CLINCH: LINKING EVIDENCE, DECISIONS, AND EXECUTION IN LONG-HORIZON CLINICAL TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical agents can make locally appropriate decisions yet still fail to complete long-horizon tasks in electronic health record (EHR) systems. Completing such tasks requires gathering patient evidence dispersed across records, revising decisions as that evidence changes, and distinguishing intended actions from operations actually completed, so that the final documentation reflects what was done. We introduce CLINCH, a clinical agent harness that maintains a revisable workspace linking task requirements, patient evidence, clinical decisions, and execution status. CLINCH preserves and recovers decision-relevant context, provides patient-specific medical knowledge at decision points, and checks task completion and documentation against runtime-recorded actions and results. Identified gaps return to the agent for further retrieval, execution, or revision. We evaluate CLINCH with DeepSeek-V4-Pro, GPT-5.6-Luna, and GLM-5.2 on PhysicianBench and MedAgentBench. Relative to the corresponding baseline agents, CLINCH improves task success by 13.0, 4.7, and 21.0 percentage points on PhysicianBench and by 11.1, 12.2, and 19.4 points on MedAgentBench, respectively. On PhysicianBench, it also reduces failed checkpoints in retrieval, clinical reasoning, action execution, and documentation for all three models.

1 INTRODUCTION

Electronic health record (EHR) systems support clinical work ranging from reviewing patient histories to placing orders and documenting care. Large language model (LLM) agents can interact with these systems through tools that query records and write to them. Completing a clinical task, however, requires a chain of dependencies to hold across many interactions: task requirements determine which patient evidence is needed, that evidence and medical knowledge inform clinical decisions, decisions must be executed as operations in the EHR, and the final documentation must report what was actually done (Figure 1). Because a task succeeds only if every link holds, a correct medical judgment or a successful tool call is not sufficient on its own.

These links break in characteristic ways. Evidence can lose its clinical status as it is retrieved, compressed, and recalled over long contexts (Jiang et al., 2024; Liu et al., 2024): records are dispersed across encounters and coded inconsistently, an unretrieved or truncated result can later be read as a negative finding, and a past medication as a current one. We call this *evidence-state distortion*. Even with accurate evidence, a decision can fail to reflect medical knowledge available to the agent, as when a grading rule is known but the patient's findings are weighed incorrectly against it (*decision–knowledge decoupling*). Finally, what the agent intends, what the EHR records, and what the note reports can drift apart: a test is recommended but never ordered, an operation is assumed to have succeeded, or the note describes work that was not done (*unverified completion*). On PhysicianBench, these patterns surface as failed checkpoints across retrieval, clinical reasoning, action, and documentation (Figure 1, left); Appendix D traces representative runs for each.

Planning, tool interfaces, and execution feedback improve multistep agent performance, and design choices beyond the base model substantially affect outcomes (Yao et al., 2023; Wang et al., 2024; Yang et al., 2024; Yao et al., 2026). For long-horizon clinical tasks, however, existing designs do not close these gaps, because each addresses a different problem. Context compression and long-context methods decide which content to keep by relevance or position (Jiang et al., 2024; Liu et al., 2024), and clinical summarization such as RetroSum re-evaluates which earlier findings matter as new

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.

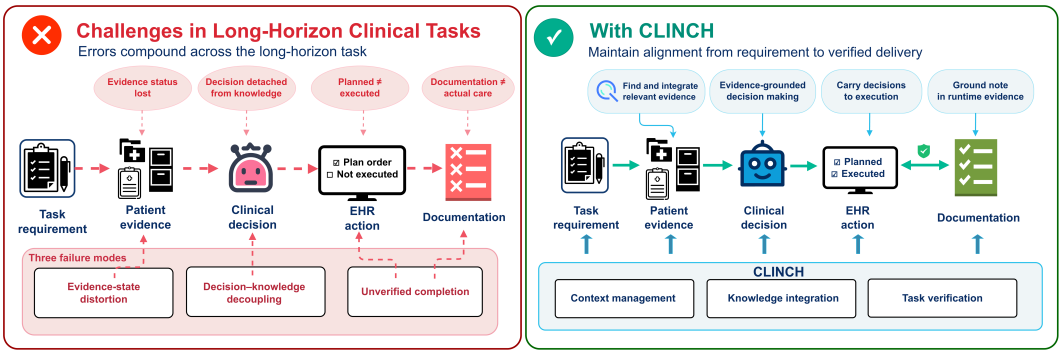


Figure 1: Left: challenges in long-horizon clinical tasks, where breaks between task requirements, evidence, decisions, actions, and documentation give rise to three failure modes. Right: CLINCH links these steps through context management, knowledge integration, and task verification.

evidence arrives (Liao et al., 2026). Evidence-state distortion, however, is a loss of qualifiers rather than of content: a summary line stating that a potassium result was "not found" is short and relevant, yet it no longer says whether the test was never ordered, fell outside the query window, or was truncated from view. Medical retrieval augmentation answers a query with relevant passages (Xiong et al., 2024), and experience-based methods retrieve procedural guidance from prior trajectories, in SELSM as entity-agnostic skills (Yang et al., 2026; Liao et al., 2025). Both indicate what is generally true or what to do next; neither maintains a record of how a rule applies to this patient's findings, so the connection must be re-established whenever the decision is revisited or documented. Execution feedback and self-reflection correct actions that fail or that the agent judges to be poor (Shi et al., 2024; Shinn et al., 2023), but an order that is never placed raises no error, and without an external check, reflection over the agent's own transcript inherits the transcript's claims. Detecting unverified completion requires comparing the requirements and the note with what the EHR actually received, which is also how agent benchmarks judge task completion (Yao et al., 2024).

In response, we introduce CLINCH, a CLINiCal agent Harness for long-horizon tasks in EHR environments. CLINCH maintains a revisable Clinical Workspace linking task requirements, patient evidence, clinical decisions, and execution status, and logs every EHR operation with its result. Each of its three mechanisms addresses one of the failures above (Figure 1, right). Clinically-Aware Context Management stores patient facts with their sources and status (acquired, unknown, or conflicting), compresses history without collapsing these distinctions, and recovers original records when a decision needs them. Decision-Anchored Knowledge Integration records each consequential decision with its evidence, patient-specific interpretation, and intended action, and conditions medical retrieval and review on that record, so that the reasoning remains available to later actions and documentation. Evidence-Grounded Task Verification compares the original requirements, the workspace, and the candidate note with the operation log, and returns omissions and inconsistencies to the agent for further retrieval, action, or revision. Our contributions are as follows:

- We identify three failure modes that arise in long-horizon EHR tasks even when individual steps are correct: *evidence-state distortion*, where retrieved evidence loses its source or status; *decision–knowledge decoupling*, where an available clinical rule is misapplied to the patient's findings; and *unverified completion*, where planned actions and records diverge.
- We introduce CLINCH, a clinical agent harness that represents the task chain as explicit, revisable state. A Clinical Workspace links task requirements, patient evidence, clinical decisions, and execution status, so that a revised decision propagates to the actions and documentation that depend on it, while a separate runtime log records what the EHR actually received.
- We evaluate CLINCH with three base models on PhysicianBench and MedAgentBench, where it improves task success by 4.7–21.0 percentage points over the corresponding baselines and reduces failed checkpoints in every PhysicianBench category; ablations examine the contribution of each mechanism group.

