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

Anonymous authors
Paper under double-blind review

Abstract

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; prospectively authored frictions 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. CareLoop allows patients' lived contexts to alter subsequent trajectories and recovers stable comparisons by aggregating structured observations across broad case coverage. Anonymous code and data: https://anonymous.4open.science/r/careloop-iclr2027-6e55. The simulation does not establish clinical effectiveness or deployment safety.

1 Introduction

A model can state a medically appropriate answer and still fail to practice medicine with a person. Symptoms may remain unspoken, patients and caregivers may hold conflicting beliefs, reports may be unreadable or delayed, and cost, transport, privacy, refusal, or family dynamics may make the nominal plan impossible. Care can stall after a correct recommendation when no one verifies execution, incorporates returned evidence, or owns unresolved risk. We call the resulting difference between medical knowledge and useful action in a patient's lived situation the knowledge-to-practice gap.

Existing medical benchmarks measure valuable components of competence, including factual knowledge, diagnosis, communication, safety, tool use, and workflow execution (Hendrycks et al., 2021; Jin et al., 2020; 2019; Pal et al., 2022; Singhal et al., 2023; Arora et al., 2025). Interactive environments add patients, health records, tools, and longitudinal actions (Schmidgall et al., 2024; Li et al., 2024; Vatanparvar et al., 2026; Liu et al., 2026b;a).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

General-agent benchmarks likewise show that plausible language is insufficient when success depends on verifiable changes to websites, applications, codebases, databases, or shared worlds (Zhou et al., 2023; Xie et al., 2024; Jimenez et al., 2023; Trivedi et al., 2024; Yao et al., 2024; Barres et al., 2025). CareLoop asks how execution-grounded evaluation changes in a partially observable socio-technical world. Task-relevant state is distributed among biology, beliefs, records, and execution; people retain control over disclosure and action; and new evidence can alter both the feasible path and the appropriate goal. The resulting evaluation problem centers on agency under limited control.

We introduce CareLoop, an executable clinical-world sandbox designed to make this gap observable. Each case is a partially observable contract that separates underlying clinical state, lived belief, recorded evidence, and executed action. Patients and family members act statefully, with their own beliefs, choices, and constraints. Delayed results, unreliable documents, medication confusion, access barriers, refusal, caregiver conflict, and broken handoffs create controlled pressure points. The tested model must obtain missing state, communicate responsively, act through a clinical workspace, observe what actually happened, revise its plan, and either close safely or preserve accountable continuity.

The evaluated object is therefore the evolving trajectory. A response matters when it changes what happens next in a clinically meaningful way. A completed workflow can remain unsafe, while an open episode can be appropriate when pending evidence has an owner, timing, verification path, and escalation condition—a form of accountable openness. The environment separately represents attempted action, realized impact, runtime termination, and clinical closure.

Our contributions are:

- A patient-centered task definition. CareLoop represents care as agency under limited control, with distributed information, patient autonomy, delayed evidence, changing goals, and accountable openness.
- A capability-eliciting environment. Prospectively authored friction and high-order opportunities force latent deficiencies to acquire observable consequences across complete trajectories.
- A process-centered measurement system. Six domains, FCC, safety-gated C-RWR, and an action-to-impact decomposition characterize how models handle friction rather than treating terminal success or a single rubric grade as sufficient.
- An evaluation at scale. We release 120 cases, 1,200 model trajectories, 4,800 blinded LLM-Judge assessments, 6,000 blinded physician assessments, and executable analyses. The results reveal substantial local evaluator uncertainty alongside stable benchmark-level model ordering.

2 Related Work

Medical evaluation. Question-answering and open-ended clinical benchmarks establish whether models possess and communicate medical knowledge (Jin et al., 2019; 2020; Pal et al., 2022; Singhal et al., 2023; Arora et al., 2025; Gong et al., 2026). AgentClinic and multi-agent hospital simulations add diagnosis, treatment, multimodal observations, and institutional interaction (Schmidgall et al., 2024; Li et al., 2024; Fan et al., 2024; Sviridov et al., 2025; Almansoori et al., 2025). PatientAgentBench evaluates patient-facing dialogue and service execution; PhysicianBench tests clinician-facing EHR workflows with executable checkpoints; HealthAgentBench spans heterogeneous end-to-end healthcare workflows (Vatanparvar et al., 2026; Liu et al., 2026b;a). CareLoop complements these efforts by treating real-world friction as an authored experimental variable and care progression through friction as the primary unit of analysis.

Executable and shared-world agent benchmarks. WebArena, WorkArena, OSWorld, SWE-bench, and AppWorld move beyond answer quality by evaluating changes to persistent external state (Zhou et al., 2023; Drouin et al., 2024; Xie et al., 2024; Jimenez et al., 2023; Trivedi et al., 2024). In τ-bench, an agent serves a simulated user while following poli-

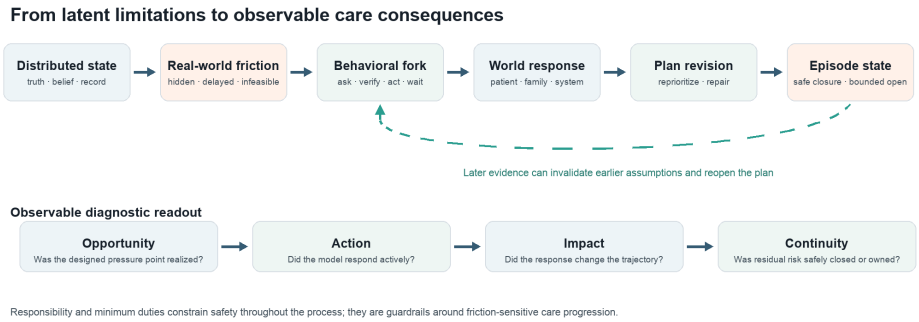


Figure 1: CareLoop turns real-world friction into diagnostic pressure. Distributed state creates behavioral forks at which the doctor-role model must discover, verify, communicate, act, track, or revise. The simulated world then changes independently. Structured review distinguishes whether an opportunity arose, whether the model acted, whether the action mattered, and whether residual risk was safely closed or retained under accountable continuity.

cies and modifying a database; τ^2-bench gives both agent and user control over the shared world, exposing coordination and verification failures (Yao et al., 2024; Barres et al., 2025). TheAgentCompany and CRMArena/Pro extend this approach to consequential organizational work and professional service systems (Xu et al., 2024; Huang et al., 2024; 2025). Together, these environments establish the value of execution-grounded evaluation. In formalizable worlds such as repository repair, an agent can often modify task-relevant state directly, and comparatively determinate external criteria can verify the result. CareLoop carries the same demand for consequences into a clinical world where state is distributed across biology, beliefs, records, and execution; independent actors shape outcomes; evidence and goals can change; and safe completion may require an accountable open state. Evaluation consequently shifts from state change under substantial control to consequential trajectories under limited control. Appendix A provides a structured comparison.

