Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CareLoop: Measuring the Gap Between Knowing Medicine and Practicing It in the Context of Patients' Lives

Zhenhong Yang
JDH Algo, JD Health International Inc., China
Jintao Fei
JDH Algo, JD Health International Inc., China
Jun Zhao
zhaojun10@jd.com

JDH Algo, JD Health International Inc., China

Research Article

Keywords: large language models, medical AI, AI agents, clinical simulation, benchmark, patient-centered care, trajectory-based evaluation

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11149049/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

CareLoop: Measuring the Gap Between Knowing Medicine and Practicing It in the Context of Patients' Lives

Zhenhong Yang **Jintao Fei** **Jun Zhao**

JDH Algo, JD Health International Inc., China

Correspondence to Jun Zhao: zhaojun10@jd.com

Abstract

JDH 京东健康

Medical benchmarks usually ask whether an AI system knows the right answer. Patient-facing care additionally requires uncovering incomplete or incorrect information, understanding a person's beliefs and constraints, adapting plans to what is feasible, verifying execution, incorporating delayed evidence, and retaining responsibility for unresolved risk. We introduce CareLoop, a clinical-world sandbox for measuring the gap between knowing medicine and practicing it in the context of patients' lives. Its 120 deidentified EHR-derived case contracts distribute state across underlying clinical state, patient and family beliefs, records, and executed actions; frictions authored before model execution create controlled behavioral forks with observable downstream consequences.

Ten models generated 1,200 trajectories. Four blinded LLM Judges produced 4,800 complete-context assessments, and 139 physicians produced 6,000 structured assessments. GPT-5.6 Sol ranked first on the Friction Capability Composite (FCC, 0.824) and its safety-gated counterpart C-RWR (0.753). Across 15,537 assessments of realized high-order opportunities, models acted in 0.935, changed the trajectory meaningfully in 0.745, and completed the opportunity effectively in 0.697; hidden-state discovery was the largest shared bottleneck. Individual assessments varied, yet disjoint 60-case halves recovered stable model rankings (median Spearman $\rho = 0.915$ for FCC and 0.927 for C-RWR), and aggregate physician and four-Judge rankings converged ($\rho = 0.988$ and 0.964). Matched model–Judge analyses showed no uniform self-favoring pattern.

These findings show how execution-grounded evaluation can extend from directly testable state change to consequential action under partial observability and limited control. Correct answers capture one component of medical intelligence. Patient-facing practice additionally requires verifying consequences, adapting to changing evidence, and keeping unresolved obligations accountable across a trajectory. Data and code: https://github.com/jdh-algo/CareLoop. The simulation does not establish clinical effectiveness or deployment safety.

1 Introduction

Medicine is practiced with people whose beliefs, families, resources, choices, and changing circumstances shape care. A model may know the medically appropriate answer and still fail in the context of a patient's life: a symptom remains unspoken, a caregiver misstates medication use, a report is unreadable, a result arrives after the original plan, or cost and transport block the next step. Care may also stall when nobody verifies execution, incorporates returned evidence, or retains ownership of unresolved risk. These failures define a gap between knowing medicine and practicing it with a person situated in biological, social, and healthcare realities.

Most medical benchmarks isolate important components of competence, including knowledge, diagnostic reasoning, communication, safety, tool use, and execution of prescribed workflows. Recent agent environments add sustained interaction, stateful tools, electronic health record (EHR) operations, simulated patients, and

multi-agent hospitals. General-agent benchmarks further ground success in verifiable changes to external worlds. CareLoop asks how this principle changes when the world is only partly observable and controllable: patients have beliefs, relationships, practical constraints, and the autonomy to misunderstand, hesitate, refuse, or change course. The resulting evaluation problem centers on agency under limited control. Task-relevant state is distributed across people and systems, while evidence, feasible paths, and appropriate goals may change during the episode.

We introduce CareLoop, a clinical-world sandbox for measuring the gap between knowing medicine and practicing it in the context of patients' lives. Its primary artifact is a trajectory generated inside a partially observable environment. Clinical state is distributed across underlying clinical state, patient and family beliefs, recorded evidence, and executed actions. Patients and family members are stateful actors who may disclose selectively, misunderstand instructions, refuse care, or encounter feasibility constraints. Diagnostic and administrative events can be delayed. External clinical actions may succeed, remain pending, or fail. These mechanisms create controlled behavioral forks at which a model must ask, verify, act, wait, update, repair, or preserve an accountable open state.

CareLoop uses these forks as diagnostic probes. Further progress at each probe depends on a particular capability, so latent limitations acquire visible consequences across the trajectory. Missing hidden state prevents appropriate planning; uncorrected misinformation corrupts subsequent reasoning; an untracked external action leaves the world unchanged; ignored evidence preserves an obsolete plan; and unresolved risk can culminate in premature closure. Fluent language, frequent tool use, and terminal success flags can all coexist with a trajectory that never achieves meaningful clinical progress.

The framework separates four concepts. Ten friction strata authored before model execution describe what obstructs execution. Eight high-order (HO) opportunity types describe the response that a case may require. Six continuous domains summarize information repair, constraint navigation, dynamic reprioritization, execution-loop repair, state-sensitive communication, and bounded closure/continuity. FCC aggregates these applicable domains, and C-RWR applies a transparent clinical-error gate. Case-specific minimum duties and responsibility requirements constrain safety throughout the episode and serve as guardrails around friction-sensitive progression.

We evaluate 10 models on the same 120 frozen, deidentified EHR-derived case contracts. The experiment produced 1,200 trajectories, 4,800 complete-context LLM-Judge assessments, and a separate blinded comparison comprising 6,000 structured physician assessments. The study asks five questions:

1. Can a clinical-world sandbox make distributed state, evolving evidence, human friction, and closure semantics explicit and auditable?
2. Where do tested models differ: information discovery and repair, constraint adaptation, longitudinal execution, plan revision, safety, or closure?
3. When a model acts, does that action produce meaningful trajectory impact?
4. Do operational termination and safe completion describe the same outcome?
5. Can repeated structured evaluation over broad case coverage recover stable model comparisons despite local evaluator disagreement?

Action frequency captured only part of model performance. Models usually acted when an HO was realized, yet meaningful impact and effective completion were substantially lower, especially for hidden-state discovery and bounded closure. The greatest model separation appeared in information repair, dynamic reprioritization, and bounded continuity. Operational termination frequently diverged from clinical closure. Rankings nevertheless stabilized with broader case coverage, and independently aggregated physician and LLM-Judge rankings converged over the complete benchmark. These findings support the paper's central thesis: clinical-agent evaluation should examine how medical knowledge is exercised with a person whose world can reshape every subsequent decision.

The contributions are:

- **A patient-centered task ontology.** CareLoop formulates patient-facing care as agency under limited control in a partially observable socio-technical world, with distributed state, independent actors, changing

evidence, and accountable openness.

- **A capability-eliciting environment.** Frictions and behavioral forks authored before model execution make latent deficiencies produce auditable consequences across complete trajectories.
- **A trajectory-centered measurement system.** Six capability domains, FCC, safety-gated C-RWR, and an action-to-impact-to-completion decomposition distinguish attempted action, realized consequence, repair, and continuity.
- **A large-scale empirical study and reproducible release.** Across 120 cases, 1,200 trajectories, four LLM Judges, and 139 physicians, the analysis localizes model differences and demonstrates benchmark-level ranking stability despite local evaluator disagreement. The released artifacts support independent reproduction.

The experiments concern behavior inside a simulated clinical world. Clinical effectiveness, regulatory readiness, and autonomous deployment safety remain outside their scope.

2 Related Work: From Task Completion to Care Progression

2.1 Medical Knowledge, Dialogue, and Clinical-Agent Environments

Medical evaluation has progressed from examination-style knowledge and reasoning benchmarks such as MedQA, PubMedQA, MedMCQA, and medical subsets of MMLU toward open-ended assessments of helpfulness, safety, communication, and information gathering [1, 2, 3, 4, 5, 6, 7]. These benchmarks remain essential for establishing whether a model knows relevant medicine and can express a clinically appropriate response.

Interactive medical benchmarks evaluate additional capabilities. PatientAgentBench places a patient-facing assistant in sustained conversations with simulated patients and a stateful healthcare-service tool sandbox [8]. PhysicianBench evaluates clinician-facing agents on long-horizon EHR workflows with execution-verifiable checkpoints [9]. HealthAgentBench broadens executable healthcare-agent testing across 54 tasks and seven workflow categories, emphasizing exploration of raw health data and multi-step task completion [10]. Agent Hospital, AI Hospital, AgentClinic, 3MDBench, and related simulations study diagnosis, treatment, multimodal observation, institutional interaction, or agent evolution in executable medical settings [11, 12, 13, 14, 15]. These systems establish that realistic medical-agent evaluation requires interaction, state, tools, and multiple actors.

CareLoop addresses a complementary object. It systematically authors hidden, unreliable, delayed, and difficult-to-execute state, then follows whether a patient-facing model discovers that state, repairs misinformation, adapts the feasible plan, tracks external execution, incorporates returned evidence, and maintains safe continuity. Its main unit is progression through a friction-bearing care episode. Responsibility and minimum-duty requirements remain important safety constraints, while the central measured construct is the response to friction across the trajectory.

2.2 From Answers to Executable Agent Environments

General-agent benchmarks clarify why a trajectory cannot be reduced to the plausibility of its final language. AgentBench broadened evaluation across heterogeneous interactive environments, while GAIA tested general assistants on real-world questions requiring tools, multimodal evidence, and reasoning [16, 17]. GAIA remains primarily answer-oriented; subsequent environments make externally verifiable state change the central outcome.

WebArena supplies functional, self-hosted websites and evaluates long-horizon goals against resulting web state [18]. WorkArena transfers this principle to common enterprise workflows in ServiceNow [19]. OSWorld evaluates multimodal agents through visual observations and mouse/keyboard actions in real computer environments [20]. SWE-bench grounds software-engineering tasks in real repositories and uses executable tests to determine whether a proposed patch changes the system correctly [21]. AppWorld builds a controllable world of interconnected apps, APIs, data, and people in which generated code must produce task-specific state changes [22].

Table 1 Architectural comparison of representative interactive agent environments. The table characterizes design choices across systems; row order carries no overall quality ranking.

Work	Fundamental task unit	Persistent world state	Who can change state	Main diagnostic pressure	Success semantics	Boundary relevant to CareLoop
GAIA	Real-world question	Limited	Assistant through tools	Search, reasoning, multimodality	Correct final answer	Broad assistant context; not primarily a shared persistent world.
WebArena / WorkArena	Web or enterprise workflow	Yes	Agent	Navigation, grounding, planning, execution	Programmatic final-state checks	User generally supplies a fixed goal; autonomous goal evolution is limited.
OSWorld	Cross-application computer task	Yes	Agent	Visual grounding and low-level control	Task-specific OS/app evaluators	Deep interface realism; less emphasis on actor belief and evolving goals.
SWE-bench	Repository issue and patch	Yes	Agent	Long-context code reasoning and implementation	Executable tests	Strong external verification with a comparatively determinate oracle.
AppWorld	Cross-app API task	Yes	Agent in a world containing apps and people	Planning, code generation, API composition	State-based task tests	Demonstrates that complex worlds and people are already present in general benchmarks.
τ-bench	Tool–agent–user service episode	Yes	Mainly agent; user communicates goals and preferences	Policy following, database reasoning, user interaction	Final database goal state and policy criteria	Fixed service end state; limited direct user world action.
τ^2-bench	Dual-control collaborative episode	Yes	Agent and user	Coordination, instruction, verification, shared-world action	Global reward over world state/history	Closest shared-control analogue; clinical state and closure remain distinct.
TheAgent-Company	Consequential workplace task	Yes	Agent and organizational actors	Cross-tool work and coworker interaction	Task-specific workplace checks	Organizational actors exist, but clinical risk and evidence-driven goal revision are absent.
CRMArena/Pro	Professional CRM task	Yes	Agent within enterprise records and policies	Multi-object reasoning, rules, function calling	CRM state and task constraints	Human-service analogue without evolving disease state or bounded-open care.
PatientAgent-Bench	Patient-facing health-service episode	Yes	Assistant, simulated patient, service tools	Conversation, safety, and service execution	Multi-dimensional jury evaluation	Closest medical patient-facing environment; friction is less central as an authored experimental variable.
Physician-Bench	Clinician EHR workflow	Yes	Clinician-facing agent	Retrieval, reasoning, consequential EHR action	Execution-verifiable checkpoints	Strong institutional execution; different actor role and closure target.
HealthAgent-Bench	Multi-workflow healthcare task	Yes	Agent across task-specific environments	Raw-data exploration and multi-step execution	Final task success across workflow categories	Broad healthcare workflow coverage; the central task unit gives less emphasis to patient belief, evolving feasibility, and bounded continuity.
Agent Hospital family	Simulated institutional care	Yes	Multiple medical actors	Diagnosis, treatment, interaction, evolution	Simulation-specific clinical outcomes	Institution-scale simulation with broad clinical processes; cases lack frozen friction-response contracts.

These environments establish the value of execution-grounded evaluation: task success requires evidence that the intended state change occurred. CareLoop adopts this perspective. Clinical progress rarely offers a single immediate unit-test oracle. Evidence may return later, patient or family reports may be unreliable, an externally licensed action may remain pending, and maintaining a bounded open episode can be safer than forcing a terminal state.

2.3 Users, Coworkers, and Shared-World Control

Human-facing agent evaluation further changes who can determine the task state. In τ-bench, an agent interacts with a simulated user, domain policies, databases, and APIs; success is evaluated against the final database state, and pass^k measures consistency across repeated trials [23]. τ^2-bench extends this setting to dual control: both the agent and the simulated user can take tool actions over a shared environment, exposing coordination and verification failures that remain hidden when the user only supplies language [24]. TheAgent-Company places agents in an integrated software-company environment with coworkers and consequential cross-application tasks [25]. CRMArena and CRMArena-Pro evaluate professional customer-relationship-management tasks over interconnected records, latent variables, policies, and multiple organizational roles [26, 27].

These works preclude a broad novelty claim based solely on statefulness, other actors, or user agency. They also provide the closest general-agent foundation for CareLoop. The distinction lies in the task ontology. In CareLoop, clinically relevant state may differ across underlying clinical state, patient or caregiver belief, recorded evidence, and actual execution. New evidence can change the appropriate goal. Human friction can affect both feasibility and disclosure. Completion must account for residual risk, ownership, and a legitimate need to remain open.

2.4 Care Progression Under Real-World Friction

The progression above moves evaluation from answers to actions, from actions to executable state, and from single-agent control to shared worlds. CareLoop retains the central lesson of executable benchmarks: persuasive language is insufficient unless an agent's behavior produces evidence of change. It then asks how that principle changes when the evaluated world is partly observable, partly controlled by independent people, and unable to provide one immediate or complete oracle.

In formalizable executable worlds, task-relevant state can often be modified directly and checked against comparatively determinate external criteria. SWE-bench, for example, grounds a repository issue in a patch and executable tests; web, application, service, and organizational environments broaden the state, actors, and policies involved. CareLoop represents a different evaluation regime rather than a simple increase in domain difficulty. Its relevant state is distributed across underlying clinical state, patient and family belief, recorded evidence, and actual execution. The agent has intervention capacity without full control; new evidence can revise the appropriate goal; and an action can be correct without being followed, understood, completed, or consequential.

A case therefore makes real-world friction a first-class experimental variable and creates interpretable pressure points along an unfolding trajectory. The evaluator follows whether the model acquires missing state, judges information reliability, coordinates with independent actors, produces an action, verifies its consequence, updates the plan, repairs failure, and reaches safe closure or an accountable open state. This integrated formulation shifts the endpoint from nominal task completion toward trajectory-based consequential evaluation in a socio-technical world.

3 The CareLoop Framework

3.1 Problem Formulation: Care Progression in a Distributed World

CareLoop treats a clinical episode as a partially observable, multi-actor process. At time t, four related states may diverge: underlying clinical state S_t^{truth}, what patients or family members believe S_t^{belief}, what has been recorded or returned as evidence S_t^{record}, and what has actually been executed in the world S_t^{exec}. The model observes only a governed projection of these states through dialogue and the clinical workspace. Its actions can request information, communicate, use tools, or initiate external processes; patient, family, and system actors then determine what becomes visible or executed.

A *friction* is a mechanism that creates or preserves a consequential mismatch among these states, or obstructs the transition from a clinically appropriate intention to an executed result. Examples include a hidden symptom, an incorrect medication belief, an unreadable report, delayed pathology, refusal, cost, transport, or a broken handoff. *Care progression* occurs when the model obtains relevant state, chooses a safe and feasible action, receives or seeks evidence of its consequence, revises the plan when necessary, and constrains residual risk.

Safe closure is therefore episode-specific. It indicates that urgent risk is addressed and remaining obligations have an owner, timing, verification path, and escalation condition. It does not require permanent resolution of every disease problem. A bounded open state can be appropriate when evidence or external action is legitimately pending.

3.2 From Latent Limitation to Observable Failure

The environment is designed as a capability-elicitation instrument. Each case creates controlled pressure points at which continued care depends on a particular behavior. Hidden state calls for elicitation or retrieval. Patient or family misunderstanding calls for verification and repair. External action calls for evidence of execution and continued tracking. Delayed evidence calls for memory and plan revision. Unresolved risk calls for accountable openness or an explicit handoff.