2 Related Work

2.1 Medical Agents and EHR Task Execution

Medical agents use patient records both to answer clinical questions and to carry out care-related tasks. EHR question-answering benchmarks cover structured records and multimodal patient data (Lee et al., 2022; Bae et al., 2023). For reasoning over structured EHR data, EHRAgent combines medical knowledge and database schemas to generate executable code, revising it in response to execution errors (Shi et al., 2024). EHRMaster organizes this reasoning around a task plan, aligns its concepts with data fields, and selects code execution or direct reasoning (Yang et al., 2025). Reinforcement learning has also been used to train agents for FHIR tool use (Knorr et al., 2026). Clinical decision-making has been studied through iterative patient-information gathering (Hager et al., 2024). MIRA uses the interaction history and prior tool results to plan emergency care, with FHIR tools for ordering tests, prescribing medications, and arranging admission in a sandbox EHR (Ferber et al., 2026). MedAgentBench and PhysicianBench evaluate patient-specific EHR interactions, with PhysicianBench assessing complete tasks across retrieval, clinical reasoning, action execution, and documentation (Jiang et al., 2025; Liu et al., 2026; Lee et al., 2025).

2.2 State and Knowledge Support for Long-Horizon Clinical Execution

During extended clinical interactions, information acquired early may become relevant only after later findings. Retaining that information and recognizing its changing significance are therefore closely related problems. RetroSum, introduced with AgentEHR, addresses them through retrospective summarization: it revisits earlier and recent interactions together to reassess clinical evidence as new observations arrive. The agent retains the full interaction history alongside the updated summary, preserving access to details while using the summary to highlight clinically relevant relationships. RetroSum also retrieves accumulated experience to guide both summarization and action selection, connecting interpretation of the current history with strategies learned from prior tasks (Shinn et al., 2023; Zhao et al., 2024; Liao et al., 2026).

The usefulness of prior experience depends on its relevance to the current patient and stage of a task. Retrieval-augmented generation has also been studied as a means of grounding language models in external knowledge, including systematic evaluation in medical question answering (Xiong et al., 2024). AgentClinic provides both retrieval of lessons from successful diagnostic cases and a separate medical-information retrieval path (Schmidgall et al., 2026). SELSM makes experience selection sensitive to the current execution state. It distills entity-agnostic logical skills from simulated FHIR trajectories, filters them by the task query, and matches them against the instruction, patient context, and accumulated actions and responses. This design addresses the ambiguity of similar intermediate observations across different clinical workflows: a successful tool response alone does not identify which skill is appropriate next. Combining the task goal with interaction state helps select guidance for the current step (Yang et al., 2026).

Retrieved experience can also support the assessment and revision of actions. ReflecTool accumulates successful trajectories and tool-level experience through reflection, then uses that experience with the current task and interaction history to evaluate and improve candidate actions (Liao et al., 2025).

These studies improve how clinical agents interpret interaction history and use experience to guide their next actions. CLINCH addresses the complementary problem of carrying evolving clinical decisions through to task completion. It links changes in patient evidence and judgment to outstanding task requirements, and checks required actions and documentation against independently recorded operations, so that revisions in reasoning are reflected in both execution and the final clinical record.

3 Method

3.1 Overview

We model a long-horizon clinical task as EHR interactions driven by an instruction I. The agent takes an action a_t to retrieve records or operate on the environment and receives a tool response

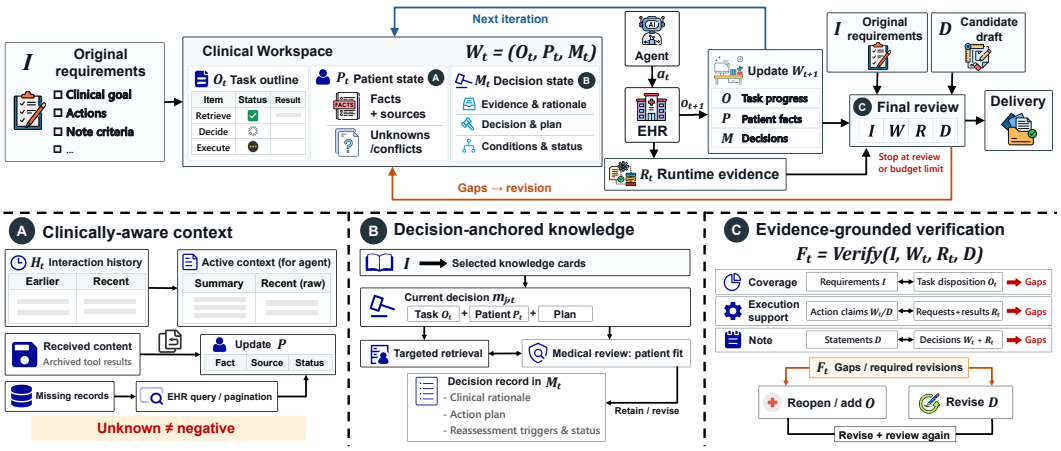


Figure 2: Method overview of CLINCH. Top: the main agent acts on the EHR and updates the Clinical Workspace, while the runtime records execution evidence; final review checks the candidate draft and returns identified gaps to the agent. Bottom: (A) clinically-aware context management, (B) decision-anchored knowledge integration, and (C) evidence-grounded task verification.

o_{t+1}. Queries reveal patient information, while operations such as order placement change the environment. Starting from $H_0 = (I)$, the interaction history grows as

$$H_{t+1} = H_t \parallel (a_t, o_{t+1}), \tag{1}$$

where $\parallel$ denotes appending. The agent must use acquired observations to revise judgments, perform required actions, and produce documentation D consistent with their outcomes.

CLINCH organizes the information, working state, and feedback available to the main agent (Figure 2). Context management retrieves and preserves clinical information; the Clinical Workspace maintains current task and patient decision state; knowledge integration supports patient-specific judgments; and verification checks requirements against execution and documentation. The runtime records tool calls, resource writes, and read-back results as R_t. Unlike this execution evidence, the workspace records the agent's current interpretation and progress. Full history H_t need not remain verbatim in the active context. Knowledge and verification feedback return to the main agent for further retrieval, action, or revision. The loop ends when the review accepts the candidate or a round or budget limit is reached.