3 CareLoop

3.1 Clinical worlds with distributed state

At time t, a case contains four related states:

$$S_t = \left(S_t^{\text{truth}}, S_t^{\text{belief}}, S_t^{\text{record}}, S_t^{\text{exec}}\right).$$

They represent underlying clinical state, what patients or family members believe, what is recorded or returned as evidence, and what has actually been executed. The doctor-role model observes a governed projection through dialogue and a clinical workspace. It can ask, communicate, use tools, or initiate external processes; patient, family, and system actors determine what is disclosed, returned, accepted, refused, delayed, or completed.

A friction creates or preserves a consequential mismatch among these states, or obstructs movement from an appropriate intention to an executed result. A care-progression event occurs when the model obtains relevant state, chooses a safe and feasible action, observes or seeks its consequence, revises as needed, and constrains residual risk. Safe closure is case-specific and allows bounded openness when unresolved obligations remain explicitly owned and monitored.

3.2 How the environment elicits failure

Each case contract specifies the opening presentation, hidden clinical state, actor-specific knowledge and behavior, workspace evidence, disclosure and reliability rules, delayed events,

friction lifecycle, external actions, minimum duties, high-order (HO) opportunities, false-closure traps, and bounded closure criteria. Ten friction strata cover misstated information, missing information, unreliable documents, delayed results, external-system gaps, medication execution, refusal, material constraints, caregiver conflict, and privacy or stigma. Four HO opportunities per case specify responses whose necessity may become realized during interaction.

The design creates an observable chain:

$$\begin{aligned}\text{authored opportunity} &\to \text{realized pressure} \to \text{model action} \\ &\to \text{world impact} \to \text{completion/continuity.}\end{aligned}$$

A model can state the correct principle yet fail to elicit the fact that makes it relevant, recommend an action without verifying execution, recognize a barrier without finding a feasible route, or end the encounter while risk remains unowned. Because the world proceeds after the initial answer, these failures affect later state rather than remaining stylistic differences.

3.3 Measurement

Four model-blinded Judges review complete, ordered trajectory packets containing the case-specific evaluation contract, dialogue, and workspace evidence. Each records typed findings for responsibility-chain items, HO opportunities, five clinical-usability dimensions, closure, and non-chain defects. The same structured constructs are used for physician review.

Six 0–1 capability domains summarize information repair, constraint navigation, dynamic reprioritization, execution-loop repair, state-sensitive communication, and bounded closure/continuity. For Judge j and trajectory i, FCC is a fixed weighted mean over domains authored for that case:

$$\mathrm{FCC}_{ij} = \frac{\sum_{d \in D_i} w_d s_{ijd}}{\sum_{d \in D_i} w_d},$$

with weights $(0.20, 0.20, 0.15, 0.20, 0.10, 0.15)$. C-RWR retains the consequence of structured clinical-error findings:

$$\text{C-RWR}_{ij} = \mathrm{FCC}_{ij}\, g_{ij}, \qquad g_{ij} \in \{1.00, 0.75, 0.35\}$$

for no, minor/moderate, or serious clinical error. The primary trajectory score is the median across four Judges; model performance averages 120 trajectory scores. The frozen 1–5 rubric remains a supplementary thresholded comparator. Appendix B gives the complete mapping and robustness analyses.

4 Experimental Design

Cases and tested models. CL120 contains 120 deidentified EHR-derived case contracts, balanced across 10 primary friction strata. Three licensed physicians reviewed candidate drafts case by case; 130 cases received unanimous approval, and 120 were frozen to improve relative clinical-domain balance. We evaluated 10 doctor-role model endpoints on every case, producing 1,200 trajectories under the same runtime, workspace, contracts, and 100-turn safeguard. Closed-success, open-at-100, and runtime-error trajectories all remained in the denominator.

LLM Judges. GPT-5.5, GPT-5.6 Sol, DeepSeek-V4-Pro, and GLM-5 each evaluated all 1,200 trajectories, yielding 4,800 complete-context assessments. Packets concealed tested-model identity and outcome-derived summaries. Transport, truncation, parse, and schema failures were retried; schema-valid substantive judgments were preserved without semantic repair. Confidence intervals use 10,000 case-paired bootstrap resamples. We tested leave-one-Judge-out aggregation, equal and perturbed domain weights, omission of friction strata, and alternative handling of untriggered HO opportunities.

Table 1: Primary friction-centered results. Scores average 120 trajectory-level medians across four Judges; intervals use 10,000 case-paired bootstrap resamples.

	Model	FCC [95% CI]	C-RWR [95% CI]
1	GPT-5.6 Sol	.824 [.799,.849]	.753 [.719,.786]
2	GPT-5.4	.786 [.758,.813]	.680 [.645,.713]
3	DeepSeek-V4-Pro	.783 [.755,.810]	.621 [.583,.659]
4	Qwen3.8-Max	.773 [.743,.801]	.619 [.581,.657]
5	MiniMax M3	.748 [.715,.779]	.586 [.545,.626]
6	GLM-5	.740 [.709,.770]	.598 [.559,.638]
7	DeepSeek-V3.2	.738 [.707,.768]	.568 [.530,.603]
8	Kimi K2	.694 [.657,.729]	.510 [.467,.552]
9	Doubao-Seed-2.0-lite	.651 [.616,.686]	.471 [.431,.511]
10	Qwen3-30B-A3B	.645 [.611,.678]	.450 [.409,.492]

Physician comparison. A total of 139 physicians from tertiary or higher-level hospitals provided 6,000 blinded structured assessments. For each case, one fixed panel of five domain-matched physicians—two residents, two attending physicians, and one associate-chief/chief physician—reviewed all 10 model trajectories. This makes the within-case 10-model ordering the primary comparison unit and controls case content and panel composition while the tested model changes. We report case-wise Spearman correlations, title-stratum agreement, Top-k overlap, aggregate rankings, and panel-cluster sensitivity. Physician judgments are an independent comparison panel, not an error-free gold standard.