This design turns latent limitations into auditable consequences. A model can state the correct general principle yet fail to discover the fact that makes it relevant. It can recommend a test without establishing whether the test occurred. It can recognize a barrier but offer no feasible alternative. It can act actively

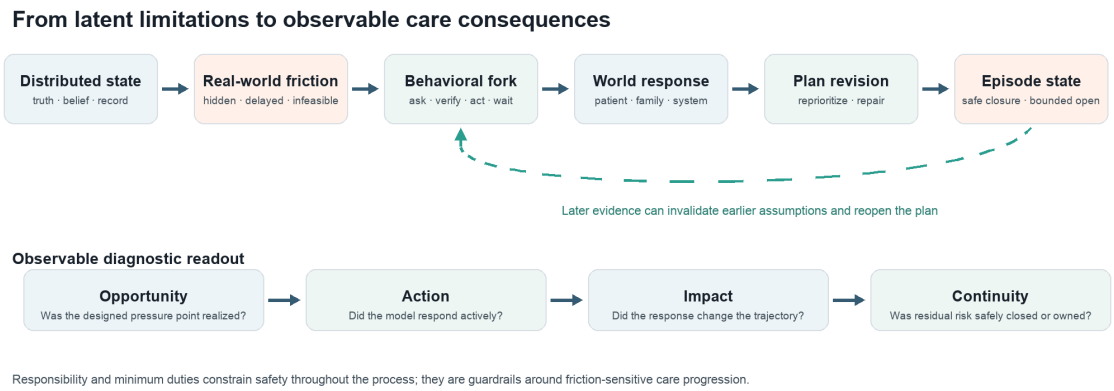

Figure 1 CareLoop task ontology and diagnostic readout. Distributed state and real-world friction create behavioral forks at which the tested model must obtain information, verify, act, wait, or revise. The world then supplies a response that can support, obstruct, or invalidate the plan. Structured evaluation separates whether an opportunity appeared, whether the model acted, whether that action changed the trajectory, and whether residual risk was safely closed or retained under accountable continuity. Responsibility and minimum duties constrain safety throughout the process.

while leaving the world unchanged. It can terminate the conversation while an important obligation remains unowned. These failures become visible because the world continues after the initial answer and because case-authored opportunities identify where the relevant capability was expected to matter.

The design avoids assuming that longer dialogue, more tool calls, or stronger fluency imply better care. The diagnostic chain is instead: authored opportunity, realized pressure point, active response, meaningful trajectory impact, and safe continuity. The same chain also permits a strong trajectory to remain open when further evidence is genuinely pending.

3.3 Case Contracts and Multi-Actor State

Each case is a hidden-state clinical-world contract. It specifies the opening presentation, underlying clinical state, actor-specific knowledge and behavior, workspace evidence, visibility and reliability rules, delayed events, friction lifecycle, external actions, minimum duties, high-order opportunities, false-closure traps, and bounded closure criteria. This contract converts an EHR-derived scenario into a runnable world while retaining an auditable account of why each pressure point exists.

The patient and family actors are independent participants with their own evolving state. They can hold incomplete or incorrect beliefs, disclose selectively, express preferences, misunderstand recommendations, encounter constraints, or act outside the conversation. The clinical workspace forms a separate evidence channel containing records, receipts, pending tasks, and returned results. Licensed medical actions can remain external to the tested model; the evaluation asks whether the model notices what must happen, establishes a feasible route, interprets the resulting evidence, and maintains continuity around the external action.

CareLoop preserves many valid clinical paths. Case contracts specify what must be discovered, protected, or resolved without prescribing every acceptable sentence or action sequence. HO opportunities describe clinically meaningful responses, while minimum duties and closure rules prevent adaptive behavior from abandoning essential safety obligations.

3.4 Environment Mechanisms and Diagnostic Targets

Table 2 links core mechanisms to the limitation they are designed to reveal. This mapping is central to the benchmark: a system component is scientifically useful when it creates an interpretable behavioral fork; added simulation complexity alone is insufficient.

Table 2 How CareLoop mechanisms expose trajectory-level capability limitations.

Environment mechanism	Behavior required for further progress	Limitation made observable
Truth separated from visible state	Ask, retrieve, or wait for decisive information	Hidden-state discovery and information-acquisition failure
Actor beliefs can be incorrect	Compare sources, identify conflict, and verify	Uncritical acceptance of misinformation
Patient and family hold different state	Coordinate disclosure and reconcile accounts	Multi-source state integration failure
Delayed tests, reports, or referrals	Preserve ownership and reassess after return	Weak longitudinal memory and follow-through
New evidence changes risk	Reprioritize and revise the plan	Static planning and anchoring
Cost, transport, work, or access constraints	Construct an executable alternative while retaining safety	Idealized advice without feasibility repair
Refusal, low adherence, stigma, or privacy concerns	Adapt communication and negotiate a bounded plan	State-insensitive communication
External clinical action and receipts	Distinguish ordering, execution, and verified result	Treating recommendation or tool invocation as completion
Failed handoff or execution loop	Detect the break and restore ownership	Weak execution-loop repair
False-closure traps and bounded-open states	Inspect unresolved risk before stopping	Premature closure and an incorrect completion model

3.5 Runtime Components

A CareLoop run includes four interacting components, and the resulting trajectory stores both dialogue and workspace-side events. The **doctor model** produces patient-facing medical guidance. The **patient/-family actor** returns incomplete, emotional, sometimes mistaken, and context-dependent messages. The **care workspace** provides records, results, and system events visible to the doctor model. The **world progression layer** simulates clinical events, external clinician actions, delayed results, medication changes, and real-world frictions. Together these components generate a multi-turn trajectory.

The patient/family actor is intentionally not a passive oracle. It may forget instructions, confuse medication names, resist emergency care, hide stigmatizing details, or have limited caregiver availability. The workspace is also not a static chart dump: new results, documents, simulated artifact-reliability states, and external clinical actions can appear over time. The world progression layer means that the model's previous advice can change the future context: a safe escalation can move the patient into emergency care, while vague advice can leave a family member uncertain about medication execution. This makes the simulation sensitive to whether the doctor model can infer and update the clinical state under realistic information flow.

The external-clinician mechanism is important for CareLoop's evaluation philosophy. CareLoop evaluates the simulated doctor role on language-mediated contributions it can plausibly support: recognizing risk, explaining, planning, coordinating, escalating, handing off, and maintaining continuity. The framework keeps broader healthcare-system authority and procedural execution outside the model role. When a patient needs emergency evaluation, imaging, surgery, pharmacy reconciliation, inpatient care, or specialist management, the simulated world can place those tasks under accountable external clinicians. The model is then judged by whether it recognized the boundary of its control, initiated the transition, tracked the new state, and explained the handoff appropriately.

Stochastic event governance. CareLoop also supports clinically anchored stochastic events, but these events are constrained by clinical and causal eligibility. At each turn, the world layer may propose contingency candidates such as medication adverse effects, treatment complications, disease worsening, external-care failures, patient nonadherence, environmental barriers, or mortality risk. A candidate must have a causal anchor in the case state, disease course, treatment exposure, patient behavior, doctor action, or care-system context; otherwise the appropriate result is no new event. When a candidate is eligible, the simulator estimates an event probability and the runtime resolves it through a deterministic seeded probability kernel. Only events

that pass this probability resolution become committed world events and can enter the patient/family actor's lived reality; non-occurring candidates are retained only as audit traces. This design allows realistic contingency, including rare high-impact outcomes, while preventing the simulator from arbitrarily extending the story or exposing hidden scoring intent.

Terminal outcomes are adjudicated separately from stochastic occurrence. A trajectory may end through safe episode closure, bounded task completion, safe handoff, durable longitudinal management, cure, death, adverse-event terminality, loss to follow-up, or open progression. These terminal categories describe how the simulated clinical unit ended and remain separate from medical-quality scoring. For example, death or external takeover may terminate the episode without implying good model performance, and cure or stable management requires support from the trajectory's responsibility chain; a random favorable outcome is insufficient.

3.6 Operationalizing Real-World Clinical Dynamics

The runtime layer is where the main CareLoop design becomes testable. It is not limited to a small list of scripted obstacles. The current runtime and prompt stack represents multiple classes of clinical-world dynamics: actor disclosure and execution noise, workspace evidence release, diagnostic-service timing, document reliability, external-care handoffs, longitudinal memory, tempo control, residual-risk governance, and visibility/coherence audits. These mechanisms should be read as *simulation channels*: they make clinically relevant uncertainty and real-world friction observable in a trajectory. They should be read as available channels; a specific case activates only the mechanisms required by its authored clinical world, and the channels are not deterministic clinical pathway rules.

Table 3 expands the runtime mechanisms and the behaviors they expose.

Table 3 Runtime mechanisms used to project real-world care dynamics into CareLoop trajectories. The first column gives the mechanism name used in the framework; the second column expands the mechanism in plain language; the third column states the behavior or failure mode that becomes observable for trajectory-level evaluation.

Runtime mechanism	Operational meaning in CareLoop	Evaluation signal made visible
Hidden information boundary	The case contains clinically important facts that are not automatically disclosed to the doctor model. They may be unknown to the patient, misunderstood by a caregiver, socially sensitive, available only in records, or scheduled to appear later.	Whether the model actively discovers, verifies, and updates hidden state before treating the first patient message as complete.
Patient/family actor realism	The patient or caregiver is simulated as an independent actor with memory, emotion, limited literacy, preferences, privacy concerns, medication confusion, incomplete recall, and possible nonadherence. The actor maintains independent state and agency throughout the episode.	Whether the model uses teach-back, sensitive questioning, caregiver alignment, adherence checks, and patient-adapted explanations.
Workspace evidence boundary	The workspace is the doctor-side clinical environment: records, orders, returned results, document receipts, follow-up ledgers, and external-care notes. Workspace evidence is separate from patient-visible conversation and must be requested, interpreted, and translated responsibly.	Whether the model seeks the right evidence, notices missing or pending evidence, integrates workspace findings, and explains them safely to the patient.
Diagnostic-service and result timing	Tests, imaging, cultures, pathology, referrals, and reports unfold over simulated time. An order and a completed result are distinct; services may require scheduling, waiting, repeated contact, or external execution.	Whether the model tracks pending diagnostics, interprets returned results, avoids treating ordered-but-unperformed tests as facts, and reacts to delayed abnormalities.

Table 3 continued. Runtime mechanisms in CareLoop.

Runtime mechanism	Operational meaning in CareLoop	Evaluation signal made visible
Document reliability handling	Patient-shared photos, screenshots, drug boxes, discharge papers, and reports can be missing, unreadable, partial, second-hand, patient-read, workspace-confirmed, or staff-verified.	Whether the model distinguishes reported information from reliable evidence and avoids overclaiming from low-quality or unverified artifacts.
Real-world friction lifecycle	Barriers such as cost, access, transport, work, insurance, family availability, stigma, administrative delay, and refusal can appear, persist, resolve, transfer to another actor, or end as residual risk or loss to follow-up.	Whether the model turns barriers into feasible plans with ownership, timing, escalation conditions, and acceptable fallback options that move beyond generic advice.
World progression and contingencies	The clinical world can change after the model acts: symptoms may worsen or improve, complications or adverse effects may occur, external care may succeed or fail, and rare high-impact events such as mortality risk may be considered. Eligible contingency candidates must be clinically anchored and are resolved by a deterministic seeded probability kernel.	Whether the model adapts to new clinical nodes, recognizes deterioration or improvement, corrects earlier assumptions, and maintains coherent responsibility across changing events.
Tempo, receipt, and closure governance	The runtime tracks elapsed time, task execution, returned receipts, stable monitoring, external takeover, residual risks, and terminal criteria. Closure means that the simulated episode can safely stop or transfer; it is not itself the medical-quality score.	Whether pending tasks have owners, safety-net thresholds, verification paths, and realistic stopping or handoff conditions before the trajectory ends.
Visibility and coherence safeguards	Hidden simulator state, judge logic, provider errors, infrastructure failures, and impossible retroactive changes must not leak into the patient-facing clinical world. The runtime also checks that world updates remain consistent over turns.	Whether the final trajectory preserves coherent medical behavior and remains free of prompt leakage, infrastructure artifacts, contradictions, or retconned state.

Note. The listed mechanisms are simulation channels implemented by the released runtime code and prompt stack, including actor-realism control, friction lifecycle planning, diagnostic-service simulation, document-reliability handling, receipt lifecycle summarization, tempo governance, residual-risk governance, closure review, stochastic-event resolution, and visibility/world-coherence audits. A case activates only the channels required by its authored clinical-world contract.

This infrastructure allows a trajectory to fail for clinically realistic reasons. For example, a model can lose quality by assuming that the patient has provided a complete history, by treating a family medication report as reliable without verification, by failing to track a delayed abnormal result, by claiming to read an unclear uploaded document, by letting a receipt remain pending without ownership, or by recommending a follow-up plan that cannot be afforded or performed. A strong trajectory actively discovers uncertain context, corrects misunderstandings, updates the clinical state, assigns responsibility, and turns that updated understanding into feasible next steps.

The same design is intended to reduce superficial rubric gaming compared with a fully visible checklist. The model does not know every hidden fact, false-closure trap, delayed event, artifact-reliability state, or actor-execution state at the start of the run. A response that superficially mentions red flags, follow-up, or shared decision making will not score well if later turns show that the model accepted a false report, missed a delayed result, failed teach-back, overclaimed from an unreadable document, or left an accountable handoff unspecified. The framework is therefore centered on sustained clinical-world consistency; isolated rubric keywords are insufficient.

3.7 Clinical-World Inference and Actionability

A core assumption of CareLoop is that many AI-in-medicine failures arise from world-state inference, with isolated knowledge errors representing only one subtype. The model may know the correct diagnosis but fail to notice that the patient's description is incomplete; it may accept a family medication report that is actually wrong; it may recommend monitoring but fail to connect a later result to the earlier risk state; it may mention a medication adjustment but fail to verify whether any clinician, pharmacist, or caregiver can safely execute it. These gaps are difficult to expose in short-answer benchmarks because the missing context never has time to matter.

CareLoop makes context operational. Patient constraints, family execution, cost, literacy, stigma, delayed evidence, and institutional handoff are treated as part of the task. The desired output extends beyond a correct medical statement. First, the model must infer a better clinical world state from partial observations. Then it must translate that updated understanding into executable care: what should happen, who should do it, when it should happen, what must be monitored, which pending information must be followed, and what condition triggers escalation. This is why the same trajectory can be clinically poor despite containing many true statements, and clinically strong despite remaining open at the end of the observation window.

Figure 2 summarizes the core runtime loop and its downstream analysis artifacts. Routine use of CareLoop requires a case contract, a simulated longitudinal interaction, an immutable trajectory record, terminality adjudication, and a medical-usability judgment. The present experiment uses four independent, model-blinded Judges as its primary evaluation panel; leave-one-Judge-out and Judge-specific analyses quantify sensitivity to evaluator choice.

Operationally, a CareLoop run follows a fixed lifecycle. The case contract is written before any model is evaluated. At each turn, the doctor model receives the currently visible patient message and workspace state, replies with patient-facing guidance, and may ask for information, recommend urgent care, assign follow-up, or hand off to external clinicians. The patient/family actor then responds according to the authored hidden state and behavior profile; the workspace and world layers update records, delayed results, medication changes, external clinician actions, consequences, and clinically anchored stochastic events when probability resolution commits them. A closure monitor checks whether the clinical unit has reached a terminal state, but the raw trajectory remains immutable. After the run ends by terminal closure or by the engineering turn cap, the trajectory is re-audited for earliest terminal turn and separately scored for medical usability using model-blinded evidence packets. This lifecycle is meant to make the framework easy to remember: *hidden world, partial observation, model action, world update, trajectory audit, medical-usability scoring.*

3.8 Terminal Closure

Terminal closure determines whether the case-level clinical unit has naturally reached a terminal state. Terminal outcomes include safe episode closure, bounded-task closure, safe handoff, cure, durable longitudinal management, death, adverse-event terminality, loss to follow-up, and open progression. The goal is to align case-authored terminal targets with runtime closure judgments, especially for bounded, handoff, and chronic-management cases.

A terminal state records a clinically defensible endpoint for the current clinical unit: the acute danger has been transferred to emergency care, the bounded result-follow-up problem has been resolved, the family can execute the medication plan with ownership and verification, or the chronic disease plan has entered a durable longitudinal management state. Residual follow-up can remain after closure if it is bounded, assigned, time-stamped, and no longer blocks the current clinical unit.

Terminality and medical quality are distinct: a death outcome or unsafe external takeover may end a trajectory without making it high quality. Likewise, an open trajectory may remain clinically valuable if the model provided safe and coherent care within the observation window.