3.2 Clinical Workspace

CLINCH maintains an explicit working state alongside interaction history:

$$\mathcal{W}_t = (O_t, P_t, M_t), \tag{2}$$

where O_t contains the dynamic task outline and progress, P_t task-relevant patient facts and unresolved information, and M_t current medical decisions and their rationale. The main agent updates these linked objects, and the runtime persists them. A revised decision can therefore change both planned actions and documentation requirements.

The *Dynamic Clinical Outline* organizes the instruction into a hierarchy of actionable, checkable items:

$$O_t = (V_t, E_t^{\text{plan}}), \qquad v_{i,t} = (\ell_{i,t}, \sigma_{i,t}, y_{i,t}), \tag{3}$$

where V_t and E_t^{plan} are nodes and parent–child relationships. Each node has a description $\ell_{i,t}$, a pending, in-progress, or completed status $\sigma_{i,t}$, and a result summary $y_{i,t}$. The agent can add or decompose items as evidence changes. Judgment and execution remain separate: recommending treatment does not complete the corresponding order-placement item.

Patient state P_t retains facts and sources, distinguishing acquired evidence, unknown information, and conflicting records. An unsuccessful query is qualified by its scope rather than treated as a

confirmed negative finding. Decisions affecting subsequent actions are recorded as

$$M_t = \{m_{j,t}\}, \qquad m_{j,t} = (e_{j,t}, i_{j,t}, q_{j,t}, a_{j,t}, x_{j,t}, u_{j,t}), \tag{4}$$

with evidence and sources $e_{j,t}$, patient-specific interpretation $i_{j,t}$, decision $q_{j,t}$, intended action $a_{j,t}$, conditions that would change the decision $x_{j,t}$, and execution status $u_{j,t}$. These links support both medical review of the rationale and verification of the resulting actions.

Updates propagate across the workspace. A new renal-function result, for example, enters P_t, may change a medication decision in M_t, and updates prescription or monitoring items in O_t. After execution, results enter R_t and the agent updates action and task status. This keeps subsequent operations and documentation aligned with revised evidence (Figure 2, top).

3.3 Clinically-Aware Context Management

Clinical information may be incompletely retrieved or no longer visible in the active context (Appendix D.1). CLINCH links acquisition, semantic preservation, and on-demand recovery (Figure 2A), supporting access to the evidence underlying the workspace's current judgments.

Terminology search and patient-record matching connect clinical concepts to candidate codes and records. Results preserve sources and query scope, distinguishing absence within a query from absence of a condition. Lengthy responses are archived in full as received; the active context shows an excerpt, a read-back reference, and pagination or truncation status. The agent can retrieve archived content or query the EHR for information not yet acquired.

As H_t grows, recent interactions remain verbatim while older ones are summarized into findings, judgments and revisions, completed actions, and outstanding items. Compression instructions preserve decision-relevant values, units, negative findings, sources, status, and clinically relevant timestamps. In particular, summaries distinguish unexamined from negative findings and planned actions from completion supported by tool results. Compression organizes existing information rather than adding clinical inferences; original records R_t remain available for checking.

When a judgment requires omitted details, the agent searches prior dialogue to locate the relevant interaction and reads the underlying record through content search or archive references. It then updates patient state and decisions from that evidence, querying the EHR further if the archive is incomplete. Summaries sustain continuity, while read-back recovers details that the summaries cannot supply.

3.4 Decision-Anchored Knowledge Integration

CLINCH anchors medical knowledge support to decisions that affect subsequent actions (Figure 2B). Outstanding items in O_t, patient conditions in P_t, and decision summaries in M_t identify what evidence is needed and where its conclusions should be retained. For a decision $m_{j,t}$, the evidence $e_{j,t}$, interpretation $i_{j,t}$, and intended action $a_{j,t}$ guide a focused medical query. Support addresses choices such as whether to order a test or adjust treatment under the current patient's conditions. The main agent can also initiate retrieval when it identifies a knowledge gap.

Auxiliary medical review reads patient state and the plan or decision, checks agreement among patient conditions, medical evidence, and intended actions, and consults sources when retrieval is enabled. Feedback identifies unconfirmed conditions, missing support, or inapplicable recommendations. The main agent gathers evidence or revises actions in response, retaining adequately supported decisions when appropriate. Patient-specific conclusions and sources are stored in M_t, with corresponding updates to patient facts and outstanding items. These records then inform later execution and documentation (Appendix D.2).

Lightweight knowledge cards provide general medical background for initial preparation. Before execution, we group clinical topics covered by the task instructions and summarize relevant public guidelines and reviews into a shared library of topic-level notes. Card-writing agents receive these topics and a common specification; patient records, agent trajectories, checkpoint results, scores, and evaluation rationales are excluded from their inputs. At run start, the system selects a small set of relevant cards from this library using the task instruction. The main agent checks their applicability to the patient, while decision-specific retrieval and medical review address questions that arise during execution.

3.5 Evidence-Grounded Task Verification

CLINCH checks completion against the original requirements, actual actions, and documentation. After the initial outline is constructed, requirements extracted from I are returned to the main agent for comparison with O_t. At delivery, an auxiliary model performs *Final Draft Review* (FDR):

$$\mathcal{F}_t = \text{Verify}(I, \mathcal{W}_t, R_t, D). \tag{5}$$

The instruction I provides an independent requirements reference, $\mathcal{W}_t$ current decisions and task disposition, R_t execution evidence, and D candidate documentation. Feedback $\mathcal{F}_t$ identifies gaps and required revisions. Including I separately allows detection of requirements that never entered the outline.

Review checks requirement coverage, action claims against actual requests and returned resources, and consistency between completed work and documentation. Evidence must support the claimed operation and outcome for the current patient. When an action's conditions are met, recommendations cannot substitute for required execution; deferred actions require a patient-specific rationale. Verification detects both unsupported completion claims and completed work omitted from the document (Appendix D.3).

The agent addresses feedback by adding omitted requirements, checking records before repeating uncertain actions, performing missing work, or revising documentation. It then updates the workspace and resubmits the candidate (Figure 2C). Medical review assesses decision appropriateness; task verification checks fulfillment and accurate reporting. The loop accepts a candidate with no outstanding review issues or ends under the review-round or task-budget limit.

4 Experiments and Analysis

We evaluate how CLINCH changes end-to-end task completion for a given base model in EHR environments. We compare baseline and CLINCH configurations across three models and two benchmarks, and examine task success and checkpoint completion under mechanism removal.

4.1 Experimental Setup

PhysicianBench contains 100 tasks and 670 checkpoints spanning retrieval, clinical reasoning, actions, and documentation (Liu et al., 2026); we use the full task set and benchmark scoring. MedAgentBench uses a fixed 60-task subset of v2 focused on record operations and conditional execution: ten instances from each of six templates, including ten retrieval tasks involving missing results and 50 write or conditional-action tasks. The subset construction and evaluation design are described in Appendix A.1.

