthcare multi-agent systems and related areas (Elkamouchi et al., 2024; Laymouna et al., 2024; Le et al., 2023; Wang et al., 2025). Our focus differs in four ways. 1) We focus on workflow-level systems with explicit state and handoffs. 2) We provide head-to-head matrices for key design choices. 3) We include operations metrics and IRL gates for integration. 4) We express challenges as tests with measurable signals **Overall**, this survey offers a workflow-centric view of LLM systems in healthcare. Success is measured by decision quality, safe handoffs, harm avoided, and efficiency under limits. We analyze core components, applications, and challenges, and we give a roadmap for integration that respects clinical realism and ethics.

2 Core Components

2.1 Roles & Interaction

Agent roles and interaction modes. Roles include doctors, nurses, patients, receptionists, researchers, and managers. Interaction spans task-focused collaboration (e.g., doctor–patient Q&A), expert-guided decision making (committee discussions), iterative optimisation (agents propose and refine), and multi-round debate with confidence-based voting. New systems simulate end-to-end flows with reception, triage, consultation, and discharge. **1. Patient-facing roles.** Patient Agents support consultation training and education; CoT and RAG stabilise persona and reduce hallucination (Yu et al., 2024). Psychological Patient Agents model mood shifts and treatment resistance via expert-guided prompts and cognitive schemata (Louie et al., 2024; Wang et al., 2024c). Resident/Population Agents navigate care pathways and support public-health simulation (Li et al., 2024c). **2. Clinical professional roles.** General Doctor Agents lead staged history taking and initial

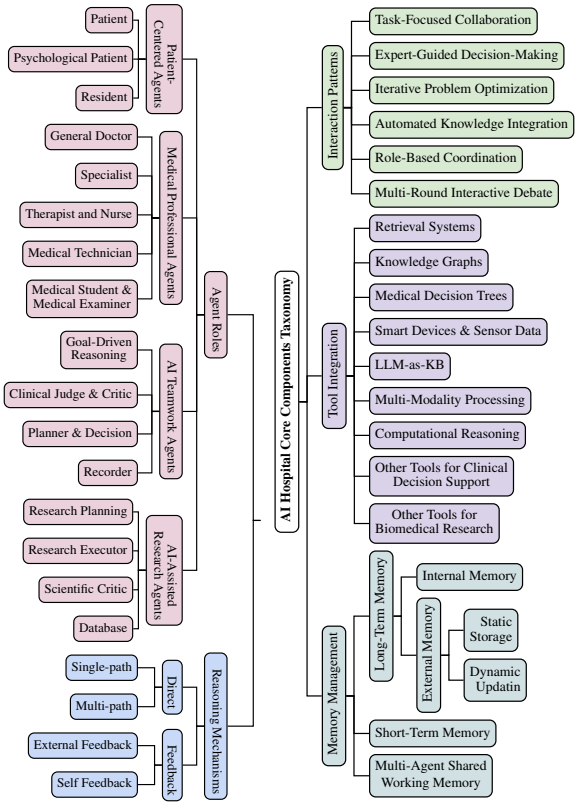

Figure 1: Taxonomy of AI hospital core components.

reasoning (Johri et al., 2023; Liu et al., 2025). Specialist Agents add domain depth and join MDTs for complex cases (Kim et al., 2024; Chen et al., 2024b). Therapist/Nurse/Technician/Student & Examiner cover counselling, triage, testing, and training (Bao et al., 2024; Schmidgall et al., 2024). **3. Teamwork and controllers.** Goal-driven reasoning / planning organises multi-step pipelines and conversations (Yu et al., 2024; Yue et al., 2024a). Clinical judge & critic constrain outputs to guidelines and provide structured feedback (Johri et al., 2023; Hong et al., 2024). Decision & recording reconcile conflicting assessments and log traceable evidence (Tang et al., 2023; Ke et al., 2024). Systems such as AgentClinic and Agent Hospital extend roles to reception, triage, and follow-ups (Schmidgall et al., 2024; Li et al., 2024c).

This module follows how hospitals organise care: outcomes depend on role division and handoffs, not only on a single clinician's reasoning. We therefore treat roles, interaction modes, and handoff points as first-class design units. Table 1 summarises when role separation is worth the overhead. Additional details are provided in Appendix A.1 and A.2.

2.2 Memory & Tools

Effective memory maintains task context, supports updates from trusted sources, and limits unneces-

Decision dimension	Single-agent	Multi-agent
Workflow span	Single step or a single stage (e.g., one triage note)	Multiple stages with handoffs (triage → consult → discharge)
Role heterogeneity	One role with limited scope and constraints	Distinct roles (nurse/doctor/specialist/ops) with different constraints
Failure containment	One failure affects the whole output; rollback is coarse	Role boundaries localise failures; easier rollback and targeted repair
Handoff / audit requirement	Light logging; static evaluation often sufficient	Handoff logs and per-role accountability are required for auditing
Escalation requirement	Often manual review only; ad hoc escalation	Explicit escalation triggers and coverage checks are needed
Ops coupling	No queue/capacity model needed; weak coupling to operations	May need time-to-disposition / throughput / queue-aware planning
Cost profile (tokens/latency)	Lower tokens per case; fewer turns; lower wall-clock latency	More turns; higher tokens per case; coordination overhead
When to choose	Prefer when the task is narrow and low-risk, or when strong HIL is available	Prefer when workflows span roles/stages, require escalation, or must be audit-ready

Table 1: Single-agent vs multi-agent as a deployment decision checklist. Multi-agent overhead is justified mainly when workflows span stages and roles and require explicit handoffs, escalation, and audit-ready accountability.

Design	Handoff recover.	Staleness	Privacy	Store (min.)	Common failure
Sliding window	Low	Low	Low–Med	T + S	Contraindication drop; context loss
RAG retrieval	Med	Med	Med	Q + EID + C	Bad retrieval; uncited claim; stale ref
Event buffer	Med–High	Low–Med	Med	EV + TS	Missing event; state mismatch
Temporal KG	High	Low	High	ENT + TL + P	Patient mix-up; entity drift; PHI over-link

Table 2: Memory designs for AI hospitals. We compare designs by (i) **handoff recoverability** (can a new agent resume the workflow state after a handoff), (ii) **staleness risk** (how easily memory becomes outdated), and (iii) **privacy risk**. **Store (min.)** lists the minimum artifacts to persist for coordination and audit: T=recent turns, S=task state, Q=query, EID=evidence IDs, C=citations, EV=structured events, TS=timestamps, ENT=entities, TL=timeline, P=provenance; PHI=protected health information. **Common failure** highlights operational errors such as missing contraindications, retrieval mistakes, stale guidelines, and patient mix-up.

Tool class	Why it is used	Deployment constraint	What must be logged	Evaluation signals
Simulated EHR	Safe experimentation; synthetic cohorts	Low compliance risk; realism gap	Scenario seed; state transitions; tool-call traces	Safety: unsafe-output rate; Process: task completion; Outcome: rubric score; Ops: tokens/latency
FHIR-backed EHR	Realistic workflows; record-grounded decisions	Access control; PHI compliance; audit requirements	FHIR queries; retrieved fields; consent/audit IDs; access events	Safety: PHI leakage; Process: record correctness; Outcome: guideline concordance; Ops: latency/availability
Calculators / guideline checkers	Enforce norms (dose/triage) and constraints	Versioning; guideline drift; input validity	Input values; tool version/guideline version; outputs; exceptions	Safety: contraindication misses; Process: correct inputs; Outcome: agreement with guideline; Ops: call rate
Order / prior-authorisation APIs	Execute high-stakes actions (orders/booking)	Authorization chain; approval state; rollback; reliability	Request payload; approval status; overrides; rollback events; timestamps	Safety: invalid orders; Process: escalation correctness; Outcome: error rate; Ops: time-to-disposition
External clinical APIs (Phenomizer, DrugBank)	Up-to-date facts; coverage beyond local stores	Source trust; rate limits; reliability; drift	Query; returned IDs; source version/date; citations; failures	Safety: wrong drug facts; Process: citation coverage; Outcome: decision impact; Ops: failure rate
Research tools (omics, lab)	Deep modelling; automation for studies	Validation burden; domain expertise; reproducibility	Data provenance; code/tool version; parameters; runtime logs	Safety: misuse risk; Process: reproducibility; Outcome: match to study endpoints; Ops: runtime/cost

Table 3: Tool stack integration for AI hospitals: deployment constraints, required audit artifacts, and evaluation signals aligned with the safety–process–outcome–operations stack.

sary exposure of sensitive data. We distinguish long-term memory (internal parameters and external stores such as EHRs and guidelines) from short-term and shared working memory used within and across agents. Tools complement memory by retrieving evidence, enforcing constraints, and executing APIs or code.

Long-term memory. Internal memory supports zero/few-shot reasoning and fills missing attributes (Li et al., 2024e). Fine-tuning with real records improves adverse event and drug predictions (Wang et al., 2024d). External memory aggregates stable sources (NIH manuals, CCD, ESI) and knowledge graphs for evidence-based care (Lu et al., 2024). Dynamic updating via RAG/APIs adds new evidence with expert feedback (Yang et al., 2024). **Short-term and shared memory.** Short-term memory preserves local coherence then clears state (Liu et al., 2025). Shared working memory synchronises teams with summaries, traces, and buffers (Hong et al., 2024). Voting and meta-doctor consolidation refine group outputs (Tang et al., 2023). These memory designs mainly differ in whether a new agent can recover the workflow state after a handoff, and in how much sensitive information is exposed. Table 2 summarises this trade-off and highlights typical failure modes such as staleness, missing contraindications, and patient mix-up. Domain adaptation and uncertainty handling matter because they make retrieval and tool use more reliable and provide a clean signal for escalation and safer handoffs. This also explains why memory is often paired with tools and provenance: retrieval and calculators reduce staleness, while tool-call traces, evidence links, and versioned outputs make handoffs recoverable and decisions auditable.