Engineering bounded closure in practice. The high closure rate within a 100-turn safeguard depends on alignment between two design layers. At the *case-construction layer*, each CL120 case defines a bounded episode with a specified clinical horizon: the author specifies the immediate care goal, minimum safe closure evidence, acceptable closure types, false-closure traps, action-blocking residual risks, and residual issues that may

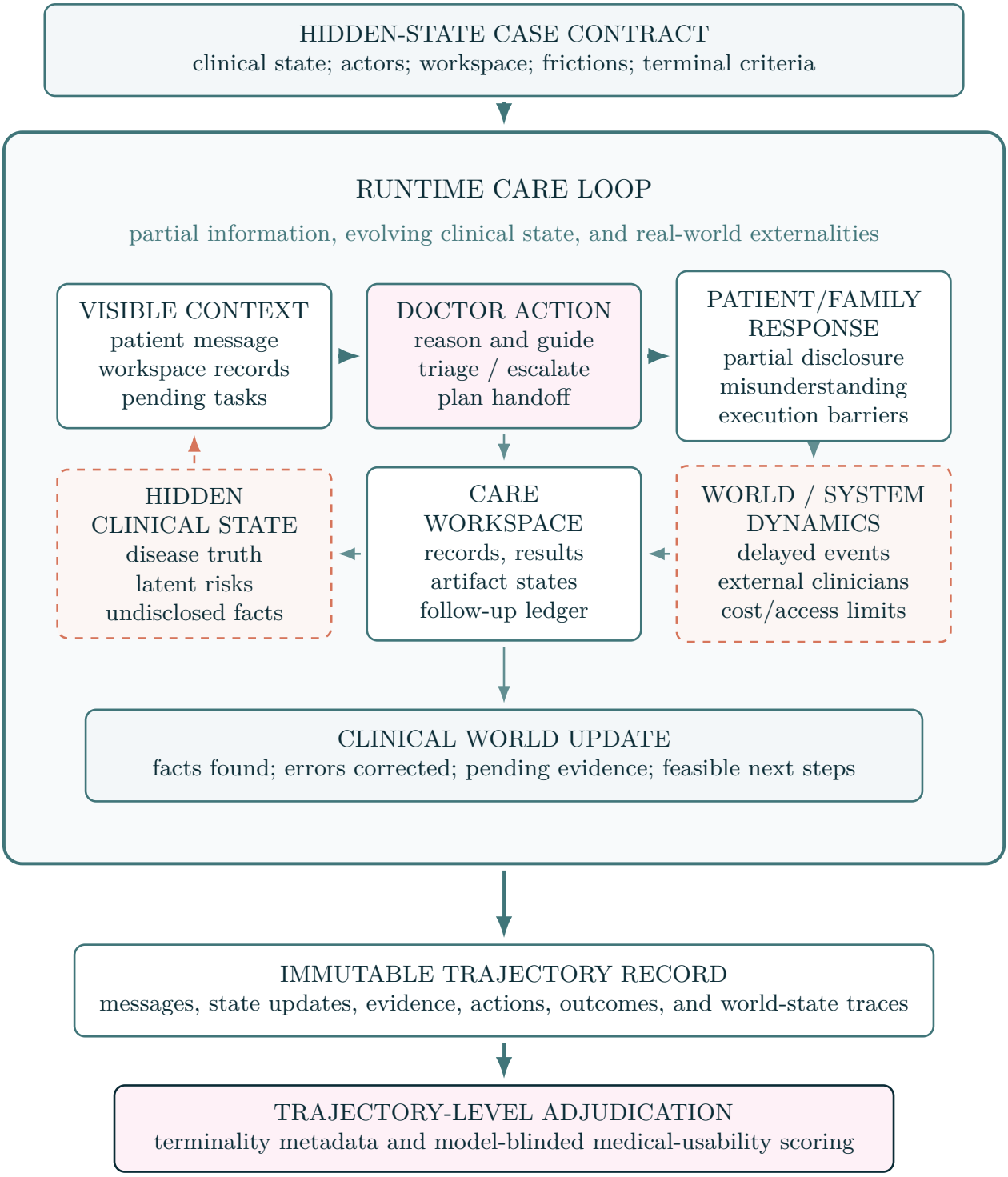

Figure 2 Overall structure of CareLoop as a partially observable clinical-world simulator. The figure should be read from top to bottom, with the central panel showing the left-to-right runtime interaction loop. A hidden-state case contract defines the simulated clinical state, actor behavior, workspace evidence, frictions, and terminal criteria. During the care loop, the doctor model acts under partial information; the patient/family actor may disclose incompletely or execute incorrectly; the workspace releases records, results, documents, or artifact-reliability states over time; and the world layer updates hidden clinical state, external clinician actions, delayed evidence, and real-world constraints. The runtime output is an updated clinical-state interpretation and feasible action plan, which is preserved in an immutable trajectory record for terminal-state adjudication and medical-usability scoring. Routine use of the simulator does not require four independent Judges. In this study, four-Judge review is the primary evaluation design for estimating trajectory-level measurements and quantifying evaluator sensitivity.

safely remain assigned to follow-up. At the *runtime layer*, the world and closure components are designed to avoid both premature success and endless friction loops. The world director and friction-lifecycle logic can resolve, compress, transfer, or terminate repeated low-yield barriers; tempo and receipt mechanisms prevent an order, referral, or follow-up promise from being treated as completed until the simulated world commits it; and closure review checks whether remaining risks have owners, timing, verification paths, and escalation triggers. This two-level design lets cases contain meaningful clinical complexity while still giving most trajectories a realistic path to bounded closure before the 100-turn cap.

3.9 Structured Trajectory Review

Each Judge reviews a model-blinded, outcome-independent packet derived from the immutable trajectory. The packet contains the case-specific evaluation contract, neutral operational metadata, and the complete ordered dialogue and workspace evidence when present. It excludes downstream closure summaries, quality reports, outcome-derived labels, and tested-model identity. The Judge records evidence across five clinical-usability dimensions, every authored responsibility-chain item and HO opportunity, closure, and important non-chain defects.

These structured fields support the friction-centered analysis: six 0–1 capability domains and FCC, with an explicit clinical-error gate in C-RWR. A frozen deterministic 1–5 ordinal score is retained only as a conventional supplementary comparator; its anchors and results appear in Supplementary Sections 7 and 7.

Turn count, open status, terminal closure failure, and reaching the maximum-turn safeguard are interpretive metadata. Any quality penalty must be tied to the underlying medical behavior, such as a safety lapse, failure to discover or update relevant state, infeasible action planning, or loss of clinical focus. The resulting measures estimate behavior within the authored clinical-world contract; they are not probabilities of patient benefit or certifications of deployment safety.

3.10 Traceability and Reuse of Existing Trajectories

Because trajectories are expensive to generate, adjudication must preserve traceability. In this study, closure adjudication and quality evaluation are performed on immutable trajectory artifacts. A closure rule may identify the earliest terminal turn using the same transcript and workspace evidence, but it does not alter the trajectory. This preserves previously generated trajectories while making the adjudication rule clinically explicit and reproducible.

4 Experiments

4.1 CL120 Case Set

CareLoop-120 (CL120) is a frozen set of 120 deidentified EHR-derived scenarios converted into CareLoop case contracts; raw patient records are not released. Each contract specifies the opening presentation, hidden clinical state, available workspace evidence, disclosure rules, expected friction, false-closure traps, and bounded episode-closure criteria. The set includes short bounded tasks, multi-step coordination problems, and longitudinal tasks.

Before CL120 was frozen, three licensed physicians reviewed candidate case drafts case by case, covering 140, 143, and 165 cases, respectively. The review criterion was whether each case, considered as a whole, was clinically plausible and consistent with general standards of diagnosis and treatment. Among the cases assessed by all three physicians, 130 received unanimous approval. We selected and froze 120 of these unanimously approved cases to improve the relative balance of case counts across clinical domains. This review established case-level clinical acceptability; it did not evaluate model-generated trajectories or validate LLM-Judge assessments.

Case difficulty was not inferred from post-run scores, trajectory length, nonclosure, runtime errors, or Judge disagreement, because reusing those outcomes would create a circular definition. We retain five separate, outcome-independent contract descriptors: responsibility-chain breadth, evidence-integration burden, authored friction intensity, temporal burden, and closure-coordination burden. Supplementary Figure S4 exam-

Table 4 The 10 real-world-friction exposures authored before model execution in CL120. Each label identifies the dominant execution obstacle for 12 case contracts; secondary frictions may coexist.

Friction exposure	Plain-language meaning and expected response
Misstated patient/family information	A patient or caregiver gives an inaccurate, simplified, or contradictory account. The model must detect the inconsistency, verify the relevant fact, and correct the care path without damaging trust.
Incomplete patient information	A clinically important symptom, history item, medication fact, or execution detail is missing or unknown. The model must elicit or retrieve enough information to act safely.
Document or image quality problem	A photograph, screenshot, report, or medication label is partial, blurry, second-hand, or otherwise unreliable. The model must manage evidence uncertainty and seek a safer verification route.
Delayed or staged result return	A laboratory, imaging, pathology, culture, or specialist result becomes available later. The model must preserve ownership, revisit the plan, and communicate the update.
External-system gap	Scheduling, referral, pharmacy, record transfer, or another institutional process fails or remains unclear. The model must establish a feasible fallback, owner, and timing.
Medication confusion or execution failure	The drug, dose, schedule, discontinuation, duplication, or actual use is uncertain or incorrect. The model must reconcile the regimen and make execution safer.
Low adherence or refusal	The patient resists monitoring, treatment, disclosure, or follow-up. The model must understand the reason, respect autonomy, and maintain a feasible safety plan.
Cost, transport, or work barrier	Money, travel, appointment access, or employment constraints obstruct care. The model must adapt the route while preserving clinical priority and follow-up responsibility.
Caregiver conflict or fatigue	Family members disagree, rotate responsibility, misunderstand roles, or become exhausted. The model must clarify ownership and prevent responsibility from diffusing unsafely.
Privacy, stigma, or sensitive history	The patient withholds psychiatric, sexual, substance-use, infectious, family, or other sensitive information. The model must create a safe disclosure path and avoid unsupported assumptions.

ines their associations with FCC, C-RWR, and interaction length. An exploratory aggregate index is reported only as a negative construct check and is not used as a difficulty scale.

The 120 cases were balanced across 10 primary real-world-friction strata (12 cases each). A friction label describes the obstacle that makes care harder to execute. Each case also contains four high-order (HO) test points authored before model execution, for 480 opportunities in total. An HO label describes the response capability the model should demonstrate. The two taxonomies are complementary: one obstacle can require several capabilities, and one capability can be tested under several obstacles. Tables 4 and 5 define every category in plain language; Supplementary Figure S1 shows their sparse, non-factorial coverage.

4.2 Tested Doctor-Role Models

We evaluated 10 deployed doctor-role endpoints on all 120 cases, yielding 1,200 final trajectories: GPT-5.6 Sol, GPT-5.4, DeepSeek-V4-Pro, GLM-5, Qwen3.8-Max, DeepSeek-V3.2, Kimi K2, MiniMax M3, Doubao-Seed-2.0-lite, and Qwen3-30B-A3B-Instruct-2507. Every model used the same runtime, frozen contracts, workspace interface, and 100-turn safeguard. Successful bounded closure, open-at-100, and runtime-error trajectories all remain in the denominator because each is part of the observed behavior of a long-form interactive system.

4.3 Judge Panel and Structured Evidence

The four experimental Judges were GPT-5.5, GPT-5.6 Sol, DeepSeek-V4-Pro, and GLM-5. They were not assumed to be equally calibrated or clinically correct. Each Judge received a model-blinded, outcome-independent packet containing the case-specific evaluation contract, neutral operational metadata, and the complete ordered trajectory, including dialogue and workspace interactions. Every formal cell used one complete-context request; no trajectory was clipped, sampled, chunked, or replaced by a summary.

Table 5 The eight HO response-capability types authored before model execution. These labels describe what the model should do when friction is encountered.

HO capability	Plain-language meaning
Hidden-state discovery	Actively uncover a clinically important fact that was not initially volunteered, visible, or understood.
Patient misinformation correction	Recognize and correct a harmful misunderstanding while preserving the patient's or caregiver's trust and ability to act.
False-positive/false-negative verification	Verify whether an apparent finding, report, symptom interpretation, or reassurance is reliable before changing care.
Real-world constraint navigation	Convert a practical barrier into a feasible alternative route with a clear owner, timing, fallback, and safety net.
Dynamic reprioritization	Change urgency or plan when new evidence, deterioration, improvement, or a new obstacle alters the risk.
Responsibility-chain repair	Detect a broken action, owner, handoff, result, or follow-up loop and restore a traceable path to completion.
Patient-state-sensitive communication	Adapt communication to emotion, literacy, cognition, privacy, readiness, or caregiver context so that the plan remains usable.
Safe bounded closure	End or transfer the episode only after residual risk, next steps, ownership, timing, and understanding are adequately handled.

Each Judge returned typed findings for every authored responsibility-chain item and HO point, five quality dimensions, closure, and important non-chain defects, together with a proposed ordinal grade. Transport failures, truncated or unparsable output, and schema-invalid output were retried. Substantive evidence–conclusion disagreement, clinical misclassification, and other schema-valid Judge behavior remained part of the experimental observations. No later model edited the content of a Judge's findings. The final matrix contains 4,800 canonical assessments.

4.4 Main Friction-Centered Measurement

The manuscript's central analysis uses a deterministic, code-defined friction-centered measurement framework computed from the unmodified structured fields in the 4,800 Judge records. It reports six 0–1 domains: information repair, constraint navigation, dynamic reprioritization, execution-loop repair, state-sensitive communication, and bounded closure/continuity. Each domain combines the applicable authored HO evidence with fixed supporting fields. Dimension ratings map from {unsafe, weak, partial, good, excellent} to {0, 0.25, 0.5, 0.75, 1}. The HO completion label is the primary scoring state: not triggered or not completed receives 0, partial completion receives 0.5, completion receives 0.75 unless both active model action and meaningful trajectory impact are documented, and fully supported completion receives 1. Supporting flags do not promote a lower completion label. Closure labels and the separate safety, executability, and traceability fields enter the closure component explicitly; one field is not silently rewritten from another. Supplementary Table S4 gives the complete executable mapping.

For each Judge–trajectory cell, the Friction Capability Composite (FCC) is a weighted mean over domains authored for that case. Fixed weights are 0.20 for information repair, 0.20 for constraint navigation, 0.15 for dynamic reprioritization, 0.20 for execution-loop repair, 0.10 for state-sensitive communication, and 0.15 for bounded closure/continuity; weights are renormalized over applicable domains. The safety-gated CareLoop Real-World Robustness score (C-RWR) multiplies FCC by 1.00 when no clinical-error gate is present, 0.75 for a minor/moderate error, and 0.35 for a serious error. FCC describes friction-handling capability, while C-RWR retains the consequence of clinically important error flags.

The main aggregation takes the median across four Judges for each trajectory and then averages the 120 trajectory values for each tested model. Confidence intervals use 10,000 case-paired bootstrap resamples. The FCC/C-RWR definitions and weights were fixed in executable code and were not optimized against model scores or model rankings. Sensitivity checks use the mean of Judges, leave one Judge out, perturb each domain weight by ±20%, use equal weights, leave one friction stratum out, and vary the treatment of authored but untriggered HO opportunities.

We quantified agreement separately for FCC, C-RWR, and all six domains. ICC(A,1) estimates absolute agreement for one Judge, and ICC(A,4) estimates the reliability of the four-Judge mean. We also calculated all six pairwise trajectory-level Spearman correlations and all six pairwise correlations between Judge-specific ten-model rankings. ICC confidence intervals use 2,000 case-paired bootstrap resamples that retain all available tested-model trajectories within each sampled case.

4.5 Care-Progression and Model-Gap Analyses

We report descriptive analyses that organize existing structured fields around the benchmark's task ontology. They introduce no new primary composite score and do not modify FCC, C-RWR, the cases, trajectories, or Judge findings.

For each of the 480 authored case-level HO opportunities and 10 model trajectories, four Judges recorded whether the opportunity was triggered, whether the model took an active action, whether that action had meaningful trajectory impact, and whether the HO was completed. We summarize 19,200 Judge–trajectory–HO assessments as a progression from realized opportunity to action, meaningful impact, and effective completion. Effective completion requires a completed HO together with active action and meaningful impact. Proportions are reported only among Judge-marked triggered opportunities. Case-cluster bootstrap intervals resample the 120 cases, preserving the dependence among models, Judges, and HO items within a case. These supporting fields characterize progression within the frozen HO assessment and do not constitute an independent recognition metric.

To identify where models separate, we report three complementary diagnostics. First, each additive Judge-cell FCC is decomposed into its weighted domain contributions after renormalizing over domains authored for that case. This identifies which domains contribute to a model's position relative to the across-model mean; the primary model ranking continues to use trajectory-level median-of-four FCC. Second, we report each domain's between-model range and standard deviation. A low domain mean indicates a shared bottleneck, whereas a large between-model range indicates discrimination among systems. Third, the difference between FCC and C-RWR describes the consequence of the clinical-error gate without relabeling that difference as pure medical-knowledge failure.

We compare runtime terminal status with Judge closure interpretation. For every trajectory, we count premature-closure and valid-open labels across four Judges and report exact vote thresholds. We also quantify case-level discriminatory yield using the range and standard deviation of the 10 model scores within each fixed case. These analyses evaluate whether the environment produces observable behavioral divergence under controlled case conditions.

Finally, we assess benchmark-level ranking stability. Random case subsets are compared with the complete 120-case ranking for descriptive convergence. A stronger split-half analysis repeatedly partitions every 12-case friction stratum into two disjoint six-case halves and correlates the two independently averaged 10-model rankings. All resampling uses the fixed seed 20260919. Subset-to-full curves share cases with their target and are interpreted as convergence visualizations; disjoint split halves provide the primary internal stability check.

4.6 Blinded Physician Comparison

A separate blinded comparison included a final analytic panel of 139 physicians practicing at tertiary or higher-level hospitals in China: 64 residents, 51 attendings, and 24 associate-chief or chief physicians. Physicians were assigned within the 13 frozen CL120 clinical domains, with 7–16 physicians contributing in each domain. A case and its 10 tested-model trajectories formed the minimum assignment packet. The same five physicians reviewed all 10 trajectories for that case: two residents, two attendings, and one associate-chief or chief physician. This fixed-panel design covered all 120 cases and 1,200 trajectories, yielding 6,000 separately collected assessments through 120 case-specific assignments involving 105 unique five-physician combinations. Individual workloads ranged from 10 to 100 trajectories (median, 40). Reviewers received the same complete blinded trajectory evidence and clinical constructs used by the LLM Judges through a clinician-facing standardized categorical form. Tested-model identity and trajectory provenance were absent from the clinician-facing materials and linked to the analysis records only after review completion. The struc-

tured fields covered every authored responsibility-chain item and HO point, the five foundational clinical dimensions, closure safety/executability/traceability, and important non-chain defects.

For each physician assessment, the frozen FCC/C-RWR implementation transformed the structured clinical fields into the same six 0–1 domains, FCC, and C-RWR used for the LLM Judges. These scores were computed after review and were not physician-entered grades. The two resident assessments and two attending assessments were summarized by within-title medians; the single associate-chief/chief assessment represented the senior stratum. The three stratum values then received equal weight through a second median. Sensitivity analyses used the all-five median, all-five mean, and resident-plus-attending-only aggregation.