Scope. All experiments evaluate behavior inside a simulation sandbox. They do not test clinical effectiveness, patient outcomes, or deployment safety. Full protocol, model versions, task taxonomies, statistical definitions, and artifact counts appear in the appendix and anonymous repository.

5 Results

5.1 The knowledge-to-practice gap separates models

Table 1 reports the primary results. GPT-5.6 Sol ranked first on 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, while C-RWR separated GPT-5.4 more clearly because clinical-error gates penalized the latter two systems. The FCC-to-C-RWR penalty ranged from 0.072 for GPT-5.6 Sol to 0.195 for Qwen3-30B. A clinical error can follow from missing state, failed verification, stale planning, incomplete execution, unsafe closure, or factual error; the penalty should not be read as a pure knowledge score.

The strongest separation occurred in information repair (between-model range 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. Pairwise trajectory-level domain correlations ranged from 0.228 to 0.714, showing that the domains are related without collapsing to a single behavior. The model–capability profile in Figure 2 shows that stronger models gained across several capabilities rather than from one isolated skill.

5.2 The environment exposes failure between action and consequence

The same fixed case produced substantial divergence across models: the median within-case FCC range was 0.414; 89 of 120 cases exceeded 0.30 and 38 exceeded 0.50. Across 15,537 Judge assessments in which an HO opportunity was realized, a model took active action in 0.935 (95% CI, 0.924–0.945), achieved meaningful trajectory impact in 0.745 (0.723–0.766), and completed the opportunity effectively in 0.697 (0.673–0.720). The action-to-impact gap was 0.190. 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

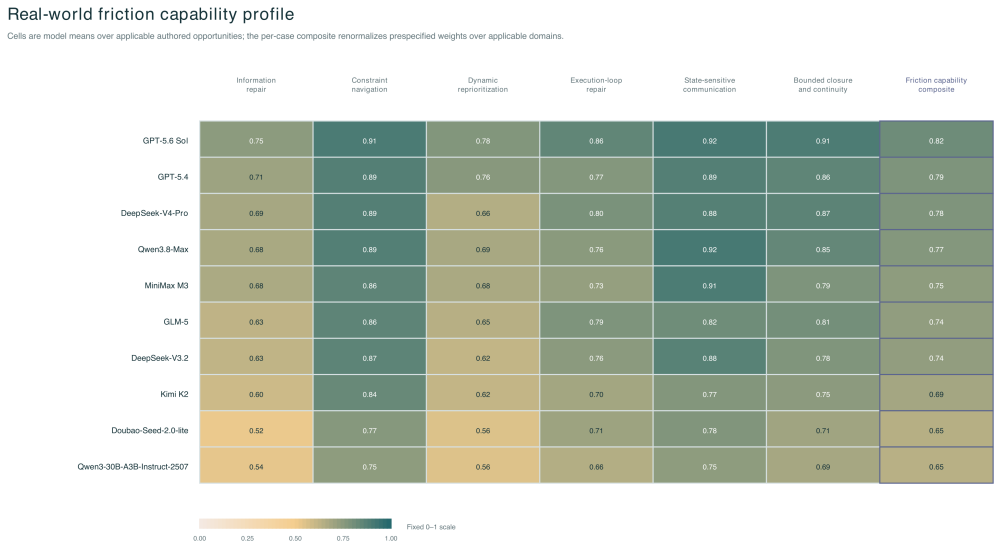


Figure 2: Six-domain capability profile and FCC. Cells average cases for which the domain was prospectively authored.

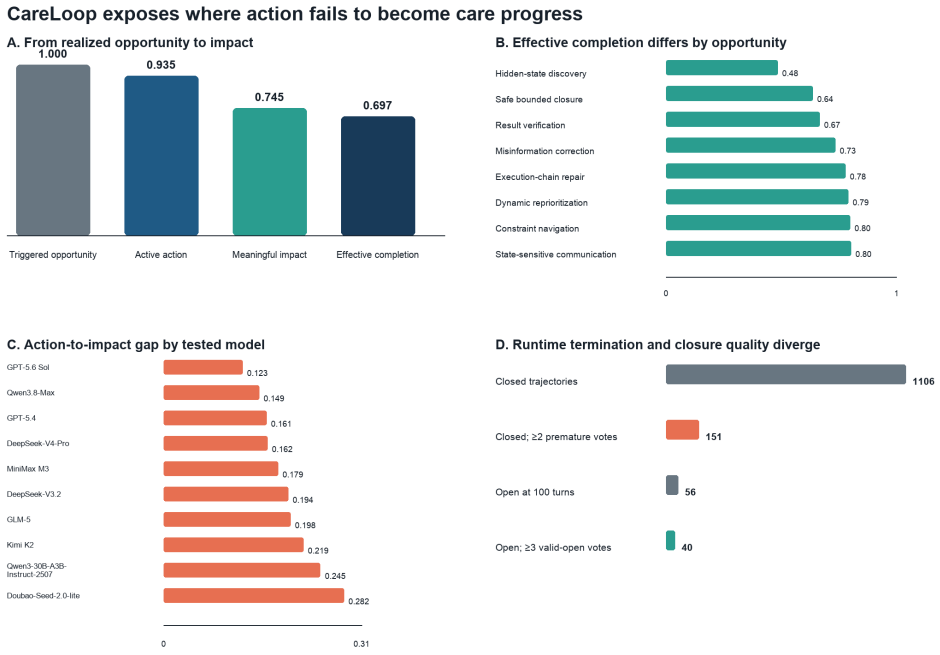


Figure 3: From response to care progression. Panels show aggregate conversion across realized HO opportunities, completion by opportunity type, each model's action-to-impact gap, and the difference between runtime termination and semantic closure.

completion with a 0.266 gap. The model-level action-to-impact gap ranged from 0.123 for GPT-5.6 Sol to 0.282 for Doubao-Seed-2.0-lite.