Module	Design choice	When it fits	Cost in practice	What to log	Typical failure
Reasoning	Single-path CoT (linear) (Li et al., 2024e; Schmidgall et al., 2024)	Short, well-specified tasks	Low tokens/latency, but weak branch coverage	tokens/latency; citation rate; tool-call trace; rationale length; error tags	Misses sparse clues; error compounding
Reasoning	Multi-path self-consistency / debate (Tang et al., 2023; Kim et al., 2024)	Uncertain diagnosis; multiple hypotheses	High tokens/latency; higher controller and review burden	#branches; vote margin; disagreement rate; escalation triggers; tokens/latency	Divergent paths; hard to reconcile; controller overload
Memory	Retrospective (static) memory (Lu et al., 2024; Wang et al., 2024c)	Stable guidelines; complete histories	Low runtime cost, but staleness and low personalisation	memory hits; last-updated; source provenance; guideline version; profile id	Fails to absorb new evidence; repeats outdated guidance
Memory	Temporal knowledge graph (Chen et al., 2024b)	Longitudinal care; rare disease	Engineering and maintenance cost; privacy and access risk	node/edge updates; timestamps; provenance links; conflict flags; PHI access events	Drift; visit mismatch; wrong merge/split across time
Control/Tooling	Tool-first (plan then call) (Yue et al., 2024a; Du et al., 2024)	Regulated steps (dose/triage)	More upfront engineering; lower flexibility in edge cases	plan steps; tool schema/version; tool errors; retries; constraint checks	Over-rigid; poor adaptation; brittle when tools fail
Control/Tooling	LLM-first (react then tools) (Yu et al., 2024)	Open-ended, personalised talk	Unstable safety; higher verification and audit cost	tool provenance; citation coverage; safety flags; refusal rate; post-hoc check status	Uncited claims; policy slips; missing tool provenance

Table 4: Design decision map for common patterns. For each choice, we list when it fits, what it costs in practice, what to log, and what usually breaks, so trade-offs can be measured and later used for evaluation and gating.

Tool integration. Tools are not only for accuracy. They make actions executable and claims auditable by attaching evidence, versions, and approval states to each step. Table 3 summarises common tool classes, their deployment constraints (privacy, permissions, reliability, rollback), what must be logged for audit, and which evaluation-stack signals they enable. Retrieval systems and knowledge graphs ground reasoning in records and guidelines (Du et al., 2024; Kim et al., 2024). Decision trees and calculators enforce safe dose/triage (Yang et al., 2024; Li et al., 2023a). LLM-as-KB supports flexible synthesis (Yue et al., 2024a; Frisoni et al., 2024). Multi-modality tools fuse text, images, and sensors (Li et al., 2024e; Yang et al., 2024). Computational reasoning tools run code and simulations for research automation (Wang et al., 2024e; Hong et al., 2024).

This module follows the separation in clinical information systems: longitudinal records (EHR) store what happened, while decision-support tools enforce or suggest what should happen. AI hospitals must therefore treat memory, tools, and privacy constraints as core infrastructure, not as optional add-ons. This grounding also motivates evidence-based evaluation, where claims should be traceable to records, guidelines, or tool outputs. Additional details are provided in Appendix A.3 and A.4.

2.3 Reasoning & Control

This module is grounded in safety-critical workflow control, where autonomy is staged and constrained by explicit gates, escalation triggers, and audit logs. We do not treat reasoning style as an abstract choice; we treat it as a control problem that determines when the system can act, when it must ask for help, and what evidence must be recorded. This grounding motivates our later IRL gates and process metrics. Table 4 summarises the key trade-offs and the logging signals needed to operationalise later evaluation and gating. **1. Direct reasoning.** Single-path logic follows structured steps or expert systems; it suits clear tasks and stable guidelines (Li et al., 2024e). CoT pipelines (e.g., AgentClinic; AI nurse simulators) add transparency but still miss alternatives (Schmidgall et al., 2024). In practice, single-path designs should at least log citation coverage and tool-call traces, because their dominant failure is silent omission and error compounding (Table 4). **2. Multi-path reasoning.** Parallel branches plus voting/debate improve robustness under uncertainty (Kim et al., 2024). Rare-disease teams and research meetings aggregate diverse expertise for better differentials (Chen et al., 2024b). Because multi-path reasoning increases coordination load, systems should log disagreement signals (e.g., vote margin and divergence rate) and tie them to escalation or reconciliation rules (Table 4). **3. Feedback-based refinement.** External feedback from clinicians, knowledge bases, and tools corrects errors and updates plans (Johri et al., 2023; Li et al., 2024d). Self-feedback and reflection structure internal critique and reduce inconsistencies (Louie et al., 2024). Symbolic controllers coordinate tool calls, maintain traces, and enforce guardrails (Hong et al., 2024). Planner agents generate parallel solutions before commit (Liu et al., 2024). Here, the key control question is not whether feedback exists, but whether it is enforced by gates (policy/evidence/uncertainty) and recorded as audit-ready traces that support rollback and post-hoc review. Additional details are provided in Appendix A.5.

Overall, the goal shifts from static accuracy to process fidelity: better decisions, safer handoffs, and measurable harm avoided. Controllers should

expose an autonomy dial through explicit gates and escalation triggers, and they should log evidence for audit. Table 4 links common reasoning and control patterns to measurable signals (e.g., citation coverage, disagreement, tool provenance), which later support our evaluation stack and IRL-style gating. Together, roles & interaction, memory & tools, and reasoning & control define the AI-hospital architecture.

3 Applications

3.1 Simulating Specific Scenarios

Clinical workflow simulation. This line of work maps to the core visit workflow, from triage to consultation and discharge, with explicit state and handoffs across roles. It relies most on visit-level working memory plus record- or guideline-grounded tools (e.g., simulated EHR surrogates, retrieval over guidelines, and basic calculators) to keep decisions verifiable. It should be evaluated primarily by process and safety signals in our stack, such as step-wise guideline adherence, handoff completeness, escalation precision/recall, and unsafe-output rate, plus operations signals like time-to-disposition and tokens per case. Representative systems include end-to-end visit simulators and benchmark-style wards that instrument stage-level weaknesses and failure modes (e.g., Liu et al. (2025)).

Psychological counselling and mental-health interaction. This category maps to longitudinal, multi-turn support and de-escalation workflows, where safety and state tracking matter more than single-turn correctness. It relies on persistent patient state (mood, goals, risk signals) and on tools or rules that constrain responses to evidence-based practice (e.g., structured CBT-aligned prompts or safety policies). It should be evaluated by safety and process metrics, including crisis-risk handling, refusal and escalation behavior, conversation policy adherence, and traceable rationales, together with outcome measures such as engagement or education gain. Representative systems operationalize these ideas via expert-guided behavior rules and structured therapy principles (e.g., Louie et al. (2024)).

Multi-disciplinary medical team (MDT) simulation. These systems map to consultation and escalation workflows, where multiple specialists coordinate under uncertainty and must reconcile conflicting hypotheses. They rely on shared working memory (summaries, evidence buffers, disagreement traces) and on retrieval/tools that support evidence grounding (guidelines, knowledge bases, and structured records). They should be evaluated by process fidelity signals such as disagreement rate, reconciliation quality, escalation triggers, and audit-ready decision traces, plus operations costs due to coordination (latency and tokens per case). Representative systems include rare-disease MDTs with dynamic memory and tool use, and moderated expert panels that switch between single-doctor and MDT modes (e.g., Chen et al. (2024b)).

Most current simulators focus on a single encounter, while multi-visit management (follow-up, adherence, chronic care) remains limited. Additional details are provided in Appendix B.1.

3.2 Solving Complex Tasks

Clinical decision making. This category maps to high-stakes consultation and differential diagnosis workflows, especially for rare or complex cases. It relies on evidence-grounded memory and tools, such as structured retrieval over guidelines and records, multimodal inputs when available, and constrained calculators/checkers. It should be evaluated by safety and outcome signals (severity-weighted errors, contraindication misses, guideline concordance), plus process signals (citation coverage, tool-call provenance, and escalation behavior). Representative systems couple reasoning agents with domain tools and structured rationales to reduce brittle errors (e.g., Yang et al. (2024)).

Triage, routing, and clinical trials. This line maps to early-stage triage and routing, and to downstream operational decisions such as matching or scheduling. It relies on rule-constrained tools and guideline-aware retrieval, because inputs are noisy and actions must respect hard constraints. It should be evaluated by process metrics (correct inputs, escalation correctness, handoff completeness), and by operations metrics (latency, time-to-disposition, and failure/rollback rates), in addition to accuracy. Representative systems implement these workflows via guideline-grounded triage and service-flow routing (e.g., Lu et al. (2024); Bao et al. (2024)).