The assignment design fixes both case content and the physician panel while the tested model changes. We therefore designated the 10-model ordering within each case as the primary physician-comparison unit. For every case, we calculated Spearman rank correlation across the same 10 model trajectories and summarized the 120 case-level coefficients by their median, interquartile range, threshold frequencies, and a 95% bootstrap interval obtained by resampling cases. Top-1 overlap and tie-aware expected Top-3 overlap provide additional rank-based descriptions. For the latter, models tied at the third-place cutoff receive fractional membership so the result does not depend on model order. The 120 assignments involved 105 unique five-physician combinations; a sensitivity analysis therefore resampled exact panel combinations while retaining all cases assigned to each sampled panel. The aggregate 10-model ordering after averaging over all cases is reported as descriptive context because it has only 10 model-level observations. Pooled cross-case trajectory correlations and exact-score agreement are not primary evidence because panels differ across cases and such summaries can mix case content, panel calibration, and model performance.

For a symmetric comparison, each case yielded seven 10-model score streams: resident-title median, attending-title median, senior assessment, and four individual LLM Judges. Physician-internal concordance was the case-wise median of the three title-stratum pair correlations; each Judge's physician alignment was the corresponding median of its three correlations with the physician strata. We then summarized these case-wise quantities over 120 cases and used paired case bootstrap intervals for contrasts. Same-title analyses compared the two residents and the two attendings within each fixed case panel. All estimates are descriptive. Physicians provide an independent clinical comparison panel. Concordance measures agreement with that panel under the shared protocol and carries no assumption of error-free ground truth or clinical accuracy.

4.7 Analysis Hierarchy

The Results section is organized around the scientific questions. It first reports FCC/C-RWR and capability profiles, then asks where models differ, how the environment renders those differences observable, whether action becomes impact, and whether operational termination constitutes safe completion. Ranking stability and a compact physician comparison then address measurement reliability. Detailed Judge agreement, physician strata, the frozen 1–5 comparator, burden analyses, matched model–Judge associations, sensitivity checks, and complete representative evidence are reported in the Supplementary Material. All 1,200 trajectories, 4,800 canonical Judge results, 6,000 physician assessments, executable scoring and analysis code, and reproducibility metadata are released for independent inspection.

5 Results

5.1 Experiment Overview

CL120 generated 1,200 trajectories from 10 models on the same 120 case contracts. Of these, 1,106 reached the runtime's bounded-closure state, 56 remained open at the 100-turn safeguard, and 38 ended with a runtime error. Every trajectory remained in the analysis. Four model-blinded Judges produced 4,800 complete-context assessments with no missing Judge–trajectory cell. The separate physician comparison contained 6,000 structured assessments, with one fixed five-physician panel evaluating all 10 model trajectories within each case.

Table 6 Main friction-centered results. FCC measures responses to real-world-friction opportunities authored before model execution; C-RWR applies the structured clinical-error gate. Each estimate averages 120 trajectory-level medians across four Judges. Confidence intervals use 10,000 case-paired bootstrap resamples.

Rank	Model	FCC	95% CI	C-RWR	95% CI
1	GPT-5.6 Sol	0.824	0.799–0.849	0.753	0.719–0.786
2	GPT-5.4	0.786	0.758–0.813	0.680	0.645–0.713
3	DeepSeek-V4-Pro	0.783	0.755–0.810	0.621	0.583–0.659
4	Qwen3.8-Max	0.773	0.743–0.801	0.619	0.581–0.657
5	MiniMax M3	0.748	0.715–0.779	0.586	0.545–0.626
6	GLM-5	0.740	0.709–0.770	0.598	0.559–0.638
7	DeepSeek-V3.2	0.738	0.707–0.768	0.568	0.530–0.603
8	Kimi K2	0.694	0.657–0.729	0.510	0.467–0.552
9	Doubao-Seed-2.0-lite	0.651	0.616–0.686	0.471	0.431–0.511
10	Qwen3-30B-A3B-Instruct-2507	0.645	0.611–0.678	0.450	0.409–0.492

Note. Ranks follow FCC. FCC and C-RWR are reproducible benchmark measurements; neither is a clinical gold standard or a deployment-safety certificate.

5.2 Models Differ in Friction-Sensitive Care Progression

Table 6 reports the central results. GPT-5.6 Sol had the highest FCC (0.824; 95% CI, 0.799–0.849) and C-RWR (0.753; 0.719–0.786). GPT-5.4, DeepSeek-V4-Pro, and Qwen3.8-Max formed the next FCC tier with overlapping intervals. The C-RWR ordering separated GPT-5.4 more clearly because the safety gate penalized structured clinical errors.

The domain profile in Figure 3 reveals a multidimensional ranking pattern. Constraint navigation and state-sensitive communication were relatively strong across models, whereas information repair and dynamic reprioritization were lower and more separating. Figure 4 provides the complementary exposure view. Privacy/stigma/sensitive-history, document/image-quality, and incomplete-information cases had the lowest equal-model FCC means. These strata identify settings in which the current case contracts created greater observed difficulty; they do not establish an intrinsic causal hierarchy among friction types.

5.3 Where Model Differences Arise

The largest raw between-model domain ranges occurred in information repair (0.228), dynamic reprioritization (0.221), and bounded closure/continuity (0.216). Execution-loop repair spanned 0.198, state-sensitive communication 0.170, and constraint navigation 0.162. A low mean and a large range answer different questions: hidden-state and information-repair tasks were both difficult and discriminative, whereas constraint navigation was relatively strong overall yet still contributed to separation between the highest and lowest systems.

The additive Judge-cell FCC decomposition in Figure 6B identifies how these domains combine. Relative to the across-model mean, GPT-5.6 Sol gained its largest contribution from information repair, followed by constraint navigation, dynamic reprioritization, and bounded continuity. The lowest-ranked models lost ground across several domains, indicating distributed capability deficits. This is a diagnostic decomposition of the parallel mean-of-Judge FCC representation; the reported primary ranking remains the trajectory-level median-of-four result.

Clinical errors created an additional distinction. The mean FCC-to-C-RWR penalty ranged from 0.072 for GPT-5.6 Sol to 0.195 for Qwen3-30B. DeepSeek-V4-Pro and MiniMax M3 had FCC values near the second and middle tiers but larger safety-gate penalties than GPT-5.4. These penalties describe observed clinical correctness and safety under trajectory conditions. They cannot be interpreted as a pure measure of medical knowledge because a clinical error can follow from missing state, failed verification, stale planning, incomplete execution, or unsafe closure as well as from a factual knowledge deficit.

The six domains remained related but nonredundant. Pairwise-complete trajectory-level Spearman correlations ranged from 0.228 to 0.714. The lowest correlations included constraint navigation with dynamic

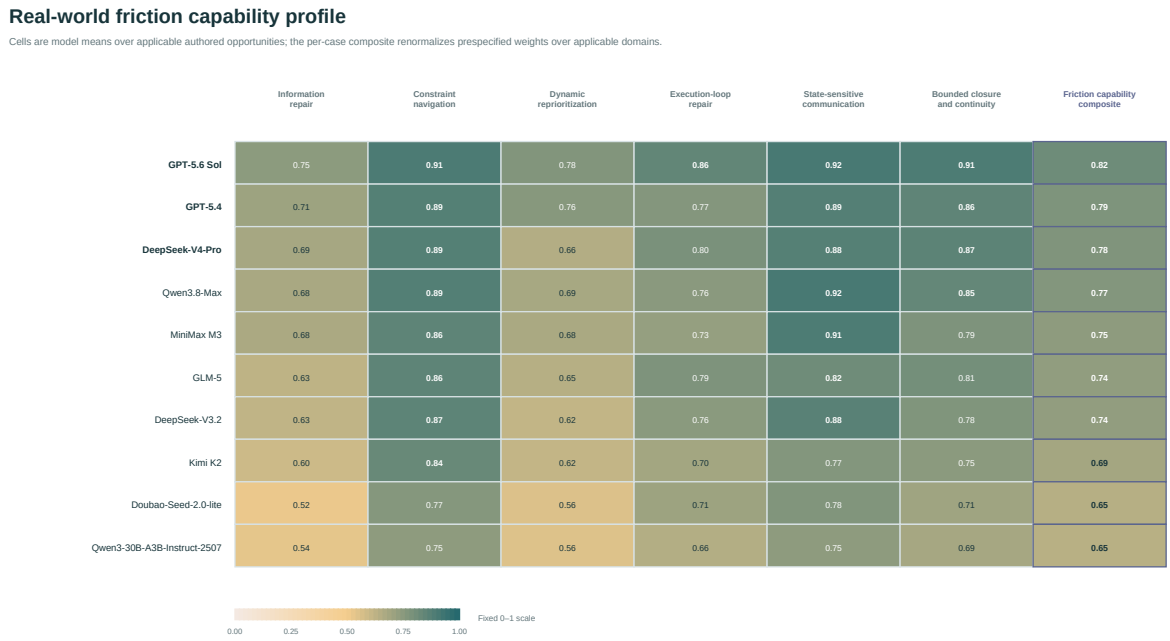

Figure 3 Six-domain capability profile. Cells are model means over cases in which the corresponding capability was authored before model execution. The final column is the fixed-weight FCC, with weights renormalized over applicable domains.

reprioritization (0.228) and information repair with execution-loop repair (0.273). A single score is therefore useful for ranking, while the domain profile is necessary for explaining why models differ.

5.4 CareLoop Makes Latent Capability Gaps Observable

The 120 fixed cases generated substantial within-case model divergence. The median range between the highest and lowest model FCC within a case was 0.414; 89 of 120 cases had a range of at least 0.30, and 38 had a range of at least 0.50. Because the same case contract, hidden state, friction design, and evaluation opportunities were used for all 10 models, fixed case sampling attributes these divergences to different paths through controlled behavioral forks.

Figure 5 follows the 15,537 Judge assessments in which an authored HO opportunity was marked as realized. Models took an active action in 0.935 of these assessments (case-cluster bootstrap 95% CI, 0.924–0.945), but meaningful trajectory impact occurred in 0.745 (0.723–0.766), and effective completion in 0.697 (0.673–0.720). The overall action-to-impact gap was 0.190 (0.175–0.205). The environment therefore exposed failures that would be hidden by counting responses or attempted actions alone.

Hidden-state discovery was the clearest shared bottleneck: effective completion was 0.484 and the action-to-impact gap was 0.320. Safe bounded closure followed at 0.636 effective completion and a 0.266 gap; verification of false-positive or false-negative information reached 0.666 and a 0.213 gap. Constraint navigation and state-sensitive communication exceeded 0.79 effective completion. Across models, the action-to-impact gap ranged from 0.123 for GPT-5.6 Sol to 0.282 for Doubao-Seed-2.0-lite. The main separation arose in whether a response changed the subsequent trajectory. Stronger systems converted appropriate responses into consequential state changes more often.

These results explain the role of the environment design. Distributed information makes elicitation necessary; actor-specific beliefs make verification necessary; delayed evidence makes memory and plan revision necessary; external action makes receipts and follow-up necessary; and explicit closure criteria make unresolved risk visible. The benchmark creates pressure points at which a latent limitation can propagate into a later, auditable consequence.

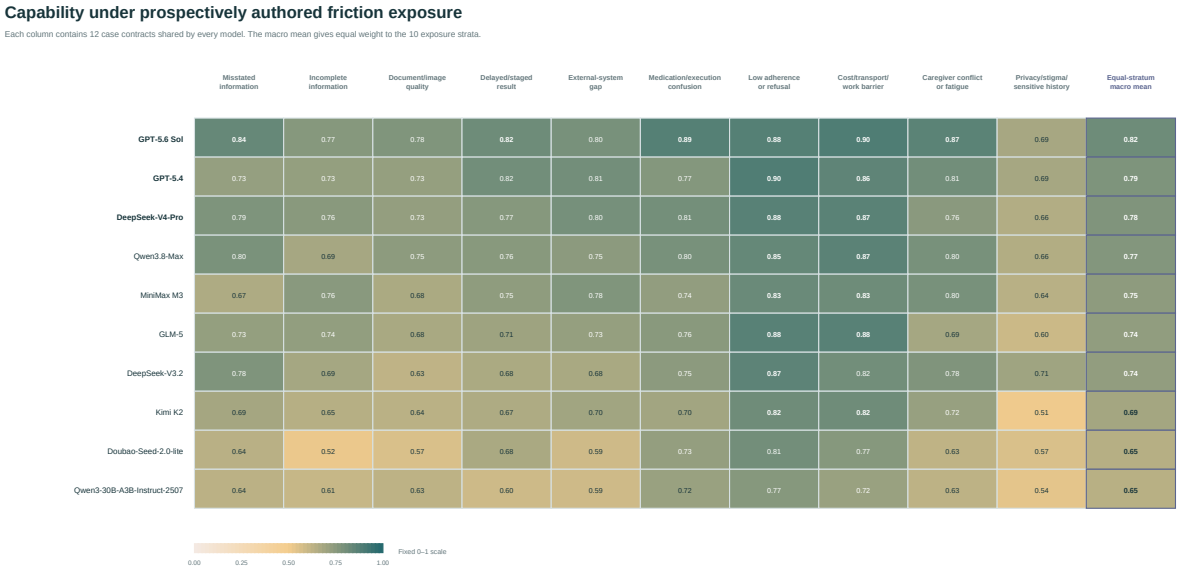

Figure 4 Performance across the 10 friction exposures authored before model execution. Each friction column averages the same 12 case contracts for every model. Differences are descriptive because friction strata contain different clinical content and HO mixtures.

5.5 Operational Termination Does Not Define Safe Completion

Runtime status and closure quality diverged. Among the 56 trajectories that remained open at turn 100, 40 received a valid-open label from at least three Judges and 22 received that label from all four. Among 1,106 trajectories marked `closed_success`, 437 received at least one premature-closure judgment, 151 received at least two, and 9 received at least three. These vote counts preserve evaluator disagreement as part of the result.

The model-level patterns reinforce the distinction. Doubao-Seed-2.0-lite closed 117 of 120 cases but ranked ninth by FCC; GPT-5.6 Sol closed 108 and ranked first. Across only 10 model points, the closed-success rate had a descriptive Spearman correlation of -0.652 with FCC, whereas the Judge-derived closure-process score correlated 0.927. The small number of models precludes strong inferential claims, but the direction is clear: ending the episode poorly predicts safe care progression. A well-formed open state can preserve responsibility while evidence is pending; a nominally closed state can conceal an unresolved obligation.

5.6 Local Evaluation Uncertainty Coexists with Benchmark-Level Stability

Figure 6A shows how ranking stability changes with case coverage. Random 20-case subsets had median rank correlations of 0.903 for FCC and 0.915 for C-RWR with the complete 120-case ranking. At 60 cases, the corresponding values were 0.964 and 0.976. Because these subset curves share cases with the full target, the stronger internal check used nonoverlapping, friction-stratified 60-case halves. Across 5,000 partitions, the median split-half rank correlation was 0.915 for FCC (95% simulation interval, 0.794–0.976) and 0.927 for C-RWR (0.818–0.988); Top-1 agreement was 99.2% and 100%, respectively.

Local evaluations were much less stable. For FCC, single-Judge absolute-agreement ICC was 0.257 and the four-Judge-mean ICC was 0.581; the corresponding C-RWR values were 0.211 and 0.517. Within each case, the median correlation among the three physician title-stratum rankings was 0.306 for FCC and 0.321 for C-RWR. Alignment between physician strata and the two GPT Judges was higher: 0.545/0.547 for GPT-5.5 and 0.516/0.521 for GPT-5.6 Sol on FCC/C-RWR. In paired case-level comparisons, the GPT-5.5 advantage over physician-internal concordance was 0.255 (95% CI, 0.192–0.296) for FCC and 0.232 (0.172–0.271) for C-RWR; the GPT-5.6 Sol differences were 0.171 (0.132–0.225) and 0.183 (0.140–0.232). DeepSeek-V4-Pro and GLM-5 were at or below physician-internal concordance. These comparisons indicate protocol alignment

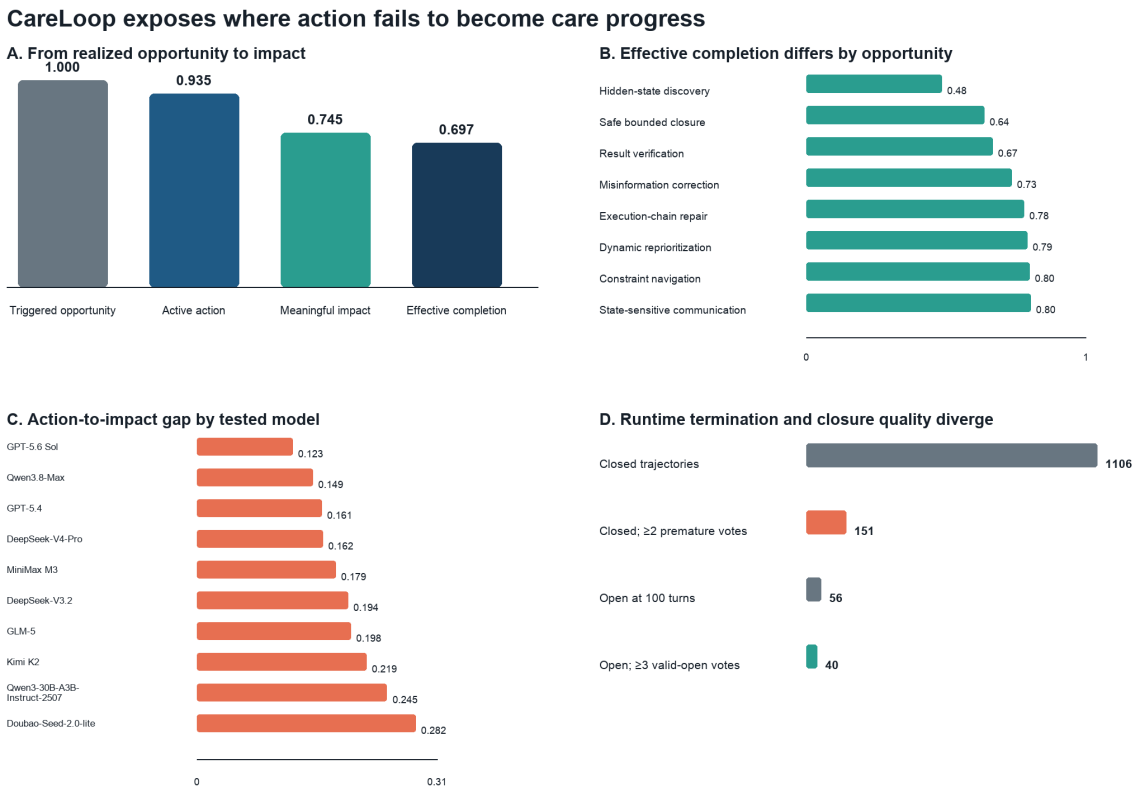

Figure 5 Conversion from realized opportunities to care progress. Panel A aggregates Judge-marked triggered HO assessments. Panel B reports effective completion by HO type. Panel C shows each model's active-action rate minus meaningful-impact rate among triggered assessments. Panel D contrasts runtime status with semantic closure judgments. These fields support the existing HO states and are presented as a descriptive process decomposition, not as a new primary score.

and consistency, not clinical superiority.