We compare baseline and CLINCH using DeepSeek-V4-Pro-0423 (hereafter DeepSeek-V4-Pro), GPT-5.6-Luna, and GLM-5.2, referred to in the text as DeepSeek, Luna, and GLM, without fine-tuning the main model. We follow the experimental setting of the PhysicianBench paper (Liu et al., 2026). We run each model under each configuration three times in the main experiments. For the ablation study, we evaluate each configuration once on all 100 PhysicianBench tasks. Tasks and scoring are shared within each benchmark. PhysicianBench uses its original agent baseline; MedAgentBench uses the same adapted runtime with additional mechanisms disabled, and official scoring functions applied through a trajectory-conversion layer. Tables 1 and 3 report recorded inference tokens per task; Appendix B gives the accounting details.

The primary metric is mean single-run task success: Pass@1 on PhysicianBench and Overall SR on MedAgentBench. The former requires all checkpoints to pass; the latter uses official binary scoring. Pass@2 averages success at least once across run pairs; Pass@3 and Pass^3 are the proportions of tasks passed at least once and in all three runs, respectively. We also report checkpoint pass rate (CP), and MedAgentBench retrieval and action success (Query SR, Action SR) and the sample standard deviation of Overall SR. Differences use unrounded values. For the main-experiment gains, we report 95% confidence intervals from a paired bootstrap over tasks (Appendix A.2).

Table 1: PhysicianBench results on 100 tasks. Pass@1 and checkpoint pass rate (CP) are three-run means ± sample SD. ΔPass@1 is the CLINCH-minus-baseline gain in percentage points with a paired task-bootstrap 95% CI. Tokens are recorded inference tokens per task (thousands). Bold marks the better Pass@1 and CP means.

	Main metrics (%) ↑		Multi-run (%) ↑				
Agent	Pass@1	CP	Pass@2	Pass@3	Pass^3	ΔPass@1 [95% CI]	Tokens (K)
DeepSeek-V4-Pro							
Baseline	21.3 ± 1.2	72.0 ± 2.3	30.3	35.0	8.0	–	524.0
CLINCH	$\mathbf{34.3} \pm 2.5$	$\mathbf{80.6} \pm 1.5$	49.0	57.0	13.0	$+13.0$ $[5.7, 20.7]$	2,299.0
GPT-5.6-Luna							
Baseline	26.7 ± 2.9	67.7 ± 1.8	35.3	40.0	14.0	–	196.4
CLINCH	$\mathbf{31.3} \pm 1.2$	$\mathbf{76.6} \pm 0.6$	42.3	48.0	15.0	$+4.7$ $[-3.0, 12.3]$	2,105.5
GLM-5.2							
Baseline	24.7 ± 3.5	76.5 ± 2.0	35.0	41.0	10.0	–	377.1
CLINCH	$\mathbf{45.7} \pm 2.9$	$\mathbf{84.6} \pm 0.2$	60.3	66.0	22.0	$+21.0$ $[13.3, 28.7]$	2,066.1

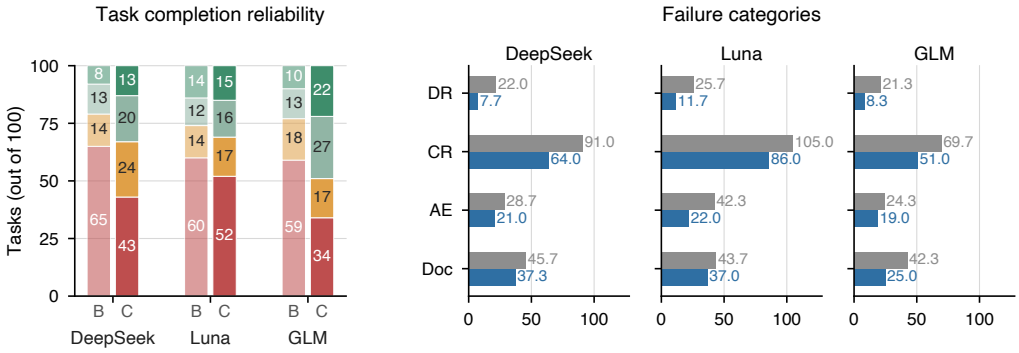


Figure 3: Task-level outcomes and failed checkpoints on PhysicianBench. Left: tasks grouped by the number of successful runs out of three for the baseline (B) and CLINCH (C); red, orange, light green, and dark green denote 0, 1, 2, and 3 successes. Right: failed checkpoints per 100 task runs by category, for the baseline (gray, top) and CLINCH (blue, bottom).

4.2 PhysicianBench Results

CLINCH improves Pass@1 by 13.0, 4.7, and 21.0 percentage points for DeepSeek, Luna, and GLM, respectively (Table 1). GLM shows the largest gain, reaching 45.7% task success. All three models also improve on CP and the multi-run metrics.

Figure 3 shows where these gains come from at the task level. With CLINCH, fewer tasks fail in all three runs (65 to 43, 60 to 52, and 59 to 34 for DeepSeek, Luna, and GLM), and GLM completes 22 tasks in all three runs, up from 10. The right panel groups failed checkpoints by the released PhysicianBench labels: Data Retrieval (DR, 103 checkpoints), Clinical Reasoning (CR, 356), Action Execution (AE, 112), and Documentation (Doc, 99), applied identically across configurations. Counts are failed checkpoints per 100 task runs, not failed-task rates or independent causal errors.

Failed checkpoints decrease in all four categories for all three models, with different patterns of improvement. DeepSeek's reasoning failures decline from 91.0 to 64.0 per 100 tasks, while Luna's action failures fall from 42.3 to 22.0, a 48.0% relative reduction. For GLM, retrieval failures decrease by 60.9%, while documentation failures fall by 40.9%, a larger relative reduction than for DeepSeek or Luna.

Luna's CP gain is comparable to those of the other models, yet its Pass@1 gain is much smaller. CLINCH governs how an agent works through a task. It keeps evidence available, anchors decisions

Table 2: Token-budget control on PhysicianBench with DeepSeek-V4-Pro. Baseline and CLINCH report three-run means; the continuation control reports one run on the same 100 tasks. Tokens are recorded usage per task (thousands).

Configuration	Pass@1 (%)	CP (%)	Tokens (K)
Baseline	21.3	72.0	524.0
Baseline + continuation	22.0	66.9	2,131.9
CLINCH	34.3	80.6	2,299.0

Table 3: Success rates on the 60-task MedAgentBench subset (%). Overall SR is the three-run mean $\pm$ sample standard deviation. Query SR and Action SR average results over three runs of ten retrieval and 50 action tasks, respectively. Pass@2 averages the three run-pair results; the other multi-run metrics and Tokens follow Table 1.