Knowledge-intensive workflows. This category maps to supporting clinical data work around the care workflow, such as EHR analytics, tool building, fact-checking, and knowledge curation. It relies on retrieval and tool-use infrastructure (code execution, structured databases, and provenance tracking). It should be evaluated by traceability and

reproducibility signals (evidence links, code/tool versions, and failure recovery). Representative examples include EHR analysis agents and tool-building pipelines that require strong provenance and verification (e.g., Shi et al. (2024b); Wang et al. (2024e)).

Scientific discovery. These systems map to research workflows adjacent to clinical practice, such as hypothesis generation, experiment planning, and domain analysis. They rely on specialised external tools (omics, molecular simulators, lab pipelines) and strict provenance so results can be validated. They should be evaluated by reproducibility and validation outcomes (match to study endpoints, error modes under tool failures), and by operations cost (runtime and compute/token budget). Representative systems automate multi-step discovery pipelines by coupling reasoning with domain tools (e.g., Swanson et al. (2024); Liu et al. (2024)). Additional details are provided in Appendix B.2.

3.3 Evaluating Agents

Evaluation in AI hospitals maps to different workflow stages, so it must score not only final answers but also interactive behaviors across steps and handoffs. It relies on instrumented logs (tool calls, evidence links, handoff traces, and escalation events) and on structured rubrics (e.g., OSCE-style checklists) to make process quality measurable. Accordingly, the most informative signals align with our stack: safety events and block time, process fidelity (handoff completeness and escalation precision/recall), outcome quality (severity-weighted errors and patient understanding), and operations (time-to-disposition, throughput, and tokens/latency). Representative forms include OSCE-style interactive exams and scalable judge-based scoring, which are increasingly used to measure reasoning and communication at scale (e.g., Arora et al. (2025)). Additional details are provided in Appendix B.3.

3.4 Synthesising Data for Training

We view data synthesis as a workflow that supports training and stress-testing AI hospitals. It relies on retrieval and adjudication infrastructure (grounded evidence, versioned prompts, and audit logs) to control validity, bias, and privacy exposure. Accordingly, it should be evaluated by traceability (evidence and provenance), diversity and coverage of edge cases, and downstream impact on safety/process/outcome metrics after training. Representative pipelines use multi-agent self-play and structured generation loops to reduce annotation cost while keeping clinical constraints explicit (e.g., Wang et al. (2023a); Tu et al. (2025)). Additional details are provided in Appendix B.4.

4 Key Challenges and Future Directions

AI-hospital challenges rarely sit in a single module; they arise as workflow failures across roles, memory, tools, and reasoning. We make these challenges actionable with two artifacts: Table 5 defines instrumentation (minimum logs and metrics) for key integration tasks, while Table 6 defines stage-wise gates that regulate autonomy from sandbox to deployment.

4.1 Cross-Module Failure Patterns

Hallucination and safety. High-confidence errors propagate across teams: unsafe plans, missed escalation, and PHI leakage often emerge only at handoffs and tool boundaries. We therefore require safety to be measurable and auditable: Table 5 logs gateway decisions, block time, PHI checks, and escalation events, and Table 6 enforces them as hard gates from IRL2 through IRL4.

Tool integration. In practice, failures come from schema drift, API mismatches, and silent fallbacks that erase provenance and make recovery impossible. Table 5 therefore mandates tool-call traces (inputs/versions/outputs) plus approval/rollback states for integration tasks, which become promotion blockers in Table 6 beyond IRL3 shadow replay.

Evaluation. Benchmark-only scores hide the process: handoffs, escalation coverage, latency tails, and human touches. We treat evaluation as instrumentation: Table 5 binds each metric to concrete log sources (gateway, templates, EHR/tool audits, queue/event logs), and Table 6 uses them as release criteria.

Data synthesis. Synthetic data scales, but it can amplify bias and collapse diversity without provenance and reviewer traces. We therefore require dataset provenance and drift-audit logs (Table 5), and we IRL-label deployment claims: synthetic-only evidence rarely justifies gates beyond IRL2/IRL3 without real-data shadow replay (Table 6).

Worked example (instrumentation template). Consider a discharge education agent at discharge → follow-up. *Roles/control:* NurseEducator delivers instructions; SafetyGateway filters; Attending-

Integration task	Key limits / failure modes	What to log (minimum artifacts)	Metrics to report (Safety / Process / Outcome / Ops)	Frequent pitfalls
Clinic scheduling / bed management	Finite capacity; peak demand; transport delays; plans ignore queue state	Queue simulator outputs; bed-board snapshots; transport timestamps; session IDs; escalation events under overload	**Safety:** unsafe-output rate; block time (if user-facing) **Process:** escalation precision/recall under overload **Outcome:** severity-weighted errors (avoid unsafe speedups) **Ops:** TTD; throughput; queue-length percentiles; end-to-end latency; tokens/case; human interventions	Reporting throughput only; ignoring severity-weighted errors; not splitting peak vs off-peak; using mean latency only (no P90/P99)
Order routing / prior authorisation	External API latency; missing forms; policy mismatch; silent fallback	Order request payloads; approval state transitions; tool/API versions; retries/errors; block reasons; timestamps; overrides/rollback events	**Safety:** unsafe-output rate; block time (unsafe orders) **Process:** guideline adherence@step; escalation P/R; order-completion rate; order-execution delay **Outcome:** error severity (wrong/late orders) **Ops:** end-to-end latency; human interventions; tokens/case	Measuring only benchmark gains; missing tool/version logs; ignoring rollback; no timing for escalation chain; fixed thresholds under drift
Shift handoffs (ED → ward)	Lossy summaries; missing critical fields; no traceability across departments	Structured handoff template fields; evidence IDs + citations; retrieved record fields; tool-call traces; cross-shift state snapshots; audit scripts	**Safety:** unsafe-output rate (handoff advice); PHI leakage rate (if EHR-linked) **Process:** handoff completeness; guideline adherence@step; citation coverage; cross-shift consistency **Outcome:** task success; severity-weighted errors **Ops:** handoff time; end-to-end latency; human interventions; tokens/case	Text looks fluent but fields missing; cross-dept schema mismatch; no evidence chain; judging final answer only (no step-level checks)
Discharge education / follow-up	One-shot counselling; no retention check; language mismatch; caregiver factors ignored	Pre/post quiz scores; OSCE-style rubrics; multilingual template version; call-center logs (7/30-day); escalation events for low comprehension; session timestamps	**Safety:** unsafe-output rate; block time (unsafe counseling) **Process:** handoff completeness (what was taught + citations); escalation P/R for low comprehension/risk **Outcome:** education gain; discharge understanding; follow-up retention proxies (7/30-day callbacks) **Ops:** end-to-end latency; human interventions; tokens/case	Single post-test only; no retention or follow-up; ignoring multilingual/caregiver effects; reporting outcome without process logs
EHR integration (FHIR)	Fragile endpoints; PHI leakage; schema drift; permission failures	FHIR access audits; consent/audit IDs; queried resources/fields; de-identification audits; chaos-test logs (latency/schema faults); citation-to-record links	**Safety:** PHI leakage rate; unsafe-output rate; block time **Process:** citation-to-record consistency; chaos-test pass rate; guideline adherence@step (record-grounded) **Outcome:** task success; error severity (wrong record/field) **Ops:** API latency/availability; end-to-end latency; human interventions; tokens/case	Ignoring access control and audit chains; not versioning schemas/tools; measuring only retrieval scores; no chaos testing under drift

Table 5: **Metric-to-workflow instrumentation map for clinical integration.** For common hospital integration tasks (bed management, prior authorization, shift handoffs, discharge education, and FHIR/EHR integration), we map real-world failure modes to the minimum logging artifacts and the metrics to report across the safety–process–outcome–operations stack, highlighting pitfalls that make offline benchmarks look good while hiding deployment-critical failures.

on-call handles mandatory escalation on high-risk meds or missing follow-up evidence. *Memory/tools:* read FHIR discharge summary/med list; cite versioned guideline snippets; call a dose/contraindication checker. *Logs (Table 5):* tool-call traces (inputs/versions/outputs), cited EIDs, handoff-template fields, gateway flags/block time, escalation events with reason codes, and session timestamps/tokens. *Metrics:* unsafe-output rate, handoff completeness, escalation P/R, discharge understanding (quiz rubric), and end-to-end latency/tokens per case. *IRL rule (Table 6):* remain IRL3 shadow replay until audit-ready handoffs pass thresholds with zero severe events; enter IRL4 pilot only if escalation recall meets the gate and P95 latency stays within budget.

4.2 Measurement for Integration Tasks

Clinical integration fails when actions are not replayable: policies ignore queues/capacity, and tool/EHR calls lose provenance or auditability. Table 5 therefore treats integration as workflow tasks with minimum logs that jointly measure safety, process, outcomes, and operations in the *same run*.

Where systems break. Breakpoints typically coincide with missing logs: handoffs omit critical fields or provenance; orders stall on authorization states without safe retry/rollback; queue state is ignored so plans fail under load; EHR/FHIR calls leak PHI or degrade under schema drift. These task-linked logs make failures testable (Table 5) and become release criteria: IRL3 requires shadow replay with audit-ready evidence, while IRL4+ adds escalation coverage and latency tails as hard gates (Table 6).

4.3 Testable Hypotheses and Release Gates

T1. Temporal and longitudinal correctness. Models mix past history with current state, so workflows drift across visits. *Test:* event-graph reconstruction accuracy; follow-up agreement; relapse detection sensitivity. *Instrumentation:* temporal KG or event-indexed logs with timestamps, EHR replay traces, and cross-handoff consistency checks (Table 5). *Gate:* evaluate primarily at IRL3 shadow replay where replayability and longitudinal consistency are required (Table 6). *Direction:* workflow-aware memory to improve shift handoffs and longitudinal follow-up, and to unlock IRL3 handoff completeness and replayable evidence chains.