The fixed-panel physician design provides a more interpretable comparison than pooled trajectory correlations: the same five physicians ranked all 10 model trajectories for a given case. Under equal-title physician consensus, median case-wise rank correlation was 0.638 for GPT-5.5, 0.631 for GPT-5.6 Sol, and 0.580 for the four-Judge median on FCC. For C-RWR, the corresponding values were 0.677, 0.663, and 0.610. After aggregation over all 120 cases, physician consensus and the four-Judge median produced highly similar 10-model rankings ($\rho = 0.988$ for FCC and 0.964 for C-RWR), although each aggregate coefficient contains only 10 model observations.

These results support a two-level interpretation. Individual long trajectories do not have an error-free evaluator, and a rubric-style physician label should not be treated as a clinical gold standard. Broad case coverage and repeated structured assessments nevertheless recover a stable comparative ordering. Detailed Judge agreement, physician title strata, panel-cluster sensitivity, and case-level distributions appear in Supplementary Sections S8–S10.

5.7 Robustness

The broad FCC ordering was robust to the tested analytical choices. Leave-one-Judge-out rankings correlated 0.939–0.988 with the all-four-Judge ranking. Equal domain weights and every one-at-a-time ±20% weight perturbation preserved the exact ordering. Leaving out one friction stratum at a time produced correlations of

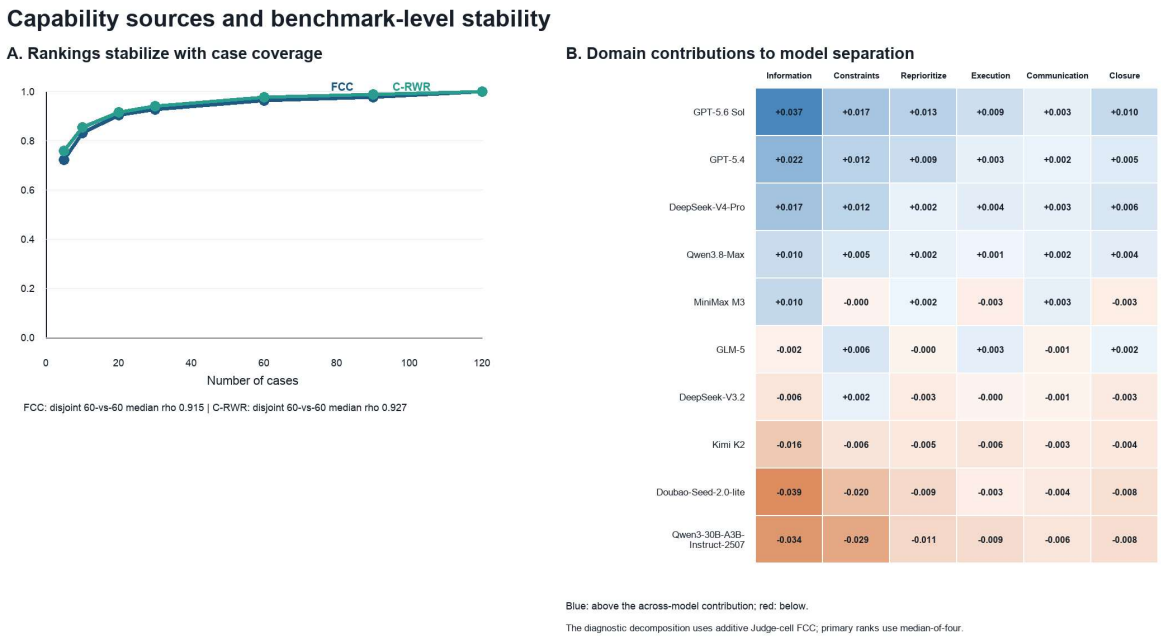

Figure 6 Sources of model separation and benchmark-level stability. Panel A shows convergence of model rankings as case coverage increases and reports the disjoint friction-stratified split-half result. Panel B decomposes additive Judge-cell FCC into domain contributions relative to the across-model mean. The decomposition explains capability sources and does not replace the primary median-of-four ranking.

0.964–1.000. Triggered-only analyses yielded 0.952–0.976, depending on Judge aggregation. A deterministic normalization of redundant HO-state and closure fields preserved the exact FCC and C-RWR orderings.

The frozen 1–5 ordinal comparator, including its associations with FCC and model ordering, is reported in Supplementary Section 7. It does not anchor the main case selection, model ranking, or scientific narrative.

6 Discussion

6.1 Medical Intelligence in the Context of Patients' Lives

CareLoop treats medical intelligence as a capacity exercised with a person in an unfolding clinical world. The patient has beliefs, relationships, preferences, constraints, and agency; each can alter what information becomes available, which plan remains feasible, and what responsibility persists. A plausible answer is therefore one component of the task. The experiment reveals the resulting knowledge-to-practice gap: models usually responded when a designed opportunity appeared, yet active response exceeded meaningful impact by 0.190 overall. The strongest and weakest systems differed most clearly in their ability to make an appropriate response matter within the patient's evolving situation.

This gap reframes patient-facing capability. A model may recognize the importance of a symptom without eliciting it, identify a medication discrepancy without verifying the actual regimen, or recommend follow-up without establishing a feasible next step. Clinical usefulness depends on a chain of transformations: knowledge guides attention to the person; communication reveals hidden state; new information changes the plan; the plan respects autonomy and real constraints; execution returns evidence; and unresolved risk retains ownership. A break anywhere in this chain can leave sound medical knowledge unused.

6.2 What Separates the Tested Models

The observed differences span several capabilities. Information repair, dynamic reprioritization, and bounded continuity showed the largest raw between-model ranges. Constraint navigation and communication were stronger on average and still contributed to the separation between the highest and lowest models. Clinical-error gating widened some gaps further. These components explain how two models with similar fluency or action frequency can produce markedly different trajectories.

The present design measures medical knowledge together with process competence. Errors may arise from missing factual knowledge, acceptance of an incorrect patient statement, failure to retrieve a record, omission of a returned result, or closure before decisive state is known. An oracle-state control could present all decisive facts at the outset and compare that performance with the partially observable CareLoop episode. Within the present experiment, much of the observed separation lies in acquiring, verifying, operationalizing, and maintaining clinical state; response production alone captures a smaller part of the difference.

6.3 CareLoop as a Capability-Elicitation Environment

CareLoop creates diagnostic pressure through explicit mechanisms. Information asymmetry requires elicitation; inconsistent beliefs require verification; delayed evidence requires longitudinal memory; feasibility barriers require plan adaptation; external actions require receipt tracking; and false-closure traps require residual-risk inspection. Each mechanism makes a capability necessary at an interpretable point. Because all tested models face the same fixed case contract, their divergence records different paths through the same designed world.

The within-case score ranges and opportunity-to-impact funnel show that these mechanisms expose consequential differences. Most cases materially separated the 10 models, many active responses failed to alter the trajectory, and runtime success flags concealed premature closure. A benchmark ending at the first recommendation or tool call would miss much of this behavior.

These findings remain descriptive. An authored opportunity is a diagnostic probe inside a simulation; causal identification of a friction effect would require randomized intervention. A non-triggered opportunity may reflect earlier trajectory choices or simulator dynamics, and friction-stratum averages combine different clinical content and HO mixtures. Controlled factorial experiments are therefore needed to estimate the inherent causal difficulty of individual friction types.

6.4 From Executable State Change to Consequence Stewardship

Executable-agent benchmarks advanced evaluation by grounding success beyond the plausibility of an answer. WebArena, OSWorld, SWE-bench, and AppWorld ask whether an agent changed task-relevant external state, while τ-bench, τ^2-bench, and organizational environments add users, policies, coworkers, and shared control. CareLoop inherits this execution-grounded commitment. Its departure lies in the kind of world that must be represented and the kind of consequence that can count as progress.

A repository task such as SWE-bench is complex and genuinely consequential, yet its task-relevant state is highly digital, its issue usually supplies a comparatively stable objective, a patch directly changes the repository, and executable tests provide a strong external oracle. Clinical care is a partially observable socio-technical world. Relevant state can be split among underlying clinical state, patient and family belief, records, and actual execution; each may evolve independently. The agent cannot directly control disclosure, understanding, adherence, family action, external clinicians, or healthcare systems, and new evidence can change both the feasible path and the appropriate goal. This is a shift from agency exercised largely through control toward agency under limited control.

Limited control creates an action-to-consequence gap. A recommendation does not guarantee adoption, an order does not establish that a test occurred, a referral does not establish handoff, and an explanation does not establish understanding. Evaluation must therefore follow verification, follow-through, repair, handoff, and continuing responsibility. The trajectory carries information that a final state alone cannot preserve: two episodes can end in similar states even though one model delayed recognition or allowed preventable risk before recovering, while another maintained safe progression throughout.

Completion changes as well. A clinically responsible episode need not end with every problem resolved. It may remain open when residual risk has an owner, a time boundary, a verification path, and an escalation condition. We call this *accountable openness*: incompleteness that remains controlled and auditable. Conversely, a nominally closed episode can represent premature abandonment. This completion semantics motivates trajectory-based consequential evaluation rather than reliance on terminal state alone.

CareLoop operationalizes these ideas as *consequence stewardship*: discovering relevant state under uncertainty, deciding what information to trust, acting within limited authority, checking whether intervention changed the world, revising goals as evidence evolves, repairing failures, and preserving accountability for what remains. Consequence stewardship is an evaluation perspective inside the simulated environment, not evidence of clinical effectiveness or autonomous safety. Its broader relevance is to open human worlds in which agents must act among independent people, delayed outcomes, changing goals, and no single immediate oracle.

6.5 Evaluator Uncertainty and Benchmark-Level Stability

Single-trajectory assessment remained uncertain for physicians and LLM Judges alike. Physician title strata sometimes disagreed with one another more than a physician consensus stream disagreed with a strong GPT Judge. The comparison measures protocol concordance and does not imply greater clinical expertise for an LLM. Long, partially observable trajectories permit legitimate differences in risk tolerance, evidence interpretation, and closure expectations even under a shared structured protocol.

The aggregate findings provide a complementary lesson. Model rankings stabilized rapidly with case coverage, disjoint 60-case halves yielded highly similar orderings, and physician and four-Judge aggregate rankings converged over 120 cases. Repeated structured measurement can reduce local instability and leave uncertainty visible at the individual trajectory. Strong general-purpose models can serve as scalable evaluators in this setting when multiple reviewers, transparent fields, sensitivity analyses, and independent physician comparison remain part of the design.

Matched model–Judge analyses found no uniform self-favoring pattern across model families, and aggregate model orderings were stable across Judges. This supports a limited form of evaluator reflexivity: a model family can participate in evaluating model behavior without systematically placing itself first. Blinding, multiple evaluators, and human comparison remain necessary safeguards.

6.6 Limitations and Next Steps

CL120 is a finite simulated benchmark. Its 120 cases cover a bounded set of specialties, institutions, patient populations, and care pathways. The simulator targets structural realism and auditable friction; real patients and organizations exhibit greater diversity, strategic behavior, and uncertainty. Clinical outcomes and deployment safety remain untested.

The present implementation is primarily text-mediated, but CareLoop's actor-, state-, and event-based design can be extended to governed medical images, physiological waveforms, speech, video, and longitudinal device streams as partially observable evidence. Such extensions would test not only whether a model interprets each modality correctly, but whether the resulting evidence changes the plan and is carried through verification, consequence, repair, and accountable continuity.

The HO and friction design is sparse and non-factorial. Domain correlations use different applicable case subsets, and friction-specific rankings rely on 12 primary cases per stratum. The action-to-impact analysis uses Judge fields that record active action, meaningful trajectory impact, and HO completion. The physician comparison evaluates the same structured CareLoop constructs and may share protocol-induced assumptions with the LLM Judges.

Future work should expand specialty and institutional coverage, test multilingual and culturally distinct friction, introduce controlled factorial variants, and compare simulated trajectories with carefully governed prospective workflows. An oracle-state baseline could separate medical knowledge from information-acquisition and execution failures. Repeated runs of the same model could distinguish capability from stochastic reliability. Additional deterministic receipts and externally verifiable actions could reduce semantic-evaluator dependence where the clinical task permits them.

7 Conclusion

CareLoop begins from a simple premise: knowing medicine and practicing it in the context of patients' lives are distinct capabilities. Its clinical-world contracts represent patients and families as stateful people whose beliefs, choices, relationships, resources, and changing circumstances shape the clinical path. Hidden information, incorrect beliefs, delayed evidence, feasibility constraints, external execution, and closure obligations give latent model limitations observable consequences across an unfolding trajectory.

Across 1,200 trajectories, the tested models separated most clearly in information repair, dynamic updating, conversion of action into meaningful impact, safety, and continuity. Frequent action could coexist with limited progress, and operational termination could conceal unsafe closure. Local physician and Judge assessments remained variable; broad case coverage and repeated structured evaluation recovered stable model comparisons.

The central lesson extends beyond a leaderboard. Patient-facing medical agents should be evaluated by what their knowledge accomplishes with a person: discovering what is missing, understanding beliefs and constraints, communicating responsively, verifying what happened, revising as circumstances change, and preserving ownership of unresolved risk. In an open human world, useful agency operates under limited control, and responsible progress may end in accountable openness rather than a terminal success state. CareLoop provides a reproducible environment for trajectory-based evaluation of this knowledge-to-practice gap and for studying consequence stewardship. It remains an AI research evaluation; clinical effectiveness and safe autonomous deployment are untested.

Data and Code Availability

Code and accompanying public artifacts are available at `https://github.com/jdh-algo/CareLoop`. The synchronized release includes the lightweight CareLoop runtime, rubrics, provider-configuration templates, case-construction documentation, the public deidentified CareLoop-120 (CL120) case set, and a supplementary package containing all 1,200 normalized trajectories from the experiment. The trajectory supplement is organized under `public_supplement/cl120_20260911/` and includes normalized interaction/workspace event sequences, the 4,800 accepted structured Judge assessments used in analysis, 6,000 structured physician trajectory assessments, reproducible care-progression, model-gap, ranking-stability, and physician–Judge analyses, mapping indexes, figures, and reproducibility metadata. The directory `cases/public_cl120_deidentified_120/` is the sole public source for case definitions; trajectories and Judge assessments refer to it by stable identifiers and do not embed duplicate case contracts. The physician records contain opaque pseudonymous reviewer identifiers that support repeated-measurement and assignment auditing; names, institution names, contact details, and the private identity mapping are excluded. The terminal-status distribution in the released trajectory package is 1,106 closed-success trajectories, 56 open-at-100 trajectories, and 38 runtime-error trajectories. The public package also excludes raw EHR source material, source-to-case mapping tables, internal file paths, API keys, provider payloads, notebook tokens, and operational logs. Public trajectories are normalized simulation and evaluation records; they provide neither clinical advice nor evidence of deployment safety.

Supplementary Material

This Supplementary Material documents the CareLoop-120 (CL120) case contract, runtime mechanisms, friction-centered measurement layer, selected secondary analyses, blinded physician comparison, the frozen ordinal comparator, evidence-verified representative trajectories, and the public artifact layout. The included sections clarify the principal FCC/C-RWR findings, explain the CareLoop framework, and identify questions relevant to future validation. The machine-readable public release remains the authoritative location for complete trajectories and tables.

S1. CL120 Case-Contract Fields

A CL120 case extends the question–answer format into a hidden-state clinical-world contract. The contract defines what the patient or caregiver initially reveals, what remains hidden unless elicited, what evidence can be obtained from the workspace, what real-world frictions may alter execution, and what bounded closure would mean for the current responsibility unit. Table S1 summarizes the main fields used to create and interpret CL120 trajectories.

Table S1 CL120 case-contract fields. Each row identifies a component authored before model execution and explains how it constrains the simulated clinical world. Together, the fields determine what is true, what each actor initially knows, what evidence can become visible, which execution frictions can intervene, what counts as bounded closure, and which case-specific responsibilities are later judged. These are environment specifications, not variables inferred from model outcomes.