Model	Configuration	Overall SR	Query SR	Action SR	Pass@2	Pass@3	Pass^3	Tokens (K)
DeepSeek-V4-Pro	Baseline	68.9 ± 1.9	100	62.7	71.1	71.7	65.0	69.9
	CLINCH	$\mathbf{80.0 \pm 1.7}$	100	**76.0**	**88.9**	**93.3**	**66.7**	293.3
GPT-5.6-Luna	Baseline	60.6 ± 1.0	96.7	53.3	67.2	70.0	50.0	52.9
	CLINCH	$\mathbf{72.8 \pm 1.9}$	**100**	**67.3**	**81.7**	**85.0**	**58.3**	137.4
GLM-5.2	Baseline	71.1 ± 1.9	100	65.3	75.0	76.7	65.0	81.0
	CLINCH	$\mathbf{90.6 \pm 1.0}$	100	**88.7**	**96.1**	**96.7**	**80.0**	191.9

to that evidence, and checks that planned actions and documentation are completed. It does not change how the model interprets that evidence. When clinical judgment is largely sound and execution is the bottleneck, closing these gaps completes the task; this is the case for GLM, 90 of whose 226 failed baseline runs miss a single checkpoint. Luna is further from completion and weakest in clinical reasoning, with the most reasoning failures of the three models under both configurations. Its remaining errors lie in the judgment itself, such as misattributing a symptom, misapplying a clinical rule, or selecting an inappropriate treatment, and they carry into orders and documentation even when these are executed consistently. CLINCH reduces Luna's retrieval and action failures but cannot supply the missing judgment. More broadly, the task-success gain across the three models follows their baseline clinical-reasoning failures: GLM, with the fewest, gains most, and Luna, with the most, gains least. Clinical reasoning is a capability of the base model and must be improved in the model itself, for example through domain-specific training. Harness design and model improvement are therefore complementary: the harness grounds a model's judgment in evidence and carries it through to completion, and its gains grow as that judgment improves.

CLINCH's gains come with increased inference token usage. To examine whether additional usage alone yields similar gains, we extend the DeepSeek baseline with generic continuation prompts: at comparable recorded token usage, it achieves 22.0% Pass@1 versus CLINCH's 34.3% (Table 2; setup in Appendix B).

4.3 MedAgentBench Results

CLINCH increases Overall SR by 11.1, 12.2, and 19.4 percentage points for DeepSeek, Luna, and GLM, respectively (Table 3). GLM reaches 90.6% success, with a task-cluster bootstrap 95% confidence interval of $[11.7, 27.8]$ points for the gain (10,000 resamples). Its Action SR rises from 65.3% to 88.7%, while both configurations achieve 100% Query SR. Across all three models, retrieval is near ceiling and the improvements primarily concern action completion.

4.4 Ablation Studies

We compare three ablated configurations, each evaluated once on 100 DeepSeek tasks, against the full system averaged over the three main-experiment runs. These configurations remove the Dynamic Clinical Outline and clinical compression, decision-anchored retrieval support, or Final Draft Review and closed-loop checks, respectively.

Table 4: Task completion and failed checkpoints in the ablation study. The full system reports the mean of three runs; each ablated configuration reports one run. Every run evaluates the same 100 tasks. Pass@1 and CP are percentages. DR, CR, AE, and Doc count failed checkpoints per 100 task runs, as in Figure 3.

Configuration	Pass@1	CP	DR	CR	AE	Doc
CLINCH (three-run mean)	34.3	80.6	7.7	64.0	21.0	37.3
w/o Dynamic Clinical Outline and clinical compression	25.0	77.9	14.0	74.0	24.0	36.0
w/o decision-anchored retrieval support	26.0	77.0	12.0	73.0	23.0	46.0
w/o Final Draft Review and closed-loop checks	32.0	79.6	11.0	61.0	23.0	42.0

Each ablated group targets one of the failure modes identified in the introduction (illustrated in Appendix D), and removing it raises the failures associated with that mode (Table 4). Without the Dynamic Clinical Outline and clinical compression, which counter *evidence-state distortion*, retrieval failures nearly double (7.7 to 14.0 per 100 tasks), and reasoning failures also rise (64.0 to 74.0), consistent with distorted evidence propagating into decisions; Pass@1 falls to 25.0%. Without decision-anchored retrieval support, which targets *decision–knowledge decoupling*, reasoning and documentation failures increase (64.0 to 73.0 and 37.3 to 46.0), and Pass@1 falls to 26.0%. Without Final Draft Review and closed-loop checks, which address *unverified completion*, documentation failures increase (37.3 to 42.0), while Pass@1 declines less, to 32.0%. No single group accounts for the full gain: the groups act on different links among evidence, decisions, and execution, and a task succeeds only when all of these links hold.

5 Conclusion

We presented CLINCH, a clinical agent harness that links task requirements, patient evidence, clinical decisions, and execution status in a revisable workspace. Across three base models, CLINCH improves task success on PhysicianBench and MedAgentBench and reduces failed checkpoints in retrieval, reasoning, action, and documentation. The gains are largest when the base model already makes sound clinical judgments and fails mainly during execution.

The main limitation of CLINCH is that it does not correct the base model's clinical judgment. If the model misreads the evidence but acts consistently with that reading, checks against the operation log will not catch the error, and further retrieval or review helps only when it exposes the mistaken interpretation. CLINCH also uses more model calls than the baseline agent, because it maintains workspace state, provides knowledge at decision points, and verifies completion. Additional token usage alone does not reproduce its gains (Table 2), and part of this cost could be reduced by triggering review only for uncertain or consequential decisions and by delegating compression and consistency checks to smaller models. Finally, we evaluate CLINCH only in simulated EHR environments. Real records are larger and noisier, workflows involve multiple clinicians, and actions in a live system require privacy safeguards and clinician oversight, so our results measure task completion, not clinical benefit.

These limitations point to directions for future work. Because the gains are largest for base models with sound clinical judgment, pairing CLINCH with stronger or clinically specialized models may convert more of its execution support into task success. The workspace and verification records also show which evidence was used, which decision was made, and which requirement remained unmet, so they could supervise training on clinical evidence interpretation, conditional decision-making, and revision of earlier judgments. Evaluation on de-identified hospital records with clinician review, followed by clinician-in-the-loop studies, would test whether the gains carry over to clinical practice. Appendix E discusses these limitations and directions in detail.

Ethics Statement

We evaluate CLINCH using existing benchmark task sets and patient records in isolated EHR environments. PhysicianBench and MedAgentBench describe their underlying data as derived from de-identified clinical records (Liu et al., 2026; Jiang et al., 2025). Agent actions in these experiments affect benchmark records rather than live clinical systems. Clinical use would require clinician oversight, institutional data governance, and prospective evaluation of effects on patient care.