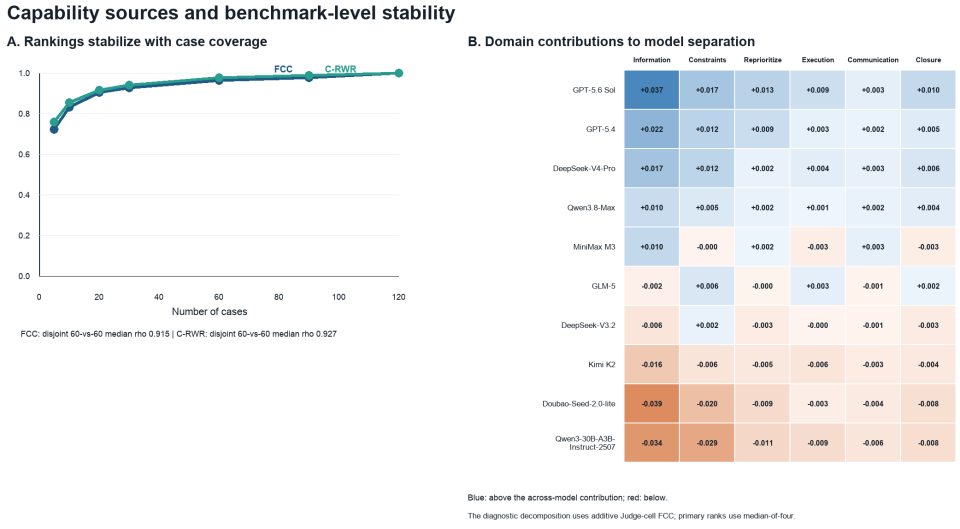


Figure 4: Benchmark-level stability and sources of model separation. Left: model rankings converge with case coverage and remain similar across disjoint 60-case halves. Right: additive domain contributions relative to the across-model mean.

Runtime status was correspondingly unreliable as a quality endpoint. Among 56 trajectories open at turn 100, 40 were judged valid-open by at least three Judges. Among 1,106 trajectories marked closed_success, 437 received at least one premature-closure judgment and 151 received at least two. Across 10 model points, closed-success rate correlated -0.652 with FCC, while the Judge-derived closure-process score correlated 0.927. Finishing the interaction and closing the clinical responsibility unit are different events.

5.3 Uncertain local judgments yield stable benchmark-level comparisons

Single-trajectory agreement was modest. For FCC, single-Judge absolute-agreement ICC was 0.257 and four-Judge-mean ICC was 0.581; C-RWR values were 0.211 and 0.517. Within a case, the median correlation among the three physician title-stratum rankings was 0.306 for FCC and 0.321 for C-RWR. GPT-5.5 aligned with physician strata at 0.545/0.547 and GPT-5.6 Sol at 0.516/0.521 for FCC/C-RWR; DeepSeek-V4-Pro and GLM-5 were at or below physician-internal concordance. These comparisons measure consistency with the shared protocol, not clinical superiority.

Aggregation changed the picture. Random 60-case subsets correlated 0.964 with the complete FCC ranking and 0.976 with the complete C-RWR ranking. Across 5,000 disjoint, friction-stratified 60-case splits, median half-to-half correlations were 0.915 (95% simulation interval, 0.794–0.976) and 0.927 (0.818–0.988), with Top-1 agreement of 99.2% and 100%. Equal-title physician consensus and the four-Judge median produced closely aligned aggregate 10-model rankings over all cases ($\rho = 0.988$ for FCC and 0.964 for C-RWR). Figure 4 summarizes the convergence and the sources of model separation.

The broad ordering survived all tested analytical perturbations. Leave-one-Judge-out rankings correlated 0.939–0.988 with the four-Judge result; equal weights and every one-at-a-time $\pm 20\%$ weight perturbation preserved the exact ordering; leaving out one friction stratum yielded correlations of 0.964–1.000. Matched model–Judge analyses found no uniform self-favoring pattern. The same model ordering is therefore visible across heterogeneous evaluators and analytical choices, even though individual trajectory labels remain uncertain.

6 Discussion

What CareLoop measures. Model scores diverged most where medical knowledge had to become clinical progress: discovering hidden state, revising after the world changed, making actions consequential, and keeping unresolved risk under accountable continuity. The simulator exposes these abilities by allowing patient beliefs, choices, constraints, documents, delayed evidence, and external execution to alter the path. The resulting trajectories explain performance gaps that static answers and nominal task-completion rates cannot localize.

From executable worlds to open human worlds. Executable-agent benchmarks advanced evaluation by grounding success in changes to external state. CareLoop retains that standard while changing the world represented by the evaluator. In SWE-bench, a patch directly changes a digital repository, and executable tests provide a strong, comparatively determinate oracle. In CareLoop, recommendations may not be adopted, orders may not lead to completed tests, referrals may not produce handoffs, and explanations may not produce understanding. Biological state, actor belief, records, and execution can diverge while patients, families, and institutions retain independent agency.

Evaluation in such settings must follow the trajectory from discovery and action through verification, consequence, revision, repair, and continuity. A terminal state can conceal preventable delay or premature closure. An unresolved episode can still be managed responsibly when risk has an owner, time boundary, verification path, and escalation condition. We call this perspective consequence stewardship: evaluating whether an agent remains accountable for the effects and unfinished obligations of its interventions under partial observability and limited control. The term describes a property measured within the simulation and makes no claim of clinical effectiveness or autonomous safety.

Implications for evaluation. Physicians and LLM Judges both showed substantial disagreement on individual long trajectories. Rubric-style human labels provide informative clinical measurements, although they do not furnish an unquestionable gold standard. Broad case coverage and repeated structured review reduced local instability and recovered stable model comparisons. In this setting, strong general-purpose models can support scalable evaluation when reviews are model-blinded, structurally constrained, replicated across Judges, compared with physician panels, and accompanied by sensitivity analysis.

Matched evaluator analyses showed no systematic tendency for a Judge to promote trajectories produced by its own model family. Other evaluator biases remain possible, yet the result provides evidence of useful reflexivity: capable models can inspect model-generated clinical behavior, including closely related behavior, without a dominant self-preference signal.

Limitations. CL120 is a finite simulation covering bounded specialties, institutions, populations, and care paths. The present implementation is text-mediated; future versions could extend its actor-, state-, and event-based design to governed images, physiological waveforms, speech, video, and device streams while retaining trajectory-level tests of whether multimodal evidence changes downstream action. Friction strata contain different clinical content and are not a factorial causal experiment. Semantic review remains necessary for many outcomes, and both physician and LLM assessments carry protocol-dependent uncertainty. The study evaluates simulation behavior, not patient outcomes, clinical effectiveness, or autonomous deployment safety. Future work should add controlled counterfactual cases, repeated generations per model, multilingual and cross-institutional settings, stronger deterministic receipts, and carefully governed prospective comparisons.