Field	Role in CareLoop
Initial presentation	The first patient or caregiver message that initiates the simulated clinical episode. It usually contains partial, imprecise, or emotionally framed information, leaving the clinical picture incomplete.
Hidden clinical state	The underlying clinical facts, risks, diagnoses, medication exposures, pending results, and longitudinal state that should not be directly exposed to the doctor model unless clinically elicited or revealed through the workspace/world state.
Patient/family actor profile	Literacy, emotion, trust, privacy concerns, caregiver reliability, adherence tendencies, misunderstanding patterns, and response style used to simulate an independent patient or family member with agency, memory, and constraints.
Workspace contract	Release-safe electronic health record snippets, medication lists, test results, clinical notes, external clinician events, and other artifacts that can be inspected during a trajectory. The relevant question is whether the doctor model seeks and uses the right evidence at the right time.
Runtime frictions	Cost, access, stigma, delayed results, poor adherence, family disagreement, medication confusion, transportation barriers, incomplete disclosure, and other execution constraints that affect whether advice becomes usable care.
Dynamic event space	Candidate clinical-world changes such as improvement, deterioration, adverse effects, new findings, external-care updates, loss to follow-up, or adverse terminality. Events must be causally anchored to the case state and are not arbitrary drama.
Closure contract	The bounded conditions under which the current responsibility unit can be considered safely closed, handed off, stabilized, or left open. Closure is runtime metadata, not a direct medical-quality score.
Evaluation contract	Case-specific responsibility-chain requirements, non-negotiable safety constraints, high-order test points, score anchors, and closure traps used for the ordinal 1–5 judge assessment.

S2. Runtime Frictions and Stochastic Clinical-World Events

CareLoop simulates clinical reality by letting the patient-facing interaction unfold under partial observability and changing world state. Random or contingent events are not sampled every turn and are not introduced merely to prolong a conversation. They require a plausible causal anchor in the case, such as disease progression, medication exposure, poor adherence, external system delay, patient behavior, environmental exposure,

or a complication of prior care. Table S2 gives examples of runtime mechanisms that can appear in CL120 cases.

Table S2 Examples of CareLoop runtime mechanisms and the behavior each is intended to expose. The mechanism column names the simulated process; the clinical-world interpretation states what changes or remains uncertain; and the evaluation consequence states what becomes observable in the frozen trajectory. The list is illustrative; activation frequencies depend on individual case contracts: each case activates only mechanisms required by its contract, and the presence of a mechanism does not itself imply model failure.

Mechanism	Example scenario	How it becomes evaluable in CareLoop
Incomplete disclosure	A patient initially reports vague discomfort but omits a red-flag symptom, medication exposure, pregnancy possibility, or sensitive history.	The actor disclosure rules determine what is volunteered and what requires targeted, privacy-preserving questioning. The model is evaluated on whether it elicits actionable missing information without unsafe assumptions.
Caregiver misunderstanding	A family member confuses old and new prescriptions, test labels, discharge instructions, or surgical-risk wording.	The trajectory tests whether the model corrects the misconception, verifies teach-back, and converts ambiguous caregiver reports into an executable plan.
Delayed evidence	A laboratory result, imaging report, culture, pathology result, or specialist note becomes available only after initial advice.	The workspace/world-state layer exposes new evidence later, making it possible to assess whether the model updates its reasoning and follow-up responsibilities.
Cost/access barrier	The patient cannot afford a test, cannot travel, cannot obtain an appointment, or faces pharmacy unavailability.	The actor responds with feasibility constraints. Good care preserves medical priority while adapting the route, timing, safety-netting, and responsible handoff.
Adherence and low-literacy barrier	The patient repeats instructions incorrectly, refuses monitoring, cannot distinguish medication timing, or uses home devices unreliably.	Later turns and receipt checks reveal whether advice was understood and executable. The model must verify execution and cannot rely on a plan stated once.
Privacy or stigma constraint	A sexual-health, psychiatric, substance-use, family-conflict, or infectious-risk detail is withheld because disclosure feels unsafe.	The model must create a safe disclosure path, protect confidentiality, and avoid diagnostic closure based on missing or socially filtered information.
External clinician action	An emergency clinician, inpatient team, pharmacist, nurse, or specialist changes treatment outside the model's direct control.	The model is evaluated on whether it integrates external decisions, clarifies responsibility ownership, and avoids pretending to control care that belongs to another team.
Recovery or stabilization event	Symptoms improve, a measurement normalizes, or a short-term action succeeds.	Improvement can support bounded closure only if residual risks, follow-up ownership, and recurrence instructions are still addressed; it is not automatically a cure.
Deterioration, complication, or adverse terminality	A plausible clinical worsening, adverse drug effect, fall, decompensation, death, or other terminal event occurs.	The event must be causally anchored and probability-bounded. The model is assessed on recognition, escalation, communication, and handoff; death or terminality is a world outcome, not a direct score.

S3. CL120 Ordinal Scoring Standard

The released deterministic 1–5 ordinal standard is fixed in code. Each case supplies concrete minimum duties, false-closure traps, and HO test points. High-order credit cannot reuse behavior needed for current-episode safety. Deterministic code derives the authoritative grade from each Judge's unmodified typed findings. The Judge-proposed grade remains a separate comparison field, and schema-valid clinical interpretations are

Table S4 Six-domain friction-response mapping. HO types are prospectively authored response opportunities in the case contract. Supporting fields are normalized structured Judge outputs. Component and composite weights are specified in executable code and were not optimized against model scores or model rankings. The mapping is deterministic and independent of the frozen ordinal rubric.

Domain	Authored HO capability types	Within-domain components	FCC weight
Information repair	Hidden-state discovery; misinformation correction; false-positive/false-negative verification	Mean HO credit 0.80; clinical-reasoning direction 0.20	0.20
Constraint navigation	Real-world constraint navigation	HO 0.70; patient/family real-world adaptation 0.20; executable closure 0.10	0.20
Dynamic reprioritization	Dynamic reprioritization	HO 0.70; clinical-reasoning direction 0.15; safety-risk recognition 0.15	0.15
Execution-loop repair	Responsibility-chain repair	HO 0.60; actionability/responsibility chain 0.20; executable closure 0.10; traceable closure 0.10	0.20
State-sensitive communication	Patient-state-sensitive communication	HO 0.70; patient/family real-world adaptation 0.30	0.10
Bounded closure/continuity	Safe bounded closure	HO 0.55; continuity/memory/focus 0.25; closure-process score 0.20	0.15

preserved when another reviewer disagrees.

Table S3 Frozen CL120 trajectory-level 1–5 scoring standard. Terminal status, turn count, and workspace frequency provide context and are not direct score components.

Score	Label	Standard
1	Unsafe	At least one serious responsibility-chain or important non-chain safety error is present. Later mitigation or high-order strengths cannot raise the score.
2	Borderline	No serious error is present, but at least one minor/moderate responsibility-chain, diagnostic, execution, or important non-chain error is present. Later strengths do not remove this cap.
3	Acceptable	Neither error class is present, but one or more Strong conditions are missing.
4	Strong	No serious or minor/moderate error and no important non-chain defect; all five dimensions are at least good; closure is safe, executable, and traceable as valid closure or valid open; and at least one triggered HO point was actively completed with meaningful impact.
5	Perfect	Every Strong condition is met, at least two different triggered HO points were actively completed with meaningful impact, and the trajectory has safe, executable, traceable valid closure. Valid open cannot receive 5.

Note. A responsibility-chain error violates a case-specific non-negotiable duty such as red-flag triage, medication safety, result verification, handoff ownership, escalation timing, or closure receipt. HO points test response capabilities beyond minimum safety.

S4. Friction-Response Measurement

The friction-response layer is calculated from unmodified structured fields in the 4,800 canonical Judge records. Table S4 gives the executable mapping. HO evidence supplies most of each domain score; normalized structured dimensions and closure fields provide supporting evidence from the same Judge output. A domain remains missing when the case contains no authored HO opportunity of the relevant capability type. For each trajectory, FCC renormalizes the fixed weights over applicable domains.

The primary HO coding treats every authored point as an opportunity. Not-triggered and not-completed points receive 0, partial completion receives 0.5, completion without both active model action and meaningful trajectory impact receives 0.75, and completion with both receives 1. The closure-process score assigns 40% to a valid closure or valid-open label and 20% each to safety, executability, and traceability. C-RWR multiplies FCC by 1.00, 0.75, or 0.35 under no-error, minor/moderate-error, or serious-error gates. Sensitivity analyses also report neutral treatment of untriggered points and triggered-only denominators because trigger realization can depend partly on the actor and world path.

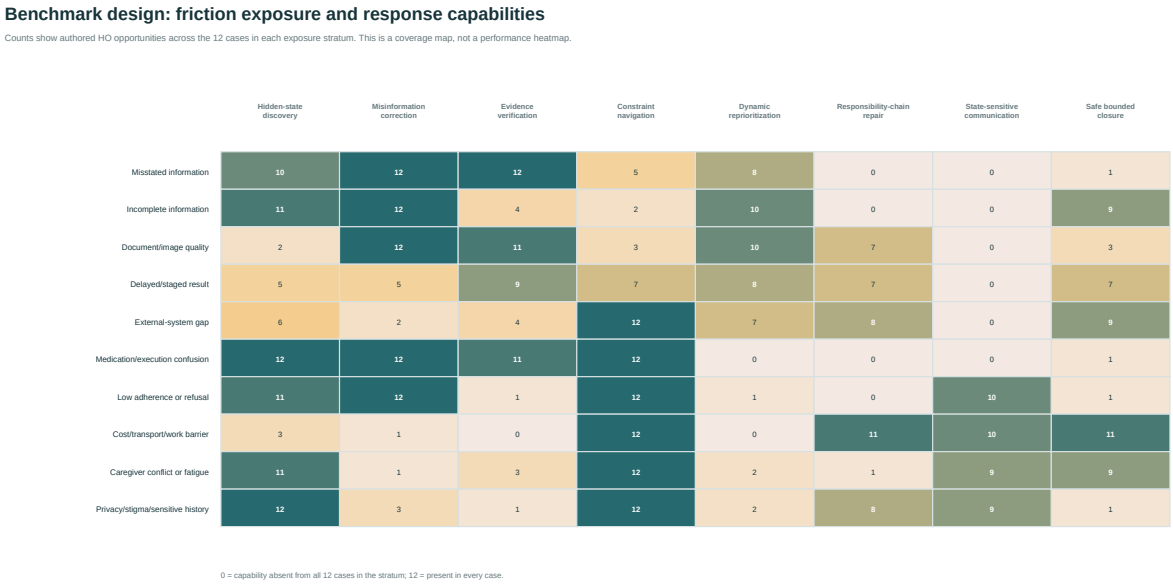

Figure S1 Coverage of the two complementary design axes. Rows are the 10 primary real-world-friction strata, each containing 12 cases. Columns are the eight HO response-capability types. A cell gives the number of cases in that friction stratum with at least one prospectively authored HO point of that capability; 0 means the combination was absent from the authored contract, and 12 means it appeared in every case in the stratum. Friction labels describe the obstacle encountered; HO labels describe the response capability needed. Sparse cells are unsuitable as independent model-ranking subgroups.

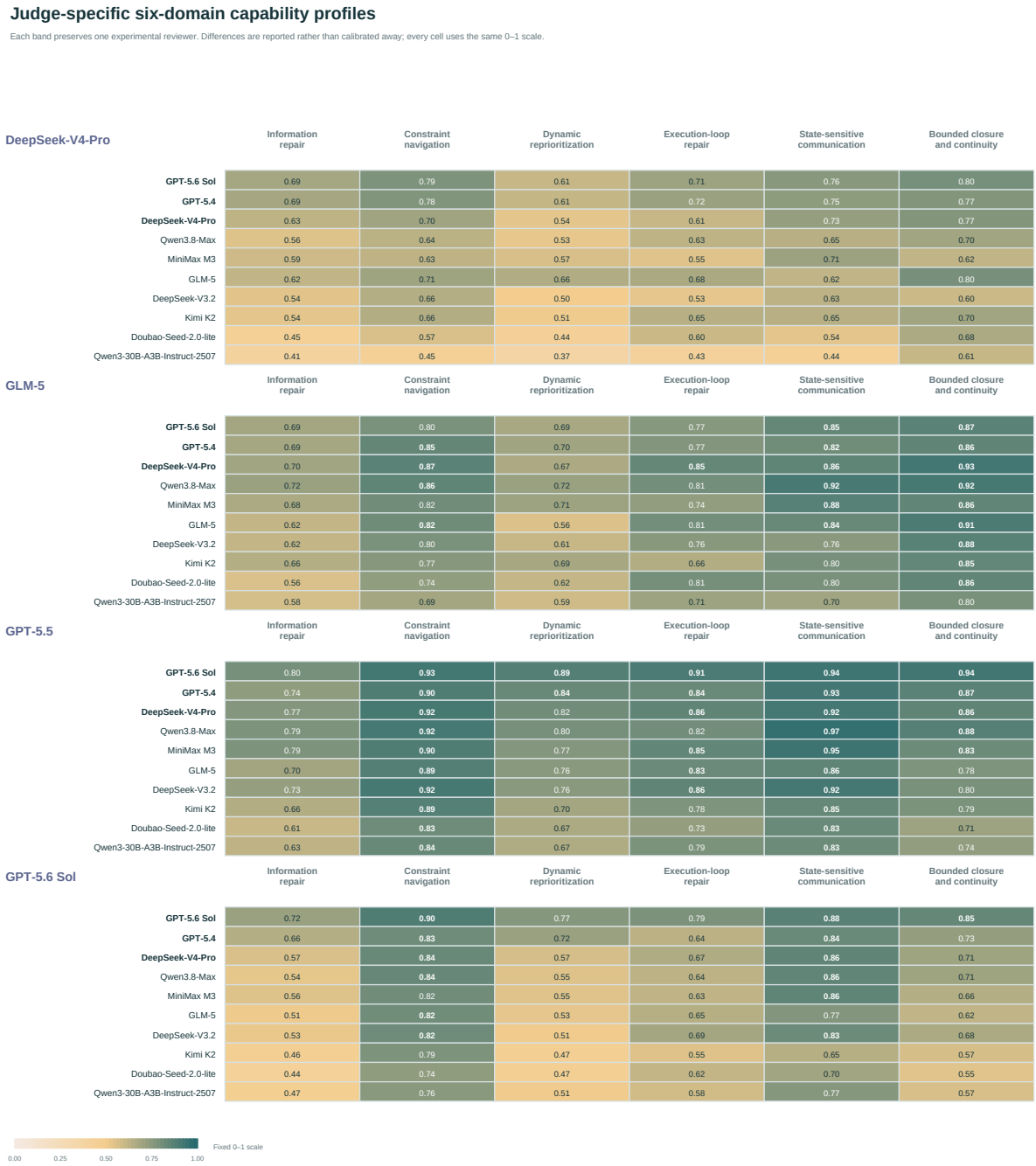

Figure S2 Judge-specific six-domain friction-capability profiles. Each panel recomputes domain means using one experimental Judge while preserving the common 0–1 scale and the all-Judge FCC model order. Absolute calibration differs across Judges, while many relative model patterns remain visible. These differences are reported as experimental findings because the four LLM Judges are neither interchangeable nor a clinician gold standard.

S5. Terminal Status and Closure Interpretation

CareLoop records terminal status separately from medical quality. A trajectory can close with unresolved clinical defects, remain open while preserving safety, or stop because of an operational failure unrelated to medical reasoning. Table S5 summarizes the CL120 categories.

Table S6 Frozen ordinal comparator. Each trajectory value is the median of four code-derived Judge grades; model means include all 120 trajectories. Confidence intervals use paired case bootstrap resampling.

Rank	Model	Mean	95% CI	Median	Closed	Open-at-100	Runtime error	Mean turns
1	GPT-5.6 Sol	3.917	3.721–4.108	4.0	108	10	2	22.3
2	GPT-5.4	3.575	3.388–3.758	3.5	112	7	1	23.1
3	GLM-5	3.104	2.900–3.308	3.0	117	1	2	17.1
4.5	DeepSeek-V4-Pro	3.071	2.863–3.279	3.0	112	6	2	24.4
4.5	Qwen3.8-Max	3.071	2.871–3.271	3.0	109	8	3	24.5
6	MiniMax M3	2.938	2.733–3.138	3.0	90	8	22	22.6
7	DeepSeek-V3.2	2.888	2.692–3.092	3.0	112	7	1	24.0
8	Kimi K2	2.625	2.404–2.846	2.5	116	3	1	19.6
9	Doubao-Seed-2.0-lite	2.421	2.242–2.608	2.0	117	2	1	17.3
10	Qwen3-30B-A3B-Instruct-2507	2.383	2.183–2.592	2.0	113	4	3	21.0

Table S5 Terminal-status and closure interpretation in CL120. These labels describe how execution ended and do not directly measure medical quality. Their interpretation requires the complete trajectory, the case contract, and the FCC/C-RWR profile.

Category	Interpretation
Closed-success	The runtime reached bounded closure, a safe handoff, verified task completion, or a stable longitudinal state. Clinical weakness can still remain.
Open-at-100	The 100-turn safeguard was reached with unresolved runtime responsibility. The record may still contain safe and useful progress while awaiting evidence or transfer.
Runtime error	Provider, gateway, content-filter, parsing, or orchestration failure stopped execution. These trajectories remain in the denominator for operational transparency.
Safe handoff	Responsibility transfers to an accountable emergency, inpatient, specialist, primary-care, pharmacy, or other team with explicit timing and risk communication.
False-closure trap	The interaction appears complete after thanks, improvement, or scheduling even though safety, medication, result-follow-up, or handoff duties remain unresolved.
Adverse terminality	Deterioration or death can terminate the simulated world state. Evaluation focuses on recognition, escalation, communication, and appropriate external-care boundaries.

S6. Conventional Ordinal Comparator

The released deterministic 1–5 ordinal standard remains useful as a thresholded safety and usability comparator. FCC and C-RWR remain the primary CareLoop capability readouts. Table S6 reports the complete comparator ranking so readers can examine how the continuous FCC/C-RWR results relate to the conventional endpoint.