T2. Operations realism under capacity limits. Plans ignore beds, staff, and transport delays, so queue state breaks otherwise-correct reasoning. *Test:* discrete-event simulation; blockage probability; change in time to disposition and throughput under stress; tail latency under load. *Instrumentation:* queue simulator outputs, bed-board snapshots, transport timestamps, and end-to-end event logs for TTD/throughput and latency tails (Table 5). *Gate:* start at IRL3 stress-tested shadow replay, then require IRL5-style ops non-regression before broader

IRL	Scenario & scope	Autonomy setting	Gate types	Escalation triggers & coverage	Required logs/audits (minimum)	Pass criteria (examples)
IRL1	Sandbox simulation; static/scripted cases; no external dependencies	HIL-default (manual review before any commit)	Manual checkpoints; safe defaults; offline review	Escalation by human judgment; coverage not yet measurable	Prompt/response logs; basic safety tags; task scripts; seed/config snapshot	Unsafe output $< 2\%$; task success $>$ baseline
IRL2	Noisy or missing-info simulation; bias and adversarial seeds	HIL-default with stricter safety gateway	Manual checkpoints + safety gateway (policy/refusal)	Trigger: human + gateway alerts; start measuring escalation events on seeded high-risk cases	Gateway logs (unsafe flags, block reasons); block time timestamps; jailbreak pass logs; citation coverage stats	Block time $<$200ms; jailbreak pass $< 0.5\%$; citation coverage above threshold
IRL3	Real-data replay (shadow mode); no live impact	HIL-default; model proposes, humans decide; no autonomous actions	Manual checkpoints + evidence gates (policy+evidence)	Trigger: missing evidence, high uncertainty, conflict; coverage measurable via labeled risk cases in replay	EHR access audit; replayable evidence chains (EID + citations); tool-call traces; handoff template checks; versioned prompts/models	Adherence@step $>$90%; handoff completeness $>$0.9; PHI leakage $= 0$ in audit
IRL4	Human-in-the-loop pilot (limited unit/time)	Auto-suggest with mandatory approval for high-stakes steps (disposition/orders)	Policy + evidence + uncertainty gates; explicit escalation rules; rollback protocol	Trigger: uncertainty/conflict/missing evidence/policy risk; coverage enforced (precision/recall) with reason codes	Escalation logs (who/why/when); human–AI disagreement postmortems; override outcomes; P95 latency logs; incident tickets	Escalation recall $>$95%; zero severe misses; P95 latency under threshold; stable unsafe-output rate
IRL5	Limited rollout; end-to-end monitoring and post-hoc audits	Auto-default with exception review; humans handle escalations and audits	Automatic gates + canary strategy; version freeze; incident response plan	Trigger: risk alerts, drift alerts, queue overload; coverage monitored continuously + periodic audits	End-to-end monitoring dashboard; incident response logs; drift monitoring; audit-ready traces (inputs, tool calls, evidence, overrides)	TTD/throughput not worse than human baseline; tokens/case within budget; no major incidents
IRL6	Scaled deployment; multi-site/multi-language	Auto-default with mature gates; humans as supervisors + auditors	Automatic gates + A/B with guardrails; multi-site drift controls	Trigger: site/language drift, policy changes, tool failures; coverage stable across sites/languages	Multi-site drift monitoring; A/B logs; monthly cost–benefit reports; periodic red-teaming; compliance audits	Sustained quarterly stability; positive ROI; zero major incidents; cross-site/language parity targets

Table 6: **Deployment gating playbook for AI-hospital agents.** We summarize a stage-by-stage path from offline sandboxing to scaled deployment (IRL1–IRL6), and at each stage specify (i) the autonomy setting (human-in-the-loop $\rightarrow$ auto-default), (ii) gate types (policy / evidence / uncertainty), (iii) escalation triggers with measurable coverage, and (iv) the minimum audit-ready logs required to justify progression—turning "readiness" into explicit, testable release criteria.

rollout (Table 6). *Direction:* operations-coupled planning that queries queue/capacity before commit, to meet IRL5 ops gates.

T3. Guideline control, deviation, and escalation. Deviations lack reasons, and single-path reasoning often fails to abstain or escalate. *Test:* variance-aware adherence per step with reason audit; win/loss accounting when deviating; escalation precision/recall on high-risk cases. *Instrumentation:* guideline versioning, per-step adherence logs, citation/EID coverage, deviation reason codes, and escalation logs (Table 5). *Gate:* validate under IRL2 adversarial/noisy testing, confirm under IRL3 replay, and enforce IRL4 escalation-recall requirements in pilots (Table 6). *Direction:* escalation learning with calibrated reasons, targeting triage/counseling/med-rec to unlock IRL4 gates.

T4. Auditability, accountability, and privacy. Without signed decisions and replay, multi-agent errors propagate, and EHR connectors add PHI and audit risks. *Test:* signed action logs; replay attribution; PHI-leakage under adversarial prompts; FHIR chaos tests; citation-to-record consistency. *Instrumentation:* signed, role-tagged action logs, per-role decision records, replay scripts, FHIR access audits, PHI leakage detectors, and chaos-testing logs (Table 5). *Gate:* require IRL3 audit-ready replay, and IRL4 pilots with mandatory escalation chains and postmortems (Table 6). *Direction:* traceability and accountability to support cross-shift handoffs and order routing, and to pass IRL3–IRL4 audit gates.

T5. Safety nets under distribution shift and promotion discipline. Guard models drift, prompts jailbreak, and token/latency spikes cause unusable workflows. *Test:* periodic red teaming; reinjection of failure seeds; jailbreak pass rate; mean time to control recovery; per-case token and tail latency budgets. *Instrumentation:* red-team seed registry, gateway logs (unsafe flags, block time, false blocks), recovery-time traces, tokens per case, and human touches with reason codes (Table 5). *Gate:* measure primarily at IRL2 for adversarial readiness, then enforce IRL4/IRL5 latency and cost budgets during pilots and rollouts (Table 6). *Direction:* playbook adoption: each system reports an IRL label and an instrumentation checklist, and autonomy increases only when gates pass.

5 Conclusion

AI hospitals provide a workflow-level paradigm for studying teams of language agents under safety, tool, and time constraints. We define the scope, distill the design space into a compact taxonomy, and connect evaluation to measurable signals across safety, process, outcomes, and operations. Progress requires longitudinal, workflow-aware memory, operations-coupled planning, reliable escalation, and end-to-end traceability, with IRL-style gates and cost reporting to prevent overclaim and support deployment-relevant comparison.

6 Limitations

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 preprints (arXiv, medRxiv, bioRxiv), so relevant work outside these channels may be missing. Because the field evolves rapidly, the taxonomy and comparisons may require updates as new deployment practices, regulations, and clinical integration standards emerge.

While this review 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, which may increase harm via missed escalation, 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 C. Jain. 2023. Conversational health agents: A personalized llm-powered agent framework. *ArXiv*, abs/2310.02374.

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*.

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*.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024a. Cod, towards an interpretable medical agent using chain of diagnosis. *ArXiv*, abs/2407.13301.

Siyuan Chen, Mengyue Wu, Ke Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *ArXiv*, abs/2305.13614.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. 2024b. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. *arXiv preprint arXiv:2412.12475*.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, et al. 2025. Verifact: Verifying facts in llm-generated clinical text with electronic health records. *arXiv preprint arXiv:2501.16672*.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. 2025. Automating evaluation of ai text generation in healthcare with a large language model (llm)-as-a-judge. *medRxiv*, pages 2025–04.

Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haohao Ying. 2024. Llms can simulate standardized patients via agent coevolution. *arXiv preprint arXiv:2412.11716*.

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.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *International Conference on Computational Linguistics*.

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 *Annual Meeting of the Association for Computational Linguistics*.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.

S. Gilbert, J. N. Kather, and A. Hogan. 2024. Augmented non-hallucinating large language models as medical information curators. *npj Digital Medicine*, 7:100.

Alex J. Goodell, MD MS Simon N Chu, Dara Rouholiman, and MD Larry F Chu. 2023. Augmentation of chatgpt with clinician-informed tools improves performance on medical calculation tasks. In *medRxiv*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling, et al. 2024. A generative pretrained transformer (gpt)–powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR medical education*, 10(1):e53961.

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.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W. John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and Zhiyong Lu. 2024. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *ArXiv*, abs/2402.13225.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *ArXiv*.

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.

Mutahira Khalid, Raihana Rahman, Asim Abbas, Sushama Kumari, Iram Wajahat, and Syed Ahmad Chan Bukhari. 2024. Accelerating medical knowledge discovery through automated knowledge graph generation and enrichment. In *International Knowledge Graph and Semantic Web Conference*, pages 62–77. Springer.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid A. Anwar, Andrew Zhang, Aidan Gilson, Maxwell Singer, Amisha D. Dave, Andrew Taylor, Aidong Zhang, Qingyu Chen, and Zhiyong Lu. 2024. Medcalc-bench: Evaluating large language models for medical calculations. *ArXiv*, abs/2406.12036.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms in medical decision making. *arXiv preprint arXiv:2404.15155*.