AI use statement

We used generative AI tools for literature review, feedback on experimental design, result analysis, and manuscript drafting, translation, and editing. We also used these tools to help write parts of the code and develop data analysis scripts. We reviewed the AI-assisted analyses and text and take responsibility for the final content and conclusions.

References

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. EHRXQA: A multi-modal question answering dataset for electronic health records with chest x-ray images. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3867–3880, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0c007ebef1d11fd48da6ce4f54687db6-Abstract-Datasets_and_Benchmarks.html.

Dyke Ferber, Lars Hilgers, Christiane Höper, Benedict Kinny-Köster, Jan-Niklas Eckardt, Katharina Egger-Heidrich, Marius Bill, Martin M. K. Schneider, Jan Clusmann, Lejla Kadric, Marcel Oehme, Maximilian Mayrhofer-Schmid, Alexander Oeser, Georg Wölflein, Isabella C. Wiest, Jan Moritz Middeke, A. John Iafrate, Daniel Truhn, Dirk Jäger, and Jakob Nikolas Kather. Towards autonomous medical artificial intelligence agents. *Nature*, 655:1282–1291, 2026. doi: 10.1038/s41586-026-10675-5. URL https://doi.org/10.1038/s41586-026-10675-5.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024. doi: 10.1038/s41591-024-03097-1. URL https://www.nature.com/articles/s41591-024-03097-1.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A realistic virtual EHR environment to benchmark medical LLM agents, 2025. URL https://arxiv.org/abs/2501.14654.

Marius S. Knorr, Robert Müller, Jan P. Bremer, and Nils Schweingruber. Reinforcement learning for tool-calling agents in fast healthcare interoperability resources (FHIR), 2026. URL https://arxiv.org/abs/2605.14126.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. EHRSQL: A practical text-to-SQL benchmark for electronic health records. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15589–15601, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/

643e347250cf9289e5a2a6c1ed5ee42e-Abstract-Datasets_and_Benchmarks.html.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Yugang Jia, and Jong Ha Lee. FHIR-AgentBench: Benchmarking LLM agents for realistic interoperable EHR question answering, 2025. URL https://arxiv.org/abs/2509.19319.

Yusheng Liao, Shuyang Jiang, Yanfeng Wang, and Yu Wang. REFLECTOOL: Towards reflection-aware tool-augmented clinical agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13507–13531, 2025.

Yusheng Liao, Chuan Xuan, Yutong Cai, Lina Yang, Zhe Chen, Yanfeng Wang, and Yu Wang. AgentEHR: Advancing autonomous clinical decision-making via retrospective summarization, 2026. URL https://arxiv.org/abs/2601.13918.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

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, 2026. URL https://arxiv.org/abs/2605.02240.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: a multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9:499, 2026. doi: 10.1038/s41746-026-02674-7. URL https://doi.org/10.1038/s41746-026-02674-7.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50208–50232. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24h.html.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.372. URL https://aclanthology.org/2024.findings-acl.372/.

John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2405.15793.

Wanrong Yang, Zhengliang Liu, Yuan Li, Bingjie Yan, Lingfang Li, Mingguang He, Dominik Wojtczak, Yalin Zheng, and Danli Shi. Empowering locally deployable medical agent via state enhanced logical skills for FHIR-based clinical tasks, 2026. URL https://arxiv.org/abs/2603.06902.

Xiao Yang, Xuejiao Zhao, and Zhiqi Shen. EHRStruct: A comprehensive benchmark framework for evaluating large language models on structured electronic health record tasks, 2025. URL https://arxiv.org/abs/2511.08206.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Yilun Yao, Xinyu Tan, Chao-Hsuan Liu, Yaoming Li, Zhengyang Wang, Wenhan Yu, Zhewen Tan, Yuxuan Tian, Guangxiang Zhao, Lin Sun, Xiangzheng Zhang, and Tong Yang. Harness-Bench: Measuring harness effects across models in realistic agent workflows, 2026. URL https://arxiv.org/abs/2605.27922.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

A Supplementary Evaluation Details

A.1 MedAgentBench Evaluation Design

PhysicianBench provides the primary evaluation of long-horizon clinical task completion and the checkpoint-based analysis of unmet requirements. MedAgentBench complements this evaluation by testing the adapted mechanisms in a separate EHR environment, with a focus on record operations and conditional execution.

MedAgentBench v2 contains 300 patient-specific tasks organized into ten templates, with 30 instances per template. We use a fixed 60-task subset spanning six templates: vital-sign recording and referral ordering, which require direct writes; conditional magnesium repletion, conditional potassium repletion with a paired laboratory order, and conditional HbA1c testing; and laboratory-result retrieval with an explicit missing-result case. The five write or conditional-action templates provide 50 tasks, while the retrieval template provides 10 tasks.

For each included template, the subset manifest selects the first ten instances after sorting the official numeric instance identifiers. The same task instances are used for every model and configuration, and each configuration is evaluated in three runs. This fixed construction keeps the contribution of each template balanced while allowing the comparison to focus on the ability to retrieve the relevant record, determine whether an action is required, and express the required operation correctly. The four templates outside this subset concern patient identification, age calculation, numerical aggregation, and latest-value retrieval; they are outside the action-oriented scope of this supplementary evaluation.

Both configurations use the same task instructions, FHIR interface, and patient records. Agent trajectories are converted into the representation expected by the benchmark's official task scorers, which then evaluate the retrieved answer or the required write operation. The conversion aligns the interface representation with the scorer without changing the task requirements or the scoring functions. The CLINCH configuration uses the mechanisms relevant to this interaction setting, including task outlining, context recovery, patient-state maintenance, operation checks, and absence grounding; mechanisms tied to clinical-document delivery are not used because these tasks do not require a clinical note. The task context supplies the relevant medical rules, and knowledge-card recall is disabled in this evaluation.

The reported aggregates are computed from the runs described here; stochastic agent execution can yield different point estimates under other run conditions.

A.2 Task-Level Confidence Intervals

We estimate uncertainty in the main-experiment gains using a paired bootstrap over tasks. Each of 10,000 resamples draws 100 PhysicianBench tasks or 60 MedAgentBench tasks with replacement, retaining all three runs of both configurations for every sampled task. We recompute the CLINCH-minus-baseline difference in mean task success for each resample. For PhysicianBench CP, we recompute each configuration's ratio of passed to total checkpoints over the sampled tasks. The 2.5th and 97.5th percentiles give pointwise 95% confidence intervals for each comparison (Tables 5 and 6). These intervals describe variation across tasks in the observed three-run collection; the standard deviations in Tables 1 and 3 describe variation across runs.

Table 5: PhysicianBench: CLINCH gains over baseline in percentage points, with pointwise 95% confidence intervals in brackets.