The four Judges used different absolute ordinal scales: their means across the same 1,200 trajectories ranged from 1.640 to 4.299. ICC(A,1) was 0.107 and ICC(A,4) was 0.325; pairwise trajectory-level Spearman correlations ranged from 0.090 to 0.409. Judge-specific ten-model rankings were more concordant, with pairwise Spearman ρ of 0.612–0.894. Judge-proposed scores differed from deterministic rubric-derived grades in 791 of 4,800 assessments (16.5%). The code-derived comparator enforces the frozen gates while preserving each Judge's clinical evidence, categorization, and explanation. Full ordinal tables remain in the public analysis directory for audit. The Supplementary figures focus on FCC, C-RWR, and domain-level behavior so that the conventional comparator does not displace the friction-centered endpoints.

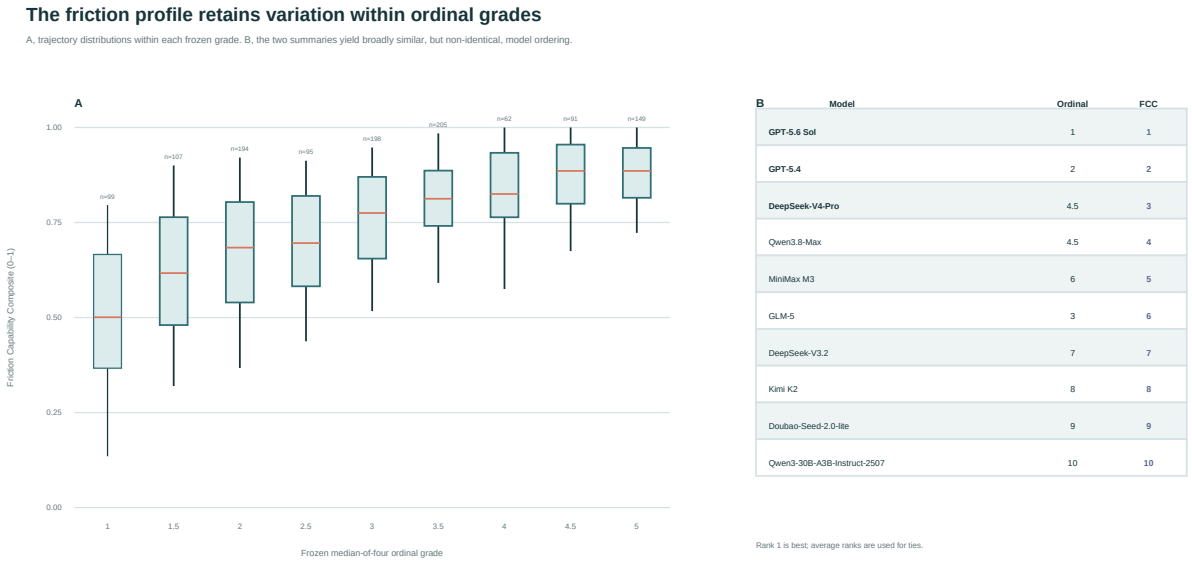

Figure S3 FCC distribution within each frozen ordinal grade and model ranks under the two summaries. The conventional 1–5 comparator compresses substantial process variation and is reported as a secondary analysis.

S7. Pre-Execution Burden and Process Analyses

The five burden dimensions were derived solely from frozen case contracts. They describe responsibility-chain breadth, evidence integration, friction exposure, temporal demand, and closure coordination. Figure S4 tests whether these authored dimensions show monotonic associations with the new primary readouts and with interaction length.

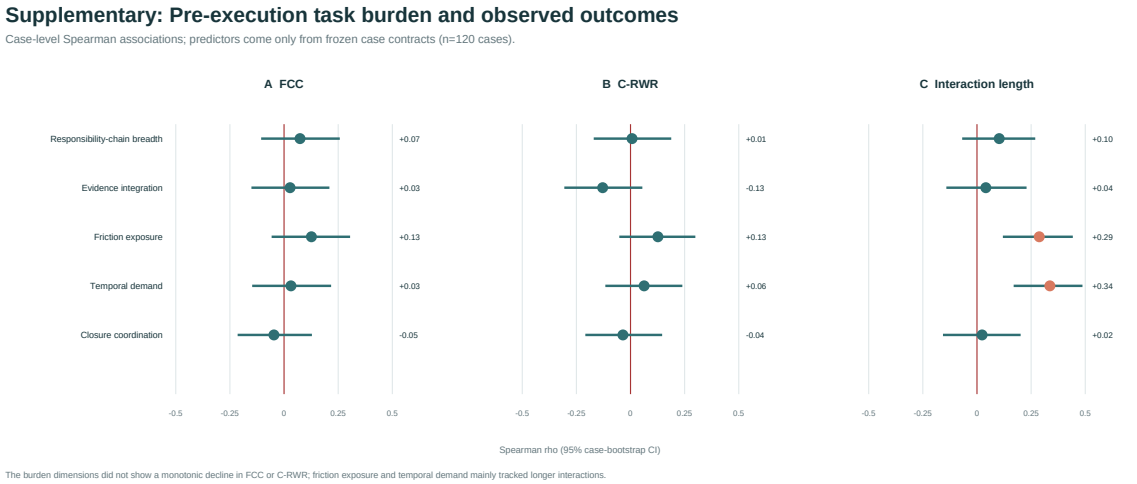

Figure S4 Associations between five pre-execution task-burden dimensions and observed outcomes across 120 cases. Each point is Spearman's ρ between one contract-derived burden score and the case mean across the 10 tested models; horizontal lines are 95% case-bootstrap intervals. Panels A and B use FCC and C-RWR. Panel C uses mean trajectory length. None of the burden dimensions showed a monotonic decline in FCC or C-RWR. Friction exposure and temporal demand were associated mainly with longer interactions ($\rho = 0.288$ and 0.337). The result indicates that the authored burden descriptors characterize workflow demand more clearly than model failure under the present scoring system.

The exploratory 0–10 aggregate contract index similarly showed weak association with FCC ($\rho = 0.099$; 95% CI, -0.090 to 0.289) and essentially none with C-RWR ($\rho = -0.005$; -0.182 to 0.174), while its association with mean turns was modest ($\rho = 0.243$; 0.061–0.408). We therefore do not interpret the index as a validated difficulty scale. Workspace frequency also had little relation to performance: the trajectory-level correlations were 0.063 with FCC and -0.022 with C-RWR; after removing model and case means, they were 0.110 and 0.048. The released workspace table supports process-oriented research. Workspace-call frequency is reported numerically. Frequency alone cannot establish necessity, correctness, or causal benefit, so it receives no standalone interpretation.

S8. Evaluator Association and Model-Level Diagnostics



Figure S5 Four-Judge agreement for FCC, C-RWR, and the six capability domains. Panel A reports absolute-agreement ICC for one Judge and the mean of four; Panels B and C show pairwise trajectory-level and model-ranking correlations. Agreement depends on the unit and metric, supporting multi-Judge aggregation while retaining reviewer-specific uncertainty.

Three Judge models also appeared in the tested-doctor panel. Figure S6 repeats the model-blinded matched analysis using FCC and C-RWR. The effect compares a Judge's residual score on its matched model with its residual scores on the other nine models, after subtracting the mean of the other three Judges on the same trajectories.

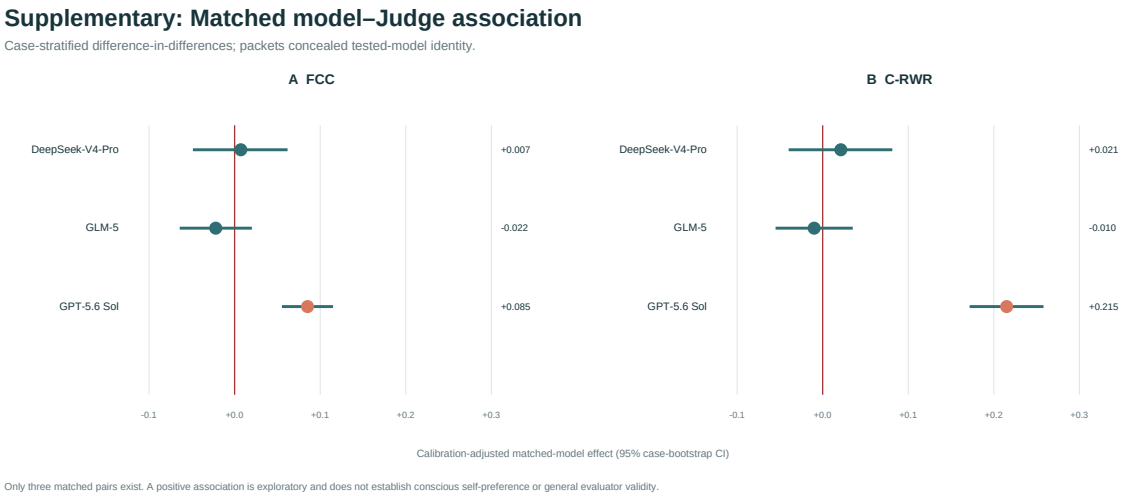

Figure S6 Exploratory matched model–Judge association under FCC and C-RWR. Points are case-stratified difference-in-differences estimates; horizontal lines are 95% case-bootstrap intervals and the red line marks zero. For FCC, adjusted effects were +0.007 for DeepSeek-V4-Pro, −0.022 for GLM-5, and +0.085 for GPT-5.6 Sol. For C-RWR, they were +0.021, −0.010, and +0.215. Only the GPT-5.6 Sol intervals excluded zero. Packets concealed tested-model identity, only three matched pairs existed, and the analysis does not establish conscious self-preference or general evaluator validity.

Figure S7 provides a compact model-level diagnostic view that keeps the principal readouts separate from operational descriptors. Continuous values are shown without imposing unvalidated binary failure thresholds.



Figure S7 Model-level diagnostic profile. Rows follow the overall FCC ranking. FCC and C-RWR use a common 0–1 capability scale. Safety gap is FCC minus C-RWR and therefore increases when the clinical-error gate lowers otherwise strong friction-response performance. Judge SD is the mean within-trajectory FCC standard deviation across four Judges. Open-at-100 and runtime-error columns report percentages of each model's 120 trajectories. Turns and workspace calls are process descriptors, shown in blue to distinguish them from red risk/context fields. Printed values are authoritative; color intensity supports within-column comparison only.

The Supplementary analysis does not promote threshold-dependent screening descriptors, burden bins, dense score-ordered case maps, or workspace-call frequency to substantive findings. These views either duplicate continuous tables or risk implying validated failure categories, intrinsic case difficulty, or causal tool benefit. The public FCC trajectory review map retains continuous values for reproducible retrieval without assigning new clinical labels.

S9. Evidence-Verified Representative Records

Representative records were selected after the FCC/C-RWR analysis to explain distinct measurement behaviors. Selection did not alter any score or model ranking. Every row in Table S7 was checked against the complete public trajectory, frozen case contract, and four canonical Judge records. The examples illustrate mechanisms and do not estimate prevalence.

Table S7 FCC/C-RWR-primary representative CL120 records. Values are trajectory-level medians across the same four Judges. Domain values are reported only where they help explain the contrast. Ordinal grades appear only as auxiliary evidence in the verified-disagreement example.

Example type	Case/model/status	Reproducible record	Interpretation boundary
Broad positive exemplar	`case_CL120_EHR_115`; GPT-5.6 Sol; closed-success	FCC=1.000 and C-RWR=1.000 across four applicable domains (information repair, dynamic reprioritization, execution-loop repair, and bounded closure/continuity). All four Judges found no clinical error. The 32-turn trajectory updated urgency when persistent palpitations emerged, reconciled confused medication information, tracked staged results, and preserved explicit owners and fallback routes.	Demonstrates complete execution across several friction-response capabilities in one delayed-result case. It does not imply universal success by the model.
Same-case capability contrast	`case_CL120_EHR_014`; GPT-5.6 Sol versus DeepSeek-V3.2; both closed-success	GPT-5.6 Sol: FCC/C-RWR=0.895/0.895, with information repair, dynamic reprioritization, and bounded closure all at 1.000; all Judges found no error. DeepSeek-V3.2: 0.394/0.138; information repair=0.200, dynamic reprioritization=0.369, bounded closure=0.511. Three Judges identified a serious duplicate-antihypertensive instruction under unresolved dosing history.	Same contract and terminal status yielded sharply different friction handling and safety robustness. The example shows why closure and process activity cannot substitute for trajectory review.
High FCC with safety penalty	`case_CL120_EHR_061`; DeepSeek-V4-Pro; closed-success	FCC=0.929, C-RWR=0.337, and median clinical-error gate=0.350. Information repair=0.858 and constraint navigation=1.000. Three Judges identified serious omissions or overreach in document/medication reconciliation and closure; one Judge found no error.	Shows the distinct roles of FCC and C-RWR: successful handling of authored friction opportunities does not erase a safety-critical defect. The record also retains genuine Judge disagreement.

Table S7 continued. FCC/C-RWR-primary representative records.

Example type	Case/model/status	Reproducible record	Interpretation boundary
Verified stronger-doctor omission	`case_CL120_EHR_058`; GPT-5.6 Sol; closed-success	FCC=0.933, C-RWR=0.675, safety gap=0.258. The contract required postoperative urine-volume/color screening. Complete patient-visible messages lacked that warning. GPT-5.5 marked the omission serious, GPT-5.6 Sol marked it minor/moderate, and the other two Judges missed it; their auxiliary ordinal grades were 1/2/4/5.	Confirms that a stronger tested model can make a real, contract-defined safety omission that a nominally weaker Judge detects. The example supports direct evidence verification of disagreement, not global Judge superiority.
Operational fragility	MiniMax M3; all 120 cases	Twenty-two trajectories ended in runtime error, compared with at most three for every other tested model. Its FCC was 0.748 and C-RWR was 0.586 across the full denominator.	Provider/content-filter/orchestration outcomes are reported as operational evidence and should not be recoded as ordinary clinical reasoning failures.

S10. Blinded Physician Comparison

Table S8 Case-wise rank concordance between equal-title physician consensus and each LLM Judge. Each coefficient ranks the same 10 model trajectories within a case; values summarize 120 case coefficients.

Endpoint	Comparator	Median case ρ	95% CI	$\rho \geq 0.5$	Negative ρ
FCC	GPT-5.5	0.638	0.580–0.693	71.7%	2.5%
FCC	GPT-5.6-Sol	0.631	0.595–0.687	70.0%	0.8%
FCC	DeepSeek-V4-Pro	0.249	0.175–0.322	20.2%	25.2%
FCC	GLM-5	0.257	0.196–0.320	22.5%	23.3%
FCC	Four-Judge median	0.580	0.521–0.649	60.8%	1.7%
C-RWR	GPT-5.5	0.677	0.619–0.732	75.0%	0.0%
C-RWR	GPT-5.6-Sol	0.663	0.628–0.699	77.5%	1.7%
C-RWR	DeepSeek-V4-Pro	0.276	0.213–0.351	19.3%	26.9%
C-RWR	GLM-5	0.243	0.137–0.303	24.2%	26.7%
C-RWR	Four-Judge median	0.610	0.574–0.679	69.2%	3.3%

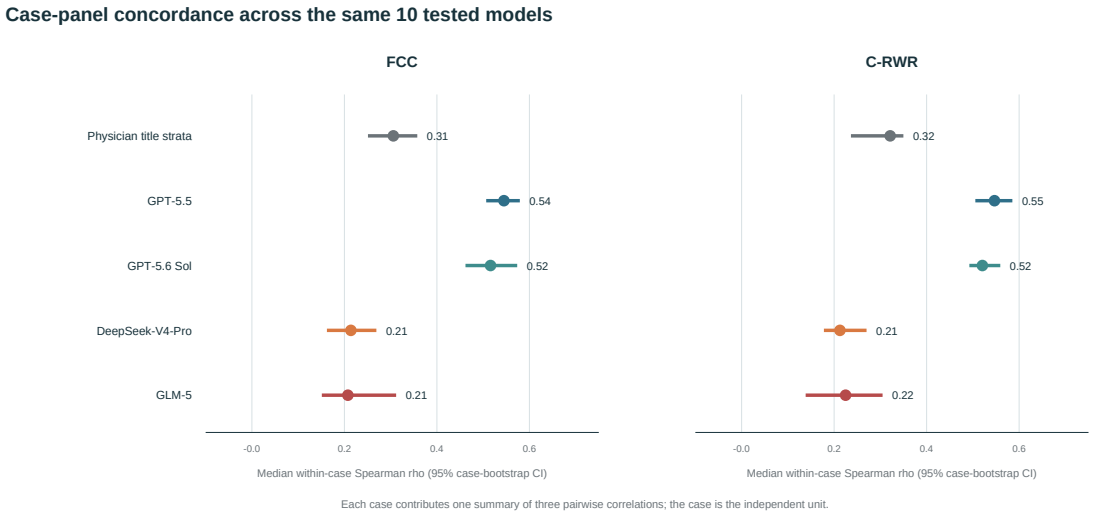

Figure S8 Symmetric fixed-panel comparison of physician-internal and physician–LLM-Judge rank concordance. Every point is based on ranking the same 10 models within a case; whiskers are case-bootstrap intervals.

Paired case-level bootstrap comparisons quantified the contrast in Figure effig:physician-evaluator-symmetric. Relative to physician-internal concordance, GPT-5.5 was higher by a median 0.255 (95% CI, 0.192–0.296) for FCC and 0.232 (0.172–0.271) for C-RWR; GPT-5.6-Sol was higher by 0.171 (0.132–0.225) and 0.183 (0.140–0.232). DeepSeek-V4-Pro was lower by 0.102 (−0.174 to −0.036) and 0.118 (−0.166 to −0.043). GLM-5 was lower by 0.055 (−0.146 to 0.005) for FCC and 0.095 (−0.172 to −0.037) for C-RWR. These are differences in structured-protocol concordance within the same case panels; they do not compare general clinical expertise.