Prerana Sanjay Kulkarni, Muskaan Jain, Disha Sheshanarayana, and Srinivasan Parthiban. 2024. Hecix: Integrating knowledge graphs and large language models for biomedical research. *arXiv preprint arXiv:2407.14030*.

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. 2023a. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Binxu Li, Tiankai Yan, Yuanting Pan, Zhe Xu, Jie Luo, Ruiyang Ji, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. Mmedagent: Learning to use medical tools with multi-modal agent. In *Conference on Empirical Methods in Natural Language Processing*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024b. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

J. Li, X. Chen, W. Liu, L. Wang, Y. Guo, M. You, others, and K. Li. 2023b. One is not enough: Multi-agent conversation framework enhances rare disease diagnostic capabilities of large language models.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024c. Agent hospital: A simulacrum of hospital with evolvable medical agents. *ArXiv*, abs/2405.02957.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024d. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Neural Information Processing Systems*.

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

Zhenzhu Li, Jingfeng Zhang, Zhou Wei, Jianjun Zheng, and Yinshui Xia. 2024f. Gpt-agents based on medical guidelines can improve the responsiveness and explainability of outcomes for traumatic brain injury rehabilitation. *Scientific Reports*, 14.

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*.

Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z Li, and Kaicheng Yu Biokgbench. 2024. A knowledge graph checking benchmark of ai agent for biomedical science. *arXiv preprint arXiv*, 2407.

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. *ArXiv*, abs/2411.15692.

Zhaocheng Liu, Quan Tu, Wen Ye, Yu Xiao, Zhishou Zhang, Hengfu Cui, Yalun Zhu, Qiang Ju, Shizheng Li, and Jian Xie. 2025. 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. *arXiv preprint arXiv:2407.00870*.

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.

Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. 2024. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6).

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.

Daniela SM Pereira, Filipe Falcão, Andreia Nunes, Nuno Santos, Patrício Costa, and José Miguel Pêgo. 2023. Designing and building oscebot® for virtual osce–performance evaluation. *Medical Education Online*, 28(1):2228550.

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*.

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 *Conference on Empirical Methods in Natural Language Processing*.

Andries P. Smit, Paul Duckworth, Nathan Grinsztajn, Kale ab Tessera, Thomas D. Barrett, and Arnu Pretorius. 2023. Should we be going mad? a look at multi-agent debate strategies for llms. In *International Conference on Machine Learning*.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Zhongyi Shui, Xiaoxuan Yu, Yizhi Zhao, Honglin Li, Yunlong Zhang, Ruojia Zhao, Xinheng Lyu, and Lin Yang. 2023. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *AAAI Conference on Artificial Intelligence*.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. 2024. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pages 2024–11.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

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*, pages 1–9.

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 Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2023a. Notechat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Annual Meeting of the Association for Computational Linguistics*.

Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024c. 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, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023b. Augmenting black-box llms with medical textbooks for clinical question answering. *ArXiv*, abs/2309.02233.

Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. 2024d. 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. 2024e. Can large language models replace data scientists in clinical research? *arXiv preprint arXiv:2410.21591*.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Tianlong Wang, Wen Tang, Yasha Wang, Chengwei Pan, Ewen M Harrison, Junyi Gao, et al. 2024f. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. *arXiv preprint arXiv:2410.02551*.

Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. 2024a. Medco: Medical education copilots based on a multi-agent framework. *arXiv preprint arXiv:2408.12496*.

Jinjie Wei, Dingkang Yang, Yanshu Li, Qingyao Xu, Zhaoyu Chen, Mingcheng Li, Yue Jiang, Xiaolu Hou, and Lihua Zhang. 2024b. Medaide: Towards an omni medical aide via specialized llm-based multi-agent collaboration. *arXiv preprint arXiv:2410.12532*.

Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. *ArXiv*, abs/2307.04986.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, MingZhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, Shaoqing Jiao, and Jiajie Peng. 2024. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *bioRxiv*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. *ArXiv*, abs/2402.13178.

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*.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 30:199–214.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, Li Zhu, and Xuandong Zhao. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *ArXiv*, abs/2406.13890.

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.

Zonghai Yao, Aditya Parashar, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, Zhichao Yang, and Hong Yu. 2024a. Mcqg-srefine: Multiple choice question generation and evaluation with iterative self-critique, correction, and comparison feedback. *arXiv preprint arXiv:2410.13191*.

Zonghai Yao, Zihao Zhang, Chaolong Tang, Xingyu Bian, Youxia Zhao, Zhichao Yang, Junda Wang, Huixue Zhou, Won Seok Jang, Feiyun Ouyang, and Hong Yu. 2024b. Medqa-cs: Benchmarking large language models clinical skills using an ai-sce framework. *ArXiv*, abs/2410.01553.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Wenyue Hua, Mingyu Jin, Guang Chen, et al. 2024. Aipatient: Simulating patients with ehrs and llm powered agentic workflow. *arXiv preprint arXiv:2409.18924*.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024a. 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. 2024b. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. *arXiv preprint arXiv:2407.09893*.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan A. Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curtis P. Langlotz, Rita Lee, Joanna Melia, Joanna Nelson, Karim Sallam, Stacey Tullis, Melissa Ann Vogelsong, John Patrick Cunningham, and William Hiesinger. 2024. Almanac - retrieval-augmented language models for clinical medicine. *NEJM AI*, 1 2.

Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. 2023. Llm-based medical assistant personalization with short- and long-term memory coordination. In *North American Chapter of the Association for Computational Linguistics*.

Wei Zhu, Wenfeng Li, Xing Tian, Pengfei Wang, Xiaoling Wang, Jin Chen, Yuanbin Wu, Yuan Ni, and Guotong Xie. 2024. Text2mdt: extracting medical decision trees from medical texts. *arXiv preprint arXiv:2401.02034*.

A Core Components

A.1 Agent Roles

Patient-Centered Agents are designed to simulate patients with different demographic backgrounds, health conditions, and communication abilities. **Patient Agent** supports various applications in AI hospitals, such as clinical training, patient education, and medical history collection. Many works (Bao et al., 2024; Wang et al., 2023a) focus on enhancing the realism of patient agents. Recent studies (Du et al., 2024; Li et al., 2024e; Yu et al., 2024; Liu et al., 2025) also leverage evolutionary learning, fine-tuning techniques, Chain-of-Thought (CoT), and Retrieval-Augmented Generation (RAG) to enhance patient agents' consistency, realism, and role-playing stability while reducing hallucinations. **Psychological Patient Agent** (PPA) simulates mental health conditions for AI-driven treatment training (Wang et al., 2024b; Wei et al., 2024a). Unlike general patient agents, PPAs must replicate mood changes, cognitive distortions, and treatment resistance, with studies focusing on authenticity through expert-guided prompt engineering (Louie et al., 2024), structured cognitive modeling (Wang et al., 2024c), and simulations fostering adaptive communication (Chen et al., 2023).
Resident Agents model general populations transitioning into patient agents when ill, autonomously navigating healthcare processes while also supporting public health simulations and epidemiological modeling by incorporating disease progression, healthcare-seeking behavior, and policy interventions (Li et al., 2024c; Williams et al., 2023).
Medical Professional Agents can perform tasks such as patient consultation, medical history collection, clinical reasoning, diagnostic decision-making, emotional support, care coordination, and auxiliary examinations. **General Doctor Agent** , often called primary care physician (PCP), performs initial patient assessments and oversees the diagnostic process. Several studies have explored various aspects of these agents, including their questioning strategies (Liu et al., 2025), autonomous learning for diagnostic optimization (Du et al., 2024), reasoning in clinical conversations (Johri et al., 2023), adaptive multi-agent collaboration (Kim et al., 2024), the role of PCP in diagnosis (Wang et al., 2024a), and their integration into AI hospital environments (Fan et al., 2024).
Specialist Agent represents domain-specific medical experts such as cardiologists, radiologists, and

hematologists, for handling complex cases and contributing expert knowledge to diagnostic and treatment decision-making. Specialist agents require high-precision reasoning, deep medical expertise, and the ability to collaborate effectively in multidisciplinary team (MDT). Many works (Chen et al., 2024b; Kim et al., 2024) highlight the benefits of structured expertise, domain-specific knowledge, and coordinated decision-making in the AI Hospital. **Therapist Agent** provides emotional support, psychological intervention and psychotherapy (Wang et al., 2024b; Qiu and Lan, 2024; Chen et al., 2023). **Nurse Agent** facilitates triage, basic care and patient coordination (Bao et al., 2024; Li et al., 2024c). **Medical Technician Agents** aid diagnostic procedures, ensuring accurate test results (Schmidgall et al., 2024). **Medical students & Examiner agent** Simulate clinical training to improve medical history collection and diagnostic skills (Li et al., 2024e; Yao et al., 2024b).

Medical AI Teamwork Agents collaborate to tackle complex AI hospital tasks beyond a single agent's capacity. They handle information extraction, reasoning, and decision-making in disease analysis, diagnosis, patient triage, medical planning, and final decisions. **Goal-Driven Reasoning Agent** coordinates multi-step reasoning using structured pipelines, dual-agent frameworks, and symbolic reasoning (Yu et al., 2024; Hong et al., 2024; Shi et al., 2024b). **Clinical Judge Agent** ensures AI-driven diagnoses meet accuracy, effectiveness, and guideline adherence (Johri et al., 2023; Yue et al., 2024a). **Critic Agent** refines reasoning, mitigates biases, and enhances reliability through structured feedback (Ke et al., 2024; Hong et al., 2024). **Planning Agent** decomposes tasks, optimizes workflows, and improves triage and structured conversations (Yue et al., 2024a; Shi et al., 2024a). **Decision Agent** mediates conflicting assessments and synthesizes insights for coherent, evidence-based diagnoses (Tang et al., 2023; Wang et al., 2024f). **Recording Agent** logs key medical insights (Ke et al., 2024; Yu et al., 2024).