7 Conclusion

CareLoop begins from a simple premise: knowing medicine and practicing it with a living person are distinct capabilities. By placing models in a world where information, beliefs, records, actions, and responsibilities can diverge, it makes the knowledge-to-practice gap observable. Across 1,200 trajectories, models differed most in repairing information, updating

plans, converting action into consequence, and preserving safe continuity. Individual judgments remained uncertain, while broad case coverage and repeated structured assessment produced stable rankings shared by physicians and multiple LLM Judges. Patient-facing medical AI should be evaluated by what its knowledge accomplishes across an evolving trajectory: acting under limited control, verifying consequences, repairing failure, and keeping unresolved obligations accountable. CareLoop operationalizes a shift from executable state change toward consequence stewardship in open human worlds.

Reproducibility statement

The anonymous repository at https://anonymous.4open.science/r/careloop-iclr2027-6e55 contains the executable runtime, 120 deidentified case contracts, 1,200 trajectories, 4,800 LLM-Judge assessments, 6,000 physician assessments, score definitions, analysis scripts, tests, and figure-generation code. Appendix B.11 maps claims to released artifacts. All primary estimates can be regenerated without model API access.

Ethics statement

All reported experiments concern simulated clinical interactions and are not used to direct patient care. The released cases are deidentified derivatives and exclude raw EHR records and source-to-case mappings. Physician reviewers evaluated deidentified simulation trajectories; public records use opaque pseudonymous identifiers and exclude names, institutions, contact details, and the private identity mapping. The framework and results do not establish clinical effectiveness, clinical safety, or fitness for deployment.

AI use statement

Generative AI tools were used to assist with research-method feedback, implementation, data cleaning and reformatting, statistical analysis, result interpretation, literature search, figure preparation, and manuscript drafting, editing, and translation. The authors reviewed the AI-assisted work, executed the released tests and analyses, checked numerical claims against the released artifacts, and take responsibility for the final content, claims, code, and data release. Generative models also served as experimental doctor-role models, simulated actors, and blinded Judges; those uses are part of the reported method rather than manuscript preparation.

References

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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, 2026a. doi: 10.48550/arXiv.2606.31179. URL https://arxiv.org/abs/2606.31179.

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, 2026b. doi: 10.48550/arXiv.2605.02240. URL https://arxiv.org/abs/2605.02240.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

A Benchmark comparison

B Additional methods and results

The following supplementary material preserves the full case schema, runtime mechanisms, measurement definitions, evaluator analyses, physician comparisons, representative records, and public-artifact index used in the study.

B.1 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 3 summarizes the main fields used to create and interpret CL120 trajectories.

Table 2: Conceptual comparison with representative executable agent environments. "Shared control" indicates that actors other than the evaluated agent can change task-relevant state.

Environment	Fundamental unit	Shared control	Success evidence	Boundary relative to CareLoop
GAIA	Real-world question	Limited	Correct answer	Broad tool reasoning without a persistent shared world.
WebArena/WorkArena	Web/enterprise task	No	Programmatic final state	Fixed user goal and mostly agent-controlled execution.
OSWorld	Computer-use task	No	App/OS evaluator	Rich interface grounding; limited actor belief or evolving goals.
SWE-bench	Repository repair	No	Executable tests	Strong oracle for a comparatively determinate state transition.
AppWorld	Cross-app API task	Partial	State-based tests	Apps and people in a controllable world; clinical risk is outside scope.
τ-bench	Service episode	Limited	Database state and policy	Simulated user supplies information and preferences.
τ^2-bench	Collaborative episode	Yes	Global state/history reward	Dual control is close in spirit; clinical evidence and closure are distinct.
TheAgentCompany	Workplace task	Yes	Task-specific checks	Consequential coworker interaction without evolving disease state.
CRMArena/Pro	Professional service task	Limited	CRM state and rules	Human-service setting without bounded clinical continuity.
PatientAgentBench	Patient-facing episode	Yes	Multidimensional jury	Closest medical analogue; friction is less explicit as an authored factor.
PhysicianBench	EHR workflow	No	Executable checkpoints	Clinician-facing institutional work rather than patient-world progression.
HealthAgentBench	Healthcare workflow	Varies	Final task success	Broad workflow coverage with a different central task unit.
Agent Hospital family	Hospital simulation	Yes	Simulation outcome	Institution-scale processes without frozen friction-response contracts.

Table 3: 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.

B.2 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 4 gives examples of runtime mechanisms that can appear in CL120 cases.

B.3 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 preserved when another reviewer disagrees.

Table 4: 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.

Table 5: 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.

B.4 Friction-Response Measurement

The friction-response layer is calculated from unmodified structured fields in the 4,800 canonical Judge records. Table 6 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.

B.5 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 7 summarizes the CL120 categories.

B.6 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 8 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

Table 6: 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

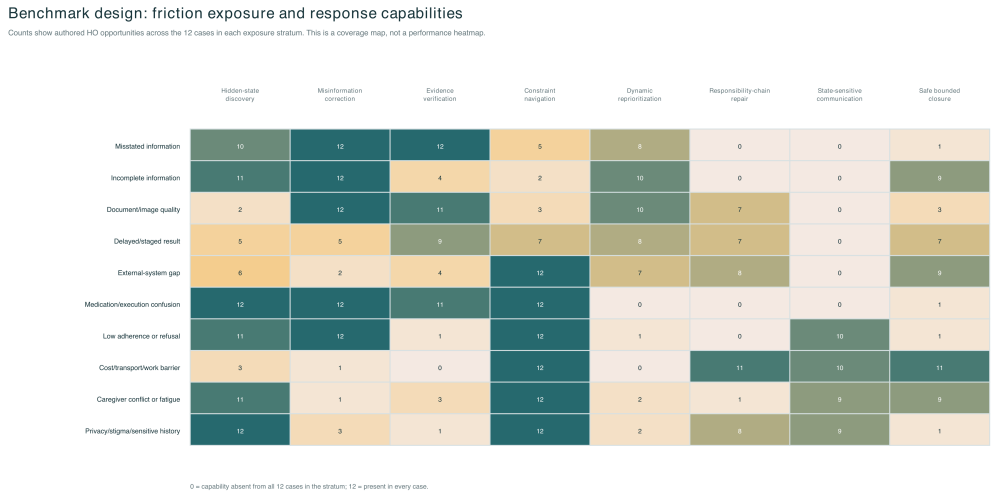


Figure 5: Coverage of the two complementary design axes. Rows are the 10 primary real-world-friction strata, each containing 12 cases. Columns are the eight HO response-capability types. A cell gives the number of cases in that friction stratum with at least one prospectively authored HO point of that 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.