Model	Pass@1	CP
DeepSeek-V4-Pro	$+13.0$ $[5.7, 20.7]$	$+8.6$ $[5.3, 11.8]$
GPT-5.6-Luna	$+4.7$ $[-3.0, 12.3]$	$+9.0$ $[3.7, 14.4]$
GLM-5.2	$+21.0$ $[13.3, 28.7]$	$+8.1$ $[5.4, 10.9]$

Table 6: MedAgentBench: CLINCH gains over baseline in percentage points, with pointwise 95% confidence intervals in brackets.

Model	Overall SR
DeepSeek-V4-Pro	$+11.1$ $[1.7, 21.1]$
GPT-5.6-Luna	$+12.2$ $[2.8, 22.2]$
GLM-5.2	$+19.4$ $[11.7, 27.8]$

B Inference Token Usage

We report average recorded inference tokens per task over the three evaluated runs of each main-experiment configuration. Counts include main-agent execution, context compaction, and online auxiliary-model calls, including independent reviewers. Recorded cached input is counted once, and reasoning tokens are included in output tokens. Evaluation and one-time knowledge preparation are excluded.

Table 7: Average recorded inference tokens per task, in thousands.

Benchmark	Model	Baseline	CLINCH
PhysicianBench	DeepSeek-V4-Pro	524.0	2,299.0
	GPT-5.6-Luna	196.4	2,105.5
	GLM-5.2	377.1	2,066.1
MedAgentBench	DeepSeek-V4-Pro	69.9	293.3
	GPT-5.6-Luna	52.9	137.4
	GLM-5.2	81.0	191.9

Means are calculated over task runs with recorded token usage. Unavailable request usage and cache components are not imputed.

For the DeepSeek ablation comparison, average recorded usage is 2,299.0K tokens for the full system across three runs. The single-run ablated configurations use 1,678.7K without the Dynamic Clinical Outline and clinical compression, 2,206.6K without decision-anchored retrieval support, and 1,141.9K without Final Draft Review and closed-loop checks.

Token-budget control. We evaluate the DeepSeek baseline once on the same 100 PhysicianBench tasks with a continuation target of 2.35M recorded tokens per task. When the model attempts to

finish below this target, we append: "You still have processing budget remaining. You may continue working." The baseline's existing termination rules and a 100-step limit remain in effect. Average recorded usage is 2.13M tokens per task, 92.7% of CLINCH's 2.30M. As shown in Table 2 in the main text, this extension does not reproduce CLINCH's task-completion gains.

These measurements characterize the performance–resource trade-off of the evaluated harness configurations. In the ablation study, removing decision-anchored retrieval support yields similar recorded token usage but lower task completion than the full reference, highlighting the importance of how inference is organized across execution stages. More selective verification and revision offer directions for improving this trade-off.

C Method Details

Figure 4 shows the system components of CLINCH and how they connect to the main agent and the EHR environment. Context summarization, online card selection and retrieval support, and Final Draft Review use separate calls to the same base model as the main agent. The separate patient-state reviewer, when enabled, uses DeepSeek-V4-Flash.

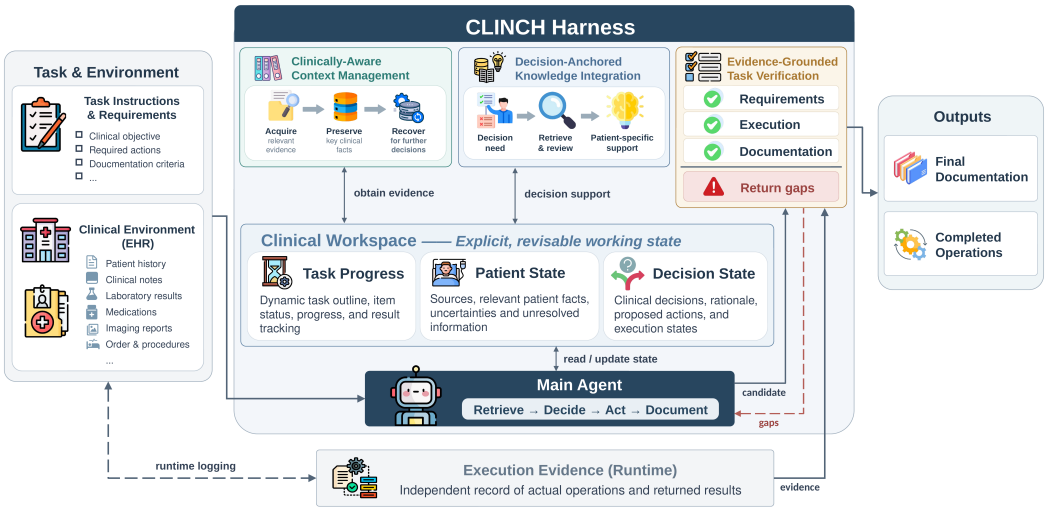


Figure 4: System architecture of CLINCH. The main agent interacts with the EHR while the Clinical Workspace maintains task progress, patient state, and decision state; context management, knowledge integration, and task verification operate around this shared state.

C.1 Dynamic Clinical Outline

The Dynamic Clinical Outline represents task requirements as a mutable hierarchy. A parent item can be decomposed into child items for evidence collection, assessment, decision-making, execution, and documentation. Each node retains its parent–child relation, a status, and a short result summary. The status distinguishes pending work from work in progress and completed work. A clinical decision and the corresponding environment operation remain separate: a treatment recommendation does not complete an order-placement item.

The main agent reads and updates the workspace through tool calls. Its current state is presented in the model context as a hierarchical outline and a compact patient-state snapshot, maintained separately from the summarized interaction history.

Evidence can change the structure of the outline. When a new result introduces a condition that affects treatment, the agent can add a retrieval or reassessment item under the relevant parent. After an operation, the corresponding node records the result and its execution status. If an existing resource satisfies an item, or an action is deferred, the node records that disposition and its rationale. The outline therefore maintains both task coverage and the current state of each requirement.

The outline is linked to patient facts, decisions, and execution evidence in the Clinical Workspace. Patient evidence supports an interpretation; the interpretation determines a decision and intended actions; and the returned execution result updates the relevant action and task nodes. Subsequent documentation can then use the revised state without reconstructing the task structure from the full interaction history.

C.2 Clinical Evidence Recovery

Clinical retrieval connects task concepts to terminology codes and EHR records. Results retain their source and query scope, allowing the agent to distinguish an empty result for one code, resource type, or time range from an established negative finding. When a hierarchical code is relevant, a failed exact query can be expanded to an appropriate code family while preserving the changed scope.

A returned result can be incomplete because only part of it is visible in the active context, or because the query returned a paginated set whose remaining records have not yet been acquired. CLINCH preserves the content already received together with a reference to the visible excerpt and its retrieval status. The agent can search the stored result for a relevant entry and read back the surrounding text, or continue pagination and query refinement when the result set itself is incomplete.