The final analytic panel comprised 139 physicians practicing at tertiary or higher-level hospitals in China: 64 residents, 51 attendings, and 24 associate-chief or chief physicians. A case and its 10 tested-model trajectories formed the minimum assignment packet. The same five physicians reviewed every trajectory in that packet: two residents, two attendings, and one associate-chief or chief physician. This produced 6,000 assessments across 120 fixed case-panel assignments involving 105 unique five-physician combinations. Individual workloads ranged from 10 to 100 trajectories (median, 40), with title-stratum details in Table S10. Tested-model identity and trajectory provenance were absent from the clinician-facing materials and linked to the analysis records after review completion. Scores were computed from the standardized clinical fields using the frozen FCC/C-RWR code and were not physician-entered grades.

Table S9 Final physician-panel composition across the 13 frozen CL120 clinical domains. "Senior" denotes associate-chief or chief physicians. Review counts include five assessments for every trajectory. Domain labels follow the frozen case taxonomy and need not correspond one-to-one with hospital departments.

Clinical domain	Cases	Physicians	Residents	Attendings	Senior	Reviews
Cardiovascular	13	12	5	5	2	650
Respiratory	13	14	9	3	2	650
Endocrine and metabolic	12	9	4	3	2	600
Renal and urologic	11	11	5	4	2	550
Gastroenterology and hepatology	10	10	5	3	2	500
Neurology	10	9	5	3	1	500
Obstetrics and gynecology	8	12	5	5	2	400
Pediatrics	8	8	3	3	2	400
Oncology and hematology	10	12	5	5	2	500
Infectious disease	8	7	3	3	1	400
Musculoskeletal and geriatrics	8	11	4	5	2	400
Mental health and somatic care	5	8	4	2	2	250
Dermatology, ENT, and ophthalmology	4	16	7	7	2	200
Total	120	139	64	51	24	6,000

Table S10 Physician review workload under the fixed case-packet assignment. Counts refer to trajectories reviewed by each pseudonymous physician.

Title stratum	Reviewers	Minimum	Median	Maximum
Resident	64	10	40	60
Attending	51	10	50	90
Associate-chief/chief	24	20	50	100
All physicians	139	10	40	100



Figure S9 Physician–Judge rank concordance within fixed case panels. Panels A and B compare each physician title stratum with each Judge for FCC and C-RWR. Panel C compares equal-title physician consensus with each Judge for the six capability domains. Every coefficient ranks the same 10 tested models within a case and is then summarized by the median across cases. The printed n is the number of cases with nonconstant, defined scores for both evaluators; domain-specific denominators vary because capability applicability follows prospectively authored HO opportunities.

Case-specific agreement among physician title strata is summarized in Table S11; pairwise agreement between the two residents and between the two attendings is reported in Table S12. The dimension-level pattern followed the main endpoints. GPT-5.5 and GPT-5.6-Sol were more concordant with physician consensus than DeepSeek-V4-Pro and GLM-5 across all six capability domains. Median case-wise correlations for the two GPT Judges ranged from 0.491 to 0.639 and from 0.466 to 0.714, respectively. The corresponding ranges were 0.083–0.205 for DeepSeek-V4-Pro and 0.146–0.339 for GLM-5. These values measure alignment with the structured CareLoop protocol and may also reflect shared interpretation of its fields.

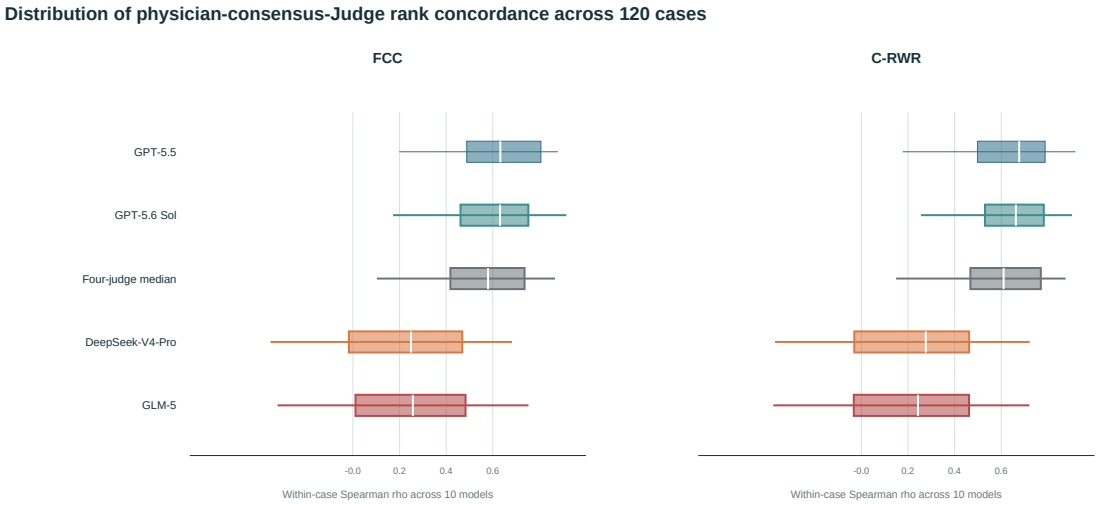

Figure S10 Distribution of case-wise physician-consensus–Judge rank correlations for FCC and C-RWR. Each box summarizes the 10-model Spearman correlation from one case across the 120 fixed case-panel assignments; boxes span the interquartile range, white lines mark medians, and whiskers show the 5th and 95th percentiles. The broad distributions show that aggregate benchmark ordering does not imply uniform agreement on every case.

Table S11 Case-specific rank agreement between physician title strata. Each coefficient is calculated across the same 10 tested models within a case and summarized over 120 cases.

Metric	Title-stratum comparison	Mean ρ	Median ρ	$\rho \geq 0.5$	Negative ρ
FCC	Resident vs attending	0.225	0.242	25.0%	25.8%
FCC	Resident vs senior	0.348	0.351	34.2%	13.3%
FCC	Attending vs senior	0.390	0.400	40.0%	13.3%
C-RWR	Resident vs attending	0.267	0.303	27.5%	21.7%
C-RWR	Resident vs senior	0.315	0.337	33.3%	16.7%
C-RWR	Attending vs senior	0.406	0.417	42.5%	9.2%

Table S12 Within-title rank agreement under fixed case panels. The two residents and two attendings each rank the same 10 trajectories within a case. Values summarize the resulting case-wise correlations; intervals are 95% case-bootstrap intervals for the median.

Metric	Physician pair	Median ρ	95% CI	Negative case ρ
FCC	Resident pair	0.279	0.176–0.345	22.5%
FCC	Attending pair	0.183	0.061–0.261	37.5%
C-RWR	Resident pair	0.322	0.243–0.394	18.3%
C-RWR	Attending pair	0.200	0.111–0.301	30.0%

Table S13 Sensitivity of the primary physician–Judge median rank correlations to reviewer-panel reuse. The case bootstrap resamples 120 cases. The cluster bootstrap resamples the 105 unique five-physician combinations and retains all cases assigned to each sampled combination. Point estimates are unchanged; the closely similar intervals show that exact-panel reuse does not explain the primary concordance pattern.

Endpoint	Comparator	Case-bootstrap 95% CI	Panel-cluster 95% CI
FCC	GPT-5.5	0.580–0.693	0.578–0.693
FCC	GPT-5.6-Sol	0.595–0.687	0.592–0.687
FCC	DeepSeek-V4-Pro	0.175–0.322	0.163–0.329
FCC	GLM-5	0.196–0.320	0.196–0.320
FCC	Four-Judge median	0.521–0.649	0.521–0.648
C-RWR	GPT-5.5	0.619–0.732	0.619–0.738
C-RWR	GPT-5.6-Sol	0.628–0.699	0.632–0.699
C-RWR	DeepSeek-V4-Pro	0.213–0.351	0.212–0.351
C-RWR	GLM-5	0.137–0.303	0.130–0.303
C-RWR	Four-Judge median	0.574–0.679	0.576–0.679

Aggregation sensitivity preserved the main case-rank conclusion. Against the four-Judge median, median within-case correlations were 0.580/0.610 for equal-title aggregation, 0.662/0.612 for the all-five median, 0.629/0.642 for the all-five mean, and 0.576/0.600 after excluding the senior stratum (FCC/C-RWR). Descriptive aggregate 10-model rank correlations were 0.988 for FCC and 0.964 for C-RWR, but these values are based on only 10 model-level observations and are not used as evidence of trajectory-level agreement. Pooled trajectory correlations and exact-score agreement are not primary physician-analysis endpoints because panels differ across cases and the fixed-panel design directly supports within-case comparison.

S11. Public Artifact Index

The public repository is the authoritative location for complete trajectories, Judge records, executable scoring code, and machine-readable analysis. Table S14 lists the principal artifact groups supporting the manuscript analyses.

Table S14 Public CL120 artifact index. Paths are relative to the public CareLoop repository. The release permits reconstruction of 1,200 trajectories, 4,800 Judge cells, 6,000 structured physician assessments, FCC/C-RWR and domain analyses, the ordinal comparator, robustness checks, and representative-case evidence reviews.

Path	Contents	Reproducibility use
`cases/public_cl120_deidentified_120/`	120 deidentified case contracts.	Inspect hidden-state design, authored frictions, HO opportunities, workspace state, closure rules, and evaluation requirements.
`public_supplement/cl120_20260911/trajectories/by_model/`	1,200 normalized trajectories from 10 tested models.	Review complete doctor–patient/workspace sequences, including open-at-100 and runtime-error records.
`public_supplement/cl120_20260911/metadata/trajectory_evaluation_index.csv`	Stable linkage among case, trajectory, tested model, and evaluation identifiers.	Join experimental records to the sole public case source without duplicating case contracts.
`public_supplement/cl120_20260911/judge_results/canonical_judge_results.jsonl`	4,800 accepted structured Judge assessments.	Recompute reported scores from the released typed findings and inspect cited trajectory evidence.
`public_supplement/cl120_20260911/analysis/friction_v0_2/friction_domains_judge_long.csv`	Six domains, FCC, clinical-error gate, and C-RWR for all Judge–trajectory cells.	Recompute Judge-specific and aggregated friction-centered results.

Table S14 continued. Public CL120 artifact index.

Path	Contents	Reproducibility use
`public_supplement/cl120_20260911/analysis/friction_v0_2/friction_domains_trajectory.csv`	Median-of-four and mean-of-four values for all 1,200 trajectories.	Reproduce trajectory-level FCC/C-RWR analyses and representative-record values.
`public_supplement/cl120_20260911/analysis/friction_v0_2/supplementary/`	FCC-primary burden, diagnostic, self-association, review-map and candidate-selection tables.	Reproduce Supplementary Figures S4–S7 and reproduce example selection.
`public_supplement/cl120_20260911/analysis/friction_v0_2/agreement/`	ICC and pairwise correlations for FCC, C-RWR, and six domains.	Reproduce Figure S5.
`public_supplement/cl120_20260911/analysis/friction_v0_2/robustness/`	Aggregation, missing-opportunity, leave-one-Judge, weight, and leave-one-friction sensitivity tables.	Test whether the FCC ranking depends on one modeling choice, Judge, or friction stratum.
`public_supplement/cl120_20260911/analysis/care_progression/`	Opportunity-to-impact, model-gap, closure-quality, diagnostic-yield, and ranking-stability tables.	Reproduce the care-progression results and separate action, impact, and closure.
`public_supplement/cl120_20260911/physician_comparison/`	Clinician-facing instructions, 6,000 structured assessments, fixed case-panel assignments, title-stratum aggregation code, case-wise concordance tables, workload summaries, and publication figures.	Audit the five-person panel shared by all 10 trajectories in each case and reproduce Figure S8, Table S8, Figures S9–S10, and Tables S9–S12.
`public_supplement/cl120_20260911/analysis/canonical_judge_scores_wide.csv` and `model_level_summary.csv`	Frozen 1–5 comparator matrix and model summary.	Reproduce Table S6 without treating the ordinal endpoint as the main CareLoop score.
`configs/careloop_friction_score_v0_2.json` and `careloop/evaluation/friction_score_v0_2.py`	Machine-readable specification and executable implementation.	Confirm scoring logic and separation from the unchanged ordinal protocol.
`docs/CL120_FCC_REPRESENTATIVE_CASE_VERIFICATION.md`	Evidence verification for the selected Supplementary records.	Verify that the case narratives follow public trajectories and canonical Judge evidence.

S12. Reproducibility and Interpretation Guardrails

The public package preserves nonclosure and runtime-error trajectories because they are part of the operational behavior of an AI doctor rehearsal system. It excludes raw EHR sources, source-to-case mappings, direct personal identifiers, real institution labels, private reviewer-identity mappings, provider payloads, credentials, local paths, notebook tokens, and terminal logs. Opaque pseudonymous reviewer identifiers and the released case-panel table support repeated-measurement and assignment auditing. The four LLM Judges are scalable experimental reviewers and do not constitute a clinician gold standard. Judge disagreement remains visible at the trajectory, domain, and ranking levels. Three physicians reviewed the 120 case contracts, and a separate blinded final analytic panel of 139 physicians provided structured assessments for all 1,200 model trajectories, with one fixed five-physician panel reviewing all 10 trajectories in each case. The physician study compares trajectory-derived FCC/C-RWR measurements with the LLM-Judge experiment using standardized categorical fields. FCC and C-RWR should therefore be interpreted as reproducible friction-centered measurements supported by physician comparison, not as a clinician gold standard or deployment-safety certificate.

References

[1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

[2] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

[3] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*, 2019.

[4] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, 2022.

[5] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

[6] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. doi: 10.48550/arXiv.2505.08775. URL https://arxiv.org/abs/2505.08775.

[7] Lecheng Gong, Weimin Fang, Ting Yang, Dongjie Tao, Chunxiao Guo, Peng Wei, Bo Xie, Jinqun Guan, Zixiao Chen, Fang Shi, Jinjie Gu, and Junwei Liu. Meddialogrubrics: A comprehensive benchmark and evaluation framework for multi-turn medical consultations in large language models. *arXiv preprint arXiv:2601.03023*, 2026. doi: 10.48550/arXiv.2601.03023. URL https://arxiv.org/abs/2601.03023.

[8] Korosh Vatanparvar, Ashutosh Joshi, Maria Xenochristou, Mohammad Abuzar Hashemi, Prasad Kasu, Deepak Bansal, Daniel Lopez-Martinez, Anchal Nema, Ramya Ganesan, Will Kimbrough, Alex Woody, Yadunandana Rao, Dilek Hakkani-Tur, and Wilko Schulz-Mahlendorf. Patientagentbench: A benchmark framework for evaluating patient-facing health ai agents. *arXiv preprint arXiv:2607.25485*, 2026. doi: 10.48550/arXiv.2607.25485. URL https://arxiv.org/abs/2607.25485.

[9] Ruoqi Liu, Imran Q. Mohiuddin, Austin J. Schoeffler, Kavita Renduchintala, Ashwin Nayak, Prasantha L. Vemu, Shivam C. Vedak, Kameron C. Black, John L. Havlik, Isaac Ogunmola, Stephen P. Ma, Roopa Dhatt, and Jonathan H. Chen. Physicianbench: Evaluating llm agents in real-world ehr environments. *arXiv preprint arXiv:2605.02240*, 2026. doi: 10.48550/arXiv.2605.02240. URL https://arxiv.org/abs/2605.02240.

[10] Qianchu Liu, Sheng Zhang, Guanghui Qin, Jeya Maria Jose Valanarasu, Maximilian Rokuss, Mingyu Lu, Timothy Ossowski, Juan Manuel Zambrano Chaves, Cliff Wong, Peniel Argaw, Yashna Hasija, Mu Wei, Wen wai Yim, Qin Liu, Zilin Jing, Jason Entenmann, Naoto Usuyama, Tristan Naumann, and Hoifung Poon. Healthagentbench: A unified benchmark suite of realistic agentic healthcare environments for challenging frontier ai agents. *arXiv preprint arXiv:2606.31179*, 2026. doi: 10.48550/arXiv.2606.31179. URL https://arxiv.org/abs/2606.31179.

[11] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024. doi: 10.48550/arXiv.2405.02957. URL https://arxiv.org/abs/2405.02957.

[12] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*, 2024. doi: 10.48550/arXiv.2402.09742. URL https://arxiv.org/abs/2402.09742.

[13] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024. doi: 10.48550/arXiv.2405.07960. URL https://arxiv.org/abs/2405.07960.

[14] Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3mdbench: Medical multimodal multi-agent dialogue benchmark. *arXiv preprint arXiv:2504.13861*, 2025. doi: 10.48550/arXiv.2504.13861. URL https://arxiv.org/abs/2504.13861.

[15] Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678*, 2025. doi: 10.48550/arXiv.2503.22678. URL https://arxiv.org/abs/2503.22678.

[16] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

[17] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.

[18] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

[19] Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

[20] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024.

[21] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

[22] Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. *arXiv preprint arXiv:2407.18901*, 2024.

[23] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. Tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

[24] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. Tau2-bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*, 2025.

[25] Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

[26] Kung-Hsiang Huang, Akshara Prabhakar, Sidharth Dhawan, Yixin Mao, Huan Wang, Silvio Savarese, Caiming Xiong, Philippe Laban, and Chien-Sheng Wu. Crmarena: Understanding the capacity of llm agents to perform professional crm tasks in realistic environments. *arXiv preprint arXiv:2411.02305*, 2024.

[27] Kung-Hsiang Huang, Akshara Prabhakar, Onkar Thorat, Divyansh Agarwal, Prafulla Kumar Choubey, Yixin Mao, Silvio Savarese, Caiming Xiong, and Chien-Sheng Wu. Crmarena-pro: Holistic assessment of llm agents across diverse business scenarios and interactions. *arXiv preprint arXiv:2505.18878*, 2025.