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%).

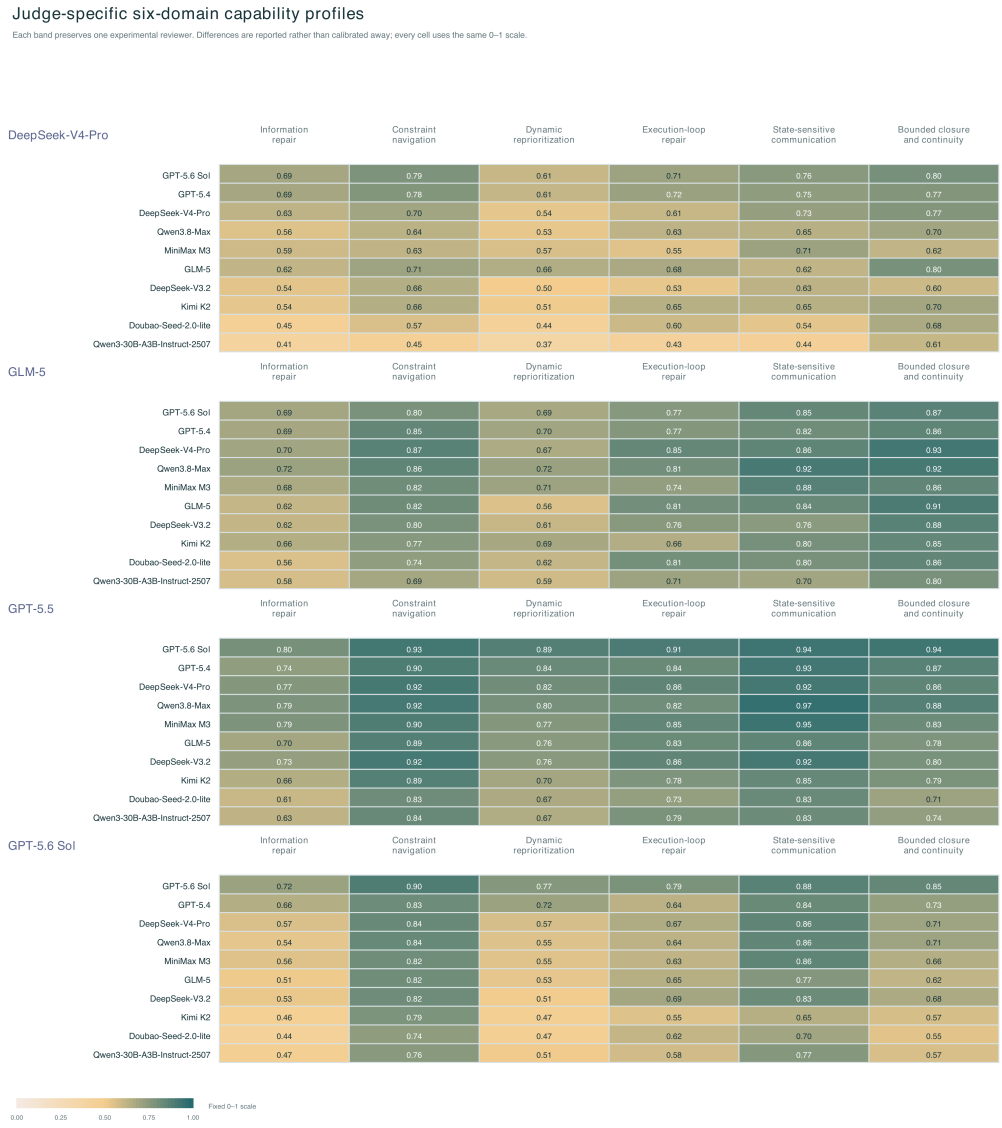


Figure 6: Judge-specific six-domain friction-capability profiles. Each panel recomputes domain means using one experimental Judge while preserving the common 0–1 scale and the all-Judge FCC model order. Absolute calibration differs across Judges, while many relative model patterns remain visible. These differences are reported as experimental findings because the four LLM Judges are neither interchangeable nor a clinician gold standard.

Table 7: 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.

Table 8: 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

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.

B.7 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 8 tests whether these authored dimensions show monotonic associations with the new primary readouts and with interaction length.

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

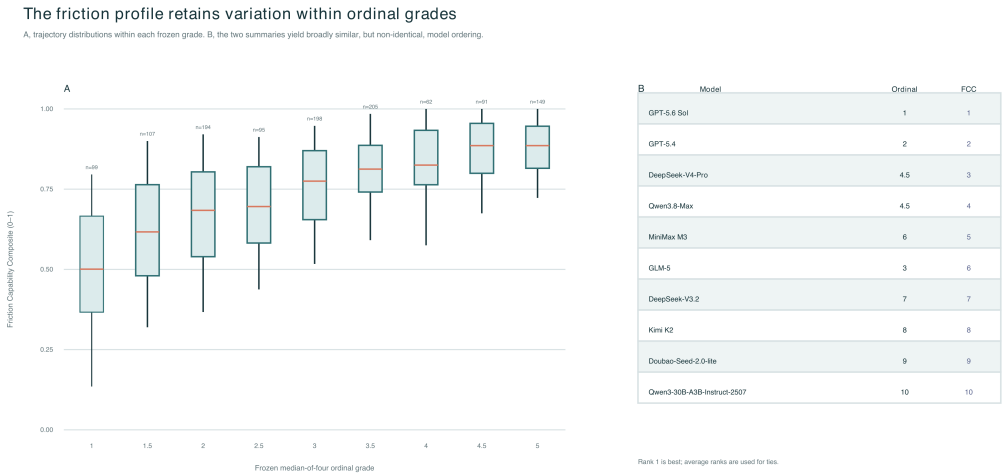


Figure 7: FCC distribution within each frozen ordinal grade and model ranks under the two summaries. The conventional 1–5 comparator compresses substantial process variation and is reported as a secondary analysis.