When the estimated history length exceeds a preset threshold, earlier interactions are summarized while recent interactions remain verbatim. The first DeepSeek CLINCH run on PhysicianBench uses an 80,000-token threshold and a 20,000-token budget for retaining recent interactions; the other CLINCH runs in the main evaluation use 192,000 and 40,000 tokens, respectively. Summaries retain decision-relevant values and units, negative findings, dates, sources, medication status, changes in interpretation, completed operations, and outstanding work. They preserve distinctions between unexamined and negative findings, and between intended actions and actions supported by tool results. When a later decision requires a detail absent from the summary, the agent locates the earlier retrieval and reads back the original result, then updates the linked patient facts and decision state. Summarization maintains continuity; read-back preserves access to the underlying evidence.

C.3 Knowledge Cards and Decision Support

Knowledge cards are prepared offline as a shared library. Clinical topics covered by the task instructions are grouped to guide searches of public clinical guidelines, reviews, and other medical sources. Card-writing agents receive the topics and a common organization specification; patient records, execution trajectories, checkpoint results, scores, and evaluation rationales are excluded from their inputs. Relevant material is condensed into short topic-level notes, retaining source references and the conditions under which the guidance applies.

At task initiation, an auxiliary model selects a small set of cards from this library by matching the instruction to the topics, titles, and aliases in the catalog. The selected material provides general background; the main agent combines it with the current patient's records to assess applicability and form patient-specific conclusions.

During execution, decision-specific retrieval and medical review provide support for the current patient and plan. The query is conditioned on relevant patient facts, the current candidate plan, and the decision that would affect later actions. Retrieved evidence is organized with its sources and candidate interpretation. Medical review checks the agreement among patient conditions, evidence, and intended action, then returns specific support or unresolved premises. The main agent can retrieve additional evidence, revise the plan, or retain an adequately supported decision. The resulting interpretation and sources are stored with the decision in the workspace so that later execution and documentation can use them.

C.4 Verification and Task Revision

Final verification compares the original task requirements, the current workspace, execution evidence, and candidate documentation. It checks whether required findings are addressed, whether claimed actions correspond to actual requests and returned resources, and whether the final document accurately reports the patient-specific decisions and actions.

When verification identifies a gap, the feedback specifies the affected requirement and the evidence supporting the finding. The main agent can retrieve missing information, check an existing resource, execute an outstanding action, or revise the document. The corresponding outline item is reopened or added when further work is required. If an action is deferred, the workspace records the patient-specific rationale and the condition for reconsideration.

The revised document and execution state are reviewed again until no further revision is requested, the limit of three review attempts per task is reached, or the task budget is exhausted. Medical review assesses the clinical fit of a decision, whereas task verification checks requirement coverage, execution grounding, and accurate delivery.

The CLINCH implementation and evaluation artifacts will be released upon publication.

D Failure Modes and Representative Cases

The introduction identifies three ways in which the links among evidence, decisions, and execution break: evidence-state distortion, decision–knowledge decoupling, and unverified completion. Sections D.1–D.3 illustrate each with baseline runs from the main PhysicianBench experiment. For each case, we compare the final output with the patient record and the checkpoint outcome, and report how the same checkpoint fared in the three CLINCH runs with the same base model. The modes describe how a failure arises rather than which checkpoint it affects, and a single run can exhibit more than one. Sections D.4 and D.5 then trace two complete tasks in which CLINCH keeps these links intact, with selected excerpts from the execution records.

D.1 Evidence-State Distortion

Evidence-state distortion occurs when the clinical status of evidence changes as it is carried through a task: a completed result becomes a pending one, or an unretrieved record becomes an absent one.

Completed negative tests reported as pending. In an infectious disease consultation for a patient with freshwater exposure during travel, the agent had to review stool studies and parasitology results before recommending targeted testing. The record contained a negative stool ova-and-parasite (O&P) examination from January 6, 2024 and a gastrointestinal PCR panel from December 30, 2023 with all pathogens not detected. The DeepSeek-V4-Pro baseline note instead reported:

> E. Parasitology Results (PENDING at time of consult):
> Prior Stool O&P (12/17/2023): CANCELED - improper collection container
> Gastroenteritis PCR Panel (12/29/2023): AWAITING RESULT

Two completed negative results thus entered the assessment as pending or unavailable, and the plan followed from that state: it called repeat stool O&P "essential given prior specimen rejection" and deferred definitive therapy to pending stool and serologic results. The data-retrieval and documentation checkpoints failed in all three baseline runs. CLINCH requires retrieval results and summaries to preserve status, source, and timestamp and to distinguish unexamined from negative findings (Section 3.3). The data-retrieval checkpoint passed in all three CLINCH runs, and one run completed the task.

Truncated results read as absent. Long record returns create a related risk. In the medication task of Appendix D.5, one GPT-5.6-Luna baseline run treated a truncated result as evidence that renal-function data were unavailable. Under CLINCH, the complete return was archived, and the agent recovered the relevant renal-function information by targeted read-back.

D.2 Decision–Knowledge Decoupling

Decision–knowledge decoupling occurs when the relevant patient findings and the applicable clinical rule are both available, but the decision does not follow from applying one to the other.

A surgical threshold applied to the wrong valve morphology. A cardiology task required evaluating aortic dimensions against established size thresholds and deciding whether referral to an aortic

specialty or surgery group was indicated. The baseline correctly reported an aortic root of 5.1–5.2 cm in a patient with a trileaflet valve, then stated:

> This exceeds the surgical threshold of 5.0 cm for non-syndromic, tricuspid aortic valve patients undergoing aortic root replacement (Class 1 recommendation).

For a non-syndromic trileaflet valve, the threshold is 5.5 cm; 5.0 cm applies to bicuspid valves or additional risk factors. The root was therefore below the surgical threshold, which calls for specialty referral and surveillance. The misapplied threshold changed the conclusion to "meets surgical threshold" and the plan to an urgent aortic surgery consultation, and the aortic assessment, aortic referral, and documentation checkpoints failed together. The aortic assessment failed in all three baseline runs; it passed in two of the three CLINCH runs, both of which completed the task. The pulmonary-toxicity case in Appendix D.4 shows the same pattern in adverse-event grading: the baseline retrieved the cough timeline, oxygen saturation, and prior imaging, yet assigned Grade 2, whereas CLINCH combined a decision-specific medical query with the patient evidence, recorded a Grade 1 interpretation, and passed all checkpoints (Section 3.4).

D.3 Unverified Completion

Unverified completion occurs when the agent's account of what was done departs from what the EHR records.

A planned request reported as a placed order. An endocrinology task concerned a breast cancer survivor on long-term anastrozole