AI-Assisted Research Agents optimize new knowledge discovery, research support, and scientific review. **Research Planning Agent** plays a crucial role in structuring research tasks and ensuring efficient problem decomposition in complex domains, leveraging hierarchical decision-making and adaptive optimization to refine research strategies and enhance scientific impact (Swanson et al., 2024; Xiao et al., 2024). **Research Executor Agent** facilitates clinical research by assisting in hypothesis testing, statistical analysis, and experiment interpretation, leveraging domain-specific expertise to optimize research workflows and minimize execution failures (Swanson et al., 2024; Xiao et al., 2024). **Scientific Critic Agent** is responsible for assessing the quality and validity of AI-generated solutions, ensuring reliable decision-making in research and clinical settings(Xiao et al., 2024). **Database Agent** is designed to retrieve, manage, and integrate medical information for improved decision-making (Shi et al., 2024b).

A.2 Interaction Patterns

AI Hospital employs different interaction patterns to enhance efficiency, reliability, and decision-making. **Task-Focused Collaboration** decomposes complex medical tasks into structured subtasks for efficiency and consistency. Modular architectures follow predefined workflows to accomplish tasks, such as the ERRG workflow (Extract, Retrieve, Rewrite, Generate) (Li et al., 2024e). Multi-agent systems like AIPatient(Yu et al., 2024), ClinicalAgent (Yue et al., 2024a), and EHRAgent (Shi et al., 2024b) assign roles and execute tasks sequentially to enhance reasoning and decision-making. **Expert-Guided Decision-Making** ensures AI-driven medical decisions are clinically reliable. Multiple studies (Du et al., 2024; Chen et al., 2024b; Kim et al., 2024; Tang et al., 2023) emphasize expert integration in decision-making, research, and medical education, ensuring domain expertise and consensus validation. **Iterative Problem Optimization (IPO)** refines problem-solving through feedback loops. AI agents iteratively adjust queries (Yu et al., 2024), refine diagnostic interactions via conversational and reflection-based corrections (Du et al., 2024; Bao et al., 2024), and critique each other's reasoning (Tang et al., 2023). Programming agents iteratively enhance code accuracy (Shi et al., 2024b). **Automated Knowledge Integration (AKI)** merges diverse medical knowledge and patient data for accurate, context-aware decision-making. Techniques include knowledge-enhanced retrieval (Shi et al., 2024a), memory-based integration (Liao et al., 2024), and Directed Acyclic Graph (DAG)-based structuring (Du et al., 2024). Multi-modal approaches combine structured and unstructured EHR data, sensor inputs, and medical evidence (Yang et al., 2024), while team-based models apply adaptive fusion (Wang et al., 2024a), confidence validation (Lu et al., 2024), and struc-

tured reasoning (Hong et al., 2024). **Role-based Coordination** assigns AI agents specific roles (e.g., physicians, therapists, or patients) to simulate medical interactions and enhance diagnosis, training, and decision-making (Du et al., 2024; Wang et al., 2024b; Qiu and Lan, 2024). Multi-disciplinary AI teams integrate specialists' insights into comprehensive diagnoses (Wang et al., 2024f; Chen et al., 2024b). Systems like AgentClinic (Schmidgall et al., 2024) and Agent Hospital (Li et al., 2024c) expand role-based AI applications to triage, reception, and follow-ups. **Multi-Round Interactive Debate** fosters structured discussions where AI agents critique, resolve disagreements, and refine conclusions (Fan et al., 2024; Li et al., 2023b; Kim et al., 2024). Approaches employ voting (Tang et al., 2023), debate strategies (Smit et al., 2023), and confidence-based stopping (Lu et al., 2024). AI-driven research teams apply debate mechanisms to synthesize findings (Swanson et al., 2024).

A.3 Tool Integration

In AI hospitals, agents use diverse tools to enhance efficiency and accuracy. For example, **Retrieval systems** ensure rapid access to medical knowledge by dynamically retrieving patient records and evidence-based guidelines, aiding both patient and doctor agents in contextual reasoning (Du et al., 2024; Kim et al., 2024). **Knowledge graphs** structure medical knowledge into interconnected networks, enabling AI systems to navigate relationships between symptoms, treatments, and medical histories for informed decision support (Li et al., 2024e; Yu et al., 2024; Chen et al., 2024b). **Medical decision trees** provide structured diagnostic pathways, ensuring AI-driven recommendations align with established clinical guidelines and expert knowledge (Yang et al., 2024; Li et al., 2023a). **LLM-as-KB** transforms LLMs into dynamic knowledge repositories, allowing AI to synthesize medical insights beyond static databases (Yue et al., 2024a; Frisoni et al., 2024). **Smart devices and sensor data** integration facilitate real-time health monitoring, merging wearable data with EHR insights to enhance predictive analytics and personalized care (Yang et al., 2024; Abbasian et al., 2023). **Multi-modality processing tools** enable AI hospitals to integrate textual, visual, and sensor data, improving tasks such as radiology interpretation and decision tree-based diagnostics (Li et al., 2024e; Yang et al., 2024; Li et al., 2024a). **Computational reasoning tools** equip AI with logical inference and code execution capabilities, supporting automated clinical research and data-driven modeling (Wang et al., 2024e; Hong et al., 2024). Finally, some **other clinical decision support tools** optimize diagnostic accuracy by leveraging external APIs, existing predictive models/systems, and structured reporting systems (Wang et al., 2024a; Li et al., 2024a). And some **other biomedical research tools** accelerate drug discovery and genomic analysis, enabling AI-powered advancements in computational biology and molecular medicine (Swanson et al., 2024; Jin et al., 2023; Liu et al., 2024).

A.4 Memory Management

AI Hospital leverages structured memory management for adaptive learning and decision-making. **Long-Term Memory (LTM)** retains knowledge across sessions, integrating internal model updates and external databases for enhanced reasoning. **Internal Memory** embedded in the model parameters serves as a foundational knowledge repository for the agent to support zero-shot and few-shot tasks. For example, Li et al. (2024e) leverages the inherent common-sense knowledge within LLMs to supplement missing information in clinical case graphs, ensuring the generation of plausible attributes based on pre-existing knowledge. Wang et al. (2024d) integrates internal memory by fine-tuning ChatGPT with real patient clinical records, resulting in more accurate adverse event and drug predictions. **External Memory** supplements AI hospital systems with structured knowledge from databases, knowledge graphs, and retrieval systems while enabling real-time adaptation. **Static Storage** maintains long-term, structured knowledge, such as NIH resources for disease-specific agents (Wang et al., 2024a), CCD for patient history (Wang et al., 2024c), and structured ESI manuals (Lu et al., 2024). Medical knowledge databases, textbooks, and diagnostic guidelines serve as stable references (Yang et al., 2024; Shi et al., 2024a; Yue et al., 2024b), while drug knowledge graphs and clinical trial registries support evidence-based decision-making (Chen et al., 2024b; Yue et al., 2024a; Liu et al., 2024). **Dynamic Updating** integrates real-time knowledge via retrieval systems and APIs, refining AI behavior with expert feedback (Louie et al., 2024), synchronizing clinical guidelines (Yang et al., 2024), and leveraging PubMed or GitHub updates (Wang et al., 2024e). Additionally, long-term memory enhances task ex-

ecution by retrieving past cases (Shi et al., 2024b; Schmidgall et al., 2024; Bao et al., 2024), preserving user preferences like recurring health concerns for personalized responses.

Short-Term Memory (STM) and **Multi-Agent Shared Working Memory (WM)** serve complementary roles in AI hospitals and medical dialogue systems, ensuring context retention, reasoning consistency, and collaborative decision-making. STM is a temporary, agent-specific memory that maintains coherence during task execution but is cleared afterward (Liu et al., 2025). Medical dialogue systems use dialogue history, entity extraction, or summaries to mitigate forgetfulness and enhance reasoning. In contrast, WM is a globally shared memory facilitating knowledge synchronization, feedback integration, and structured reasoning across agents. It supports dynamic inference buffers, execution trace retention, and cross-agent coordination. For instance, Lu et al. (2024) updates summary reports for diagnostic consistency, while Hong et al. (2024) structures symbolic inference steps. WM also optimizes iterative decision-making (Kim et al., 2024; Xiao et al., 2024), reducing redundancy by storing shared task outcomes (Xiao et al., 2024). Feedback integration enhances refinement, as seen in expert voting (Tang et al., 2023), meta-doctor consolidation (Wang et al., 2024f), and structured critique cycles (Swanson et al., 2024).