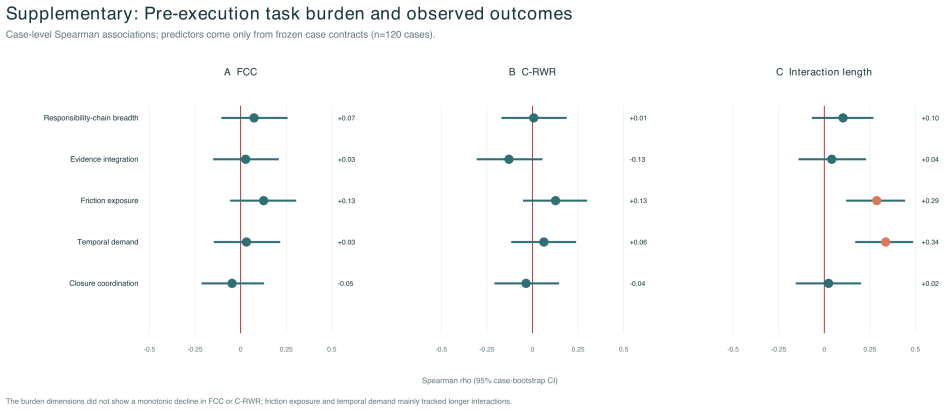


Figure 8: Associations between five pre-execution task-burden dimensions and observed outcomes across 120 cases. Each point is Spearman's ρ between one contract-derived burden score and the case mean across the 10 tested models; horizontal lines are 95% case-bootstrap intervals. Panels A and B use FCC 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.

frequency is reported numerically. Frequency alone cannot establish necessity, correctness, or causal benefit, so it receives no standalone interpretation.

B.8 Evaluator Association and Model-Level Diagnostics

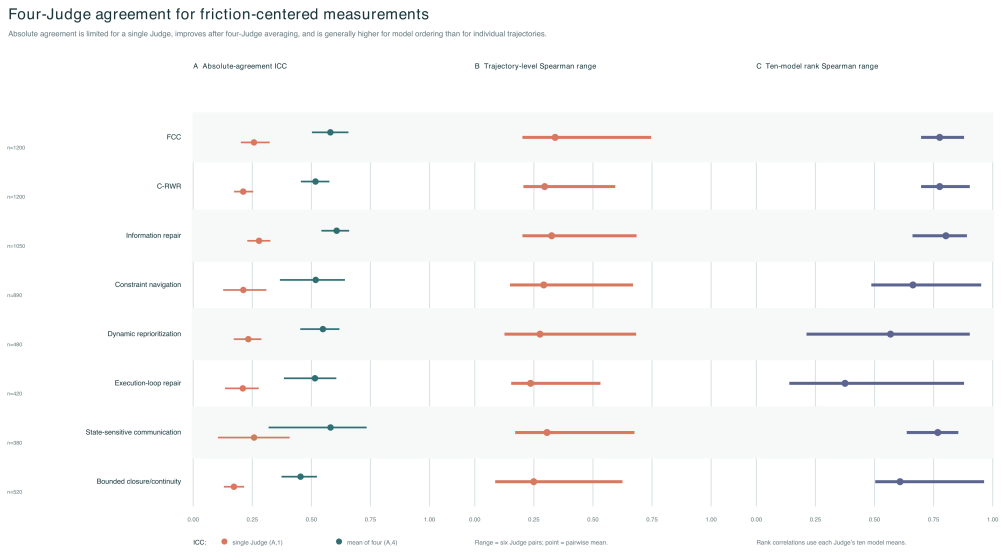


Figure 9: Four-Judge agreement for FCC, C-RWR, and the six capability domains. Panel A reports absolute-agreement ICC for one Judge and the mean of four; Panels B and C show pairwise trajectory-level and model-ranking correlations. Agreement depends on the unit and metric, supporting multi-Judge aggregation while retaining reviewer-specific uncertainty.

Three Judge models also appeared in the tested-doctor panel. Figure 10 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 11 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.

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.

B.9 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 9 was checked against the complete public trajectory, frozen case contract, and four canonical Judge records. The examples illustrate mechanisms and do not estimate prevalence.

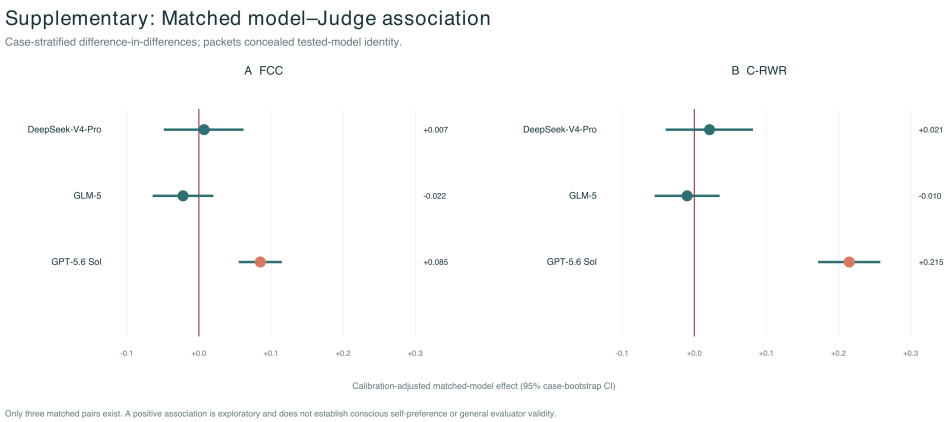


Figure 10: Exploratory matched model–Judge association under FCC and C-RWR. Points are case-stratified difference-in-differences estimates; horizontal lines are 95% case-bootstrap intervals and the red line marks zero. For FCC, adjusted effects were +0.007 for DeepSeek-V4-Pro, −0.022 for GLM-5, and +0.085 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.

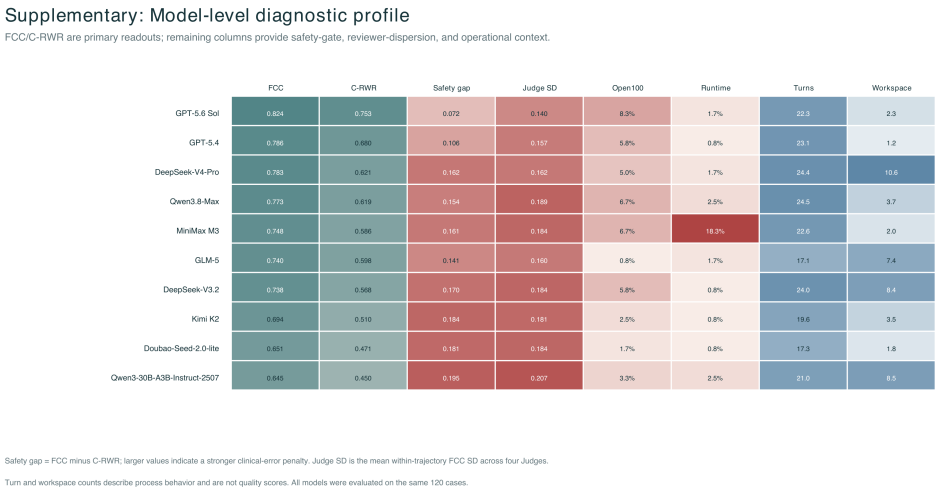


Figure 11: Model-level diagnostic profile. Rows follow the overall FCC ranking. FCC and C-RWR use a common 0–1 capability scale. Safety gap is FCC minus C-RWR and therefore increases when the clinical-error gate lowers otherwise strong friction-response performance. Judge SD is the mean within-trajectory 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.

Table 9: 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	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	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	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.
Verified stronger-doctor omission	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.

B.10 Blinded Physician Comparison

Paired case-level bootstrap comparisons quantified the contrast in Figure 12. 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

Table 10: 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%

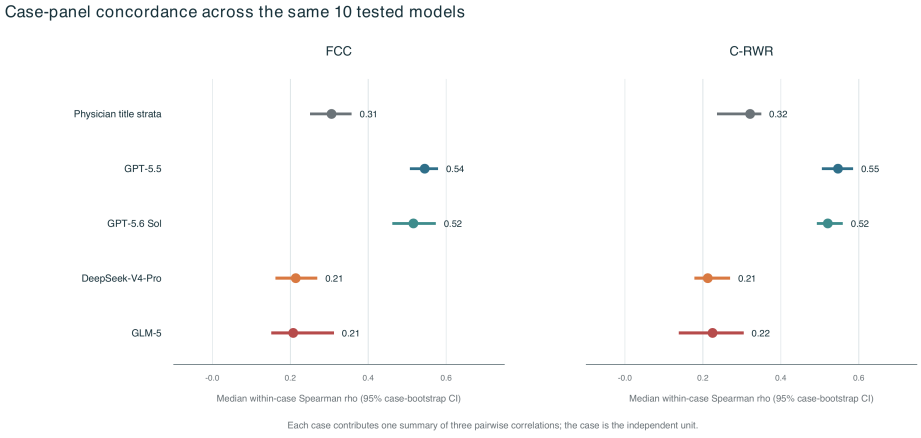


Figure 12: Symmetric fixed-panel comparison of physician-internal and physician–LLM-Judge rank concordance. Every point is based on ranking the same 10 models within a case; whiskers are case-bootstrap intervals.

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 12. 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 11: 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 12: 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

Case-specific agreement among physician title strata is summarized in Table 13; pairwise agreement between the two residents and between the two attendings is reported in Table 14. 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.

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



Figure 13: 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.

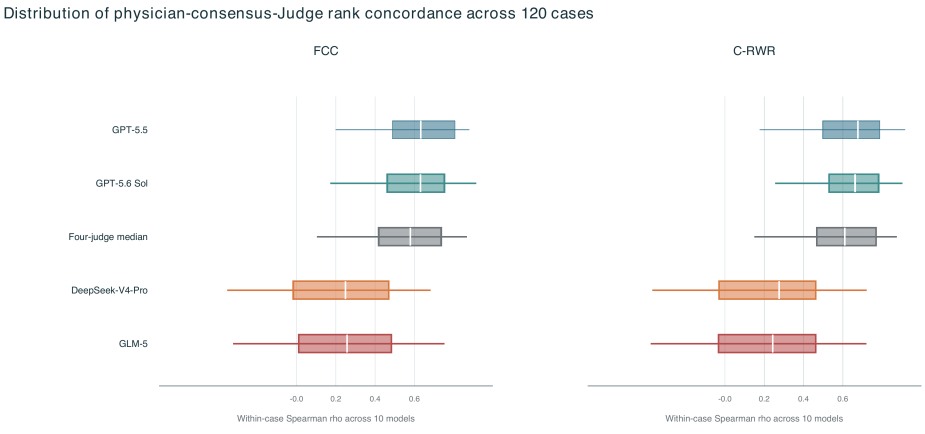


Figure 14: Distribution of case-wise physician-consensus–Judge rank correlations for FCC and C-RWR. Each box summarizes the 10-model Spearman correlation from one case across the 120 fixed case-panel assignments; boxes span the interquartile range, white lines mark medians, and whiskers show the 5th and 95th percentiles. The broad distributions show that aggregate benchmark ordering does not imply uniform agreement on every case.

Table 13: 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 14: 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 15: 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

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.

B.11 Public Artifact Index

The public repository is the authoritative location for complete trajectories, Judge records, executable scoring code, and machine-readable analysis. Table 16 lists the principal artifact groups supporting the manuscript analyses.

Table 16: 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 de-identified 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.
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 8–11 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 9.
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.

Table 16 continued. Public CL120 artifact index.

Path	Contents	Reproducibility use
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 12, Table 10, Figures 13–14, and Tables 11–14.
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 8 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.

B.12 Reproducibility and Interpretation Guardrails

The public package preserves nonclosure and runtime-error trajectories as part of observed system behavior. It excludes raw EHR sources and mappings, direct identifiers, institution labels, private reviewer mappings, provider payloads, credentials, local paths, notebook tokens, and terminal logs. Opaque reviewer identifiers and the released case-panel table support assignment auditing. Four LLM Judges serve as scalable experimental reviewers, with disagreement retained at trajectory, domain, and ranking levels. Three physicians reviewed the case contracts; a separate blinded panel of 139 physicians assessed all 1,200 trajectories, with one fixed five-physician panel per case. FCC and C-RWR are reproducible friction-centered measurements supported by physician comparison, not a clinician gold standard or deployment-safety certificate.