A.5 Reasoning Mechanisms

Direct: derives conclusions through structured logic without external feedback. **Single-path** follows a linear progression, where each step builds on the previous one, as seen in ERRG (Li et al., 2024e), cognitive conceptualization maps (Wang et al., 2024c), ClientCAST (Wang et al., 2024b), and medical diagnostic frameworks like MDAgents (Kim et al., 2024) and expert systems (Yan et al., 2024). CoT-based approaches include AgentClinic (Schmidgall et al., 2024), AI nurse simulators (Bao et al., 2024), CoT-driven coding (Wang et al., 2024e), least-to-most reasoning in clinical agents (Yue et al., 2024a), and Chain-of-Diagnosis models (Chen et al., 2024a). **Multi-path** enables parallel inference for flexible decision-making, integrating multi-agent systems like EvoPatient (Du et al., 2024), RareAgents (Chen et al., 2024b), MDAgents (Kim et al., 2024), and MedAgents (Tang et al., 2023). Other methods leverage multi-agent collaboration (Wang et al., 2024f), expert self-consistency (Li et al., 2024d), and symbolic reasoning (Wang et al., 2024a; Hong et al., 2024). Additionally, LLM planners (Liu et al., 2024) generate parallel solutions before validation, while simulated medical research meetings (Swanson et al., 2024) synthesize discussions into optimal decisions.

Feedback-Based: adjusts reasoning by integrating feedback to refine. **External Feedback** enhances AI agents by incorporating real-time data, expert input, and structured resources, enabling agents to refine their understanding through interactions and external tools (Chen et al., 2024a; Johri et al., 2023). Medical consultation systems iteratively update diagnoses through patient interactions, while decision-making agents query external resources like Phenomizer and DrugBank for real-time clinical knowledge (Li et al., 2024d). **Self Feedback** enables AI agents to refine reasoning internally by evaluating logic, correcting inconsistencies, and iteratively improving outputs (Louie et al., 2024; Yu et al., 2024). Reflection-based techniques such as Reflection CoT and self-play mechanisms further enhance AI models by structuring error analysis and collaborative discussions (Schmidgall et al., 2024). Applications extend to code generation, drug discovery, medical research, and medical exam question generation (Wang et al., 2024e).

B Applications

B.1 Simulating Specific Scenarios

Clinical Workflow Simulation employs multi-agent to model patient care, from consultation to diagnosis. Some works simulate the full consultation workflow, where patient, doctor, and evaluator agents interact. Liu et al. (2025) segmented consultations into four stages and identifies the weakest stage as the limiting factor, akin to Liebig's law. Johri et al. (2023) proposed CRAFT-MD, using doctor agents interacting with structured patient agents and an automatic grading system. Li et al. (2024d) developed MEDIQ, integrating abstention strategies, rationale generation, and self-consistency to refine diagnosis. Fan et al. (2024) introduced AI Hospital, where doctor agents engage in multi-round discussions, mediated by a Central Agent to resolve disagreements. Schmidgall et al. (2024) presented AgentClinic, a multimodal benchmark incorporating cognitive biases and incomplete information to evaluate LLM-based doctor agents. Another direction expands simulations beyond consultation to the entire patient journey.

Bao et al. (2024) developed PIORS, an outpatient reception system using a Service Flow-aware Medical Scenario Simulation framework to enhance department recommendations. Li et al. (2024c) proposed Agent Hospital, a fully autonomous system covering disease onset to recovery. Its MedAgent-Zero framework enables doctor agents to refine their diagnostic accuracy via case-based learning and RAG, mirroring real-world physicians' iterative knowledge refinement and boosting medical evaluation performance. Given the communication-centric nature of mental health care, a large body of work also focuses on **Psychological Counseling and Mental Health Interaction**, which can be viewed as a specialized form here. Examples include Roleplay-doh (Louie et al., 2024), which turns expert feedback into behavior rules; PATIENT-Ψ (Wang et al., 2024c), which incorporates CBT principles; and Chen et al. (2023), which aligns interactions with DSM-5 criteria.

Multi-Disciplinary Medical Team Simulation replicates real-world medical teams' collaborative processes, optimizing communication, information sharing, and decision-making for complex clinical scenarios. For rare disease, Chen et al. (2024b) introduced RareAgents, where a patient agent presents symptoms, an attending physician agent assembles an MDT, and specialists iteratively refine diagnoses using dynamic memory and medical toolkits. Similarly, Kim et al. (2024) proposed MDAgents, employing a hierarchical collaboration strategy where a single doctor handles simple cases, while MDTs, moderated by an external knowledge-integrating agent, address complex ones. Tang et al. (2023) introduced MEDAGENTS, structuring MDT collaboration into four phases—expert recruitment, independent analysis, collaborative consultation, and final decision-making—to enhance reasoning without training. In EHR modeling, Wang et al. (2024f) proposed ColaCare, where DoctorAgent processes structured EHR data with medical guidelines, while MetaAgent refines clinical decisions through iterative assessments, improving predictive modeling by integrating numerical predictions with textual reasoning.

Simulated Patients for Medical Education improve student training in communication, clinical reasoning, and diagnosis within a controlled setting. Advances in LLM-driven simulations enhance fidelity and interactivity. Du et al. (2024) introduced EvoPatient, a multi-agent framework where doctor-patient agents iteratively co-evolve using RAG and personality traits. Wei et al. (2024a) proposed MEDCO, integrating structured training, interdisciplinary collaboration, and multimodal inputs with memory and peer discussion modules. For assessment, Mehandru et al. (2024) proposed AI-SCE for process-focused training, while Yao et al. (2024b) introduced MedQA-CS with simulated student interactions and structured evaluations.

Other Medical Process Optimization and Cross-Disciplinary Simulation AI-driven methodologies have been explored for optimizing medical processes and enabling cross-disciplinary simulations. Swanson et al. (2024) introduced a multi-agent "Virtual Lab," where LLM-powered agents (e.g., principal investigator, biologist, scientific critic) collaborate using biomedical tools like ESM and AlphaFold-Multimer to design nanobody treatments for SARS-CoV-2 variants, showcasing AI's potential in accelerating interdisciplinary research. Similarly, Williams et al. (2023) proposed a generative AI-enhanced epidemic modeling platform, where LLM-driven agents autonomously assess health status and public health data to simulate pandemic dynamics, improving traditional agent-based modeling. These works demonstrate AI's role in advancing scientific discovery and public health modeling through intelligent agent-based decision-making.

B.2 Solving Complex Tasks

Many AI Hospital works leverage multi-agent frameworks to enhance diagnosis, triage, research, and discovery in dynamic clinical settings.

Clinical Decision-Making: AI hospitals improve diagnostic accuracy and transparency, especially for rare or complex diseases. Systems like RareAgents (Chen et al., 2024b), MMedAgent (Li et al., 2024a), and DrHouse (Yang et al., 2024) integrate tools, memory, and retrieval for consistent, multimodal reasoning. Others focus on interpretability: DiagnosisGPT (Chen et al., 2024a), ArgMed-Agents (Hong et al., 2024), and MedAgents (Tang et al., 2023) use structured reasoning or argumentation to reduce bias and enhance trust.

Triage and Clinical Trials: Agent-based systems like TriageAgent (Lu et al., 2024), PIORS (Bao et al., 2024), and ClinicalAgent (Yue et al., 2024a) improve emergency triage, outpatient routing, and trial matching using guideline-based retrieval and reasoning strategies.

Knowledge-Intensive Workflows: AI agents support data science tasks such as EHR analysis (Shi

et al., 2024b), code generation (Wang et al., 2024e), fact-checking (Yue et al., 2024b), and question generation (Yao et al., 2024a), streamlining clinical research.

Scientific Discovery: Multi-agent labs like Virtual-Lab (Swanson et al., 2024), CellAgent (Xiao et al., 2024), and DrugAgent (Liu et al., 2024) automate biomedical discovery, integrating reasoning agents with domain tools to accelerate hypothesis generation, molecular analysis, and drug development.

B.3 Evaluating Agents

AI hospital evaluations are shifting from static benchmarks to interactive, multi-agent simulations that capture real-time reasoning, collaboration, and patient engagement (Johri et al., 2023; Schmidgall et al., 2024; Li et al., 2024d). Recent work emphasizes state-aware evaluation, using patient simulators like SAPS (Liao et al., 2024) and role-play settings (Louie et al., 2024; Wang et al., 2024b) to test an agent's adaptability and coherence across turns. Multi-agent frameworks such as AI Hospital (Fan et al., 2024) and ClinicalLab (Yan et al., 2024) assess inter-agent collaboration, dispute resolution, and cross-department knowledge exchange. Multimodal evaluation is also gaining traction: MMedAgent (Li et al., 2024a) combines imaging and text-based reasoning, while others assess tool-assisted clinical calculations (Khandekar et al., 2024). Finally, OSCE-style benchmarks like MedQA-CS (Yao et al., 2024b), OSCEBot (Pereira et al., 2023), and AI-SCE (Mehandru et al., 2024) offer comprehensive, scenario-based evaluations of real-world clinical skills. In parallel, LLM-as-Judge frameworks are gaining traction across both general (Li et al., 2024b; Gu et al., 2024) and clinical NLP (Tu et al., 2025; Arora et al., 2025; Croxford et al., 2025; Chung et al., 2025) settings, enabling scalable evaluation of agent reasoning, conversation quality, and task outcomes.

B.4 Synthesizing Data for Training

Synthetic data generation in AI hospitals supports realistic, privacy-preserving training for medical LLMs. Multi-agent co-evolution frameworks (Du et al., 2024; Li et al., 2024c) simulate diagnostic dialogues, refine agent reasoning, and improve generalization to benchmarks. NoteChat (Wang et al., 2023a) transforms clinical notes into role-played, polished conversations via planning, simulation, and feedback. AMIE (Tu et al., 2025) uses self-play and auto-feedback to enhance history-taking and reasoning. These methods reduce annotation costs while maintaining clinical validity, enabling scalable training for downstream applications.

ID	Paper Title	Venue	Code/Data	Study
1	Exploring the Inquiry-Diagnosis Relationship with Advanced Patient Simulators	arXiv	link	(Liu et al., 2025)
2	Leveraging Large Language Model as Simulated Patients for Clinical Education	arXiv	No	(Li et al., 2024e)
3	A GPT-Powered Chatbot as a Simulated Patient to Practice History Taking	JMIR Med Edu	No	(Holderried et al., 2024)
4	Designing and building OSCEBot ® for virtual OSCE – Performance evaluation	Med Edu Online	No	(Pereira et al., 2023)
5	Roleplay-doh: Enabling domain-experts to create LLM-simulated patients	EMNLP24	link	(Louie et al., 2024)
6	AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow	arXiv	link	(Yu et al., 2024)
7	LLMs Can Simulate Standardized Patients via Agent Coevolution	arXiv	link	(Du et al., 2024)
8	PATIENT-Ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals	EMNLP24	link	(Wang et al., 2024c)
9	Towards a Client-Centered Assessment of LLM Therapists by Client Simulation	arXiv	link	(Wang et al., 2024b)
10	Automatic Interactive Evaluation for LLMs with State Aware Patient Simulator	arXiv	link	(Liao et al., 2024)
11	Guidelines For Rigorous Evaluation of Clinical LLMs For Conversational Reasoning	medRxiv	link	(Johri et al., 2023)
12	RAREAGENTS: Autonomous Multi-disciplinary Team for Rare Disease Diagnosis	arXiv	No	(Chen et al., 2024b)
13	TWIN-GPT: Digital Twins for Clinical Trials via Large Language Model	TOMM24	No	(Wang et al., 2024d)
14	Interactive Agents: Simulating Counselor-Client Psychological Counseling	arXiv	link	(Qiu and Lan, 2024)
15	DrHouse: An LLM-empowered Diagnostic Reasoning System	IMWUT	No	(Yang et al., 2024)
16	MDAgents: Adaptive Collaboration of LLMs for Medical Decision-Making	NeurIPS 2024	link	(Kim et al., 2024)
17	MEDAGENTS: Large Language Models as Collaborators for Medical Reasoning	ACL Findings 2024	link	(Tang et al., 2023)
18	ColaCare: Enhancing EHR Modeling through LLM Multi-Agent Collaboration	arXiv	link	(Wang et al., 2024f)
19	Mitigating Cognitive Biases in Clinical Decision-Making via Multi-Agent LLMs	JMIR	No	(Ke et al., 2024)
20	AgentClinic: A Multimodal Agent Benchmark for Simulated Clinical Environments	arXiv	link	(Schmidgall et al., 2024)
21	TRIAGEAGENT: Multi-Agents for LLM-Based Clinical Triage	EMNLP Findings 2024	link	(Lu et al., 2024)
22	PIORS: Personalized Intelligent Outpatient Reception Using Multi-Agents	arXiv	link	(Bao et al., 2024)
23	Can Large Language Models Replace Data Scientists in Clinical Research?	arXiv	No	(Wang et al., 2024e)
24	The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies	BioRxiv	link	(Swanson et al., 2024)
25	MEDCO: Medical Education Copilots Using Multi-Agent Framework	arXiv	No	(Wei et al., 2024a)
26	Should We Be Going MAD? Multi-Agent Debate Strategies for LLMs	ICML2024	link	(Smit et al., 2023)
27	Beyond Direct Diagnosis: Multi-Specialist Agent Consultation for Diagnosis	arXiv	No	(Wang et al., 2024a)
28	ClinicalAgent: Clinical Trial Multi-Agent System with LLM Reasoning	BCB '24	link	(Yue et al., 2024a)
29	Enhancing Clinical Trial Patient Matching via Multi-Agent Knowledge Augmentation	arXiv	No	(Shi et al., 2024a)
30	ArgMed-Agents: Explainable Clinical Decision Reasoning via Argumentation	BIBM2024	No	(Hong et al., 2024)
31	Synergistic Multi-Agent Framework with Trajectory Learning	AAAI25	link	(Yue et al., 2024b)
32	Empowering Biomedical Discovery with AI Agents	Cell	No	(Gao et al., 2024)
33	MedDM: LLM-executable Clinical Guidance Tree for Decision-Making	arXiv	No	(Li et al., 2023a)
34	Text2MDT: Extracting Medical Decision Trees from Texts	arXiv	No	(Zhu et al., 2024)
35	BioKGBench: A Knowledge Graph Benchmark	arXiv	link	(Lin et al., 2024)
36	Medical Graph RAG: Safe LLMs via Graph Retrieval-Augmented Generation	arXiv	link	(Wu et al., 2024)
37	HeCiX: Integrating Knowledge Graphs and LLMs for Biomedical Research	arXiv	No	(Kulkarni et al., 2024)
38	KRAGEN: Knowledge Graph-Enhanced RAG for Biomedical Problem Solving	Bioinformatics	link	(Matsumoto et al., 2024)
39	Accelerating Medical Knowledge Discovery via Automated Knowledge Graphs	KGSWC 2024	No	(Khalid et al., 2024)
40	Augmented Non-Hallucinating LLMs as Medical Information Curators	npj Digital Medicine	No	(Gilbert et al., 2024)
41	Benchmarking Retrieval-Augmented Generation for Medicine	ACL Findings 2024	link	(Xiong et al., 2024a)
42	Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions	arXiv	link	(Xiong et al., 2024b)
43	Almanac — Retrieval-Augmented Language Models for Clinical Medicine	NEJM AI	link	(Zakka et al., 2024)
44	Augmenting Black-box LLMs with Medical Textbooks for Biomedical QA	EMNLP Findings 2024	link	(Wang et al., 2023b)
45	To Generate or Retrieve? Effectiveness of Artificial Contexts in Medical QA	ACL 2024	link	(Frisoni et al., 2024)
46	AgentMD: Empowering Language Agents for Risk Prediction	arXiv	link	(Jin et al., 2024)
47	MedCalc-Bench: Evaluating LLMs for Medical Calculations	NeurIPS 2024	link	(Khandekar et al., 2024)
48	Augmenting ChatGPT with Clinician-Informed Tools for Medical Calculations	medRxiv	No	(Goodell et al., 2023)
49	GeneGPT: Augmenting LLMs with Domain Tools for Biomedical Information	Bioinformatics	link	(Jin et al., 2023)
50	EHRAgent: Code-Empowered LLMs for Few-shot Complex Tabular Reasoning	EMNLP 2024	link	(Shi et al., 2024b)
51	MMedAgent: Learning to Use Medical Tools with Multi-modal Agent	EMNLP Findings 2024	link	(Li et al., 2024a)
52	Conversational Health Agents: A Personalized LLM-Powered Agent Framework	arXiv	link	(Abbasian et al., 2023)
53	PathAsst: A Generative AI Assistant for Pathology Analysis	AAAI Technical Track	link	(Sun et al., 2023)
54	GPT-agents Based on Medical Guidelines for Traumatic Brain Injury Rehabilitation	Scientific Reports	No	(Li et al., 2024f)
55	CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis	arXiv	No	(Xiao et al., 2024)
56	DrugAgent: Automating AI-Aided Drug Discovery via LLM Multi-Agent Collaboration	arXiv	No	(Liu et al., 2024)
57	Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents	arXiv	No	(Li et al., 2024c)
58	AI Hospital: Benchmarking LLMs in a Multi-agent Medical Interaction Simulator	COLING 2025	link	(Fan et al., 2024)
59	ClinicalLab: Aligning Agents for Multi-Departmental Clinical Diagnostics	arXiv	link	(Yan et al., 2024)
60	LLM-empowered Chatbots for Psychiatrist and Patient Simulation	arXiv	No	(Chen et al., 2023)
61	MediQ: Question-Asking LLMs and a Benchmark for Interactive Clinical Reasoning	NeurIPS 2024	link	(Li et al., 2024d)
62	Epidemic Modeling with Generative Agents	arXiv	link	(Williams et al., 2023)
63	NoteChat: A Dataset of Synthetic Patient-Physician Conversations	ACL 2024 Findings	link	(Wang et al., 2023a)
64	Evaluating Large Language Models as Agents in the Clinic	npj Digital Medicine	No	(Mehandru et al., 2024)
65	MedQA-CS: Benchmarking LLMs Clinical Skills Using an AI-SCE Framework	arXiv	link	(Yao et al., 2024b)
66	Towards Conversational Diagnostic AI	arXiv	No	(Tu et al., 2025)
67	LLM-based Medical Assistant Personalization with Short- and Long-Term Memory	NAACL 2024	link	(Zhang et al., 2023)
68	CoD: Towards an Interpretable Medical Agent Using Chain of Diagnosis	arXiv	link	(Chen et al., 2024a)
69	Multi-Agent Conversation Framework Enhances Rare Disease Diagnosis in LLMs	Preprint	link	(Li et al., 2023b)
70	MEDAIDE: Towards an Omni Medical Aide via Specialized LLM-based Multi-Agent Collaboration	Preprint	No	(Wei et al., 2024b)
71	RAG-Gym: Optimizing Reasoning and Search Agents with Process Supervision	Preprint	link	(Xiong et al., 2025)
72	MCQG-SRefine: Multiple Choice Question Generation and Evaluation with Iterative Self-Critique, Correction, and Comparison Feedback	NAACL 2025	link	(Yao et al., 2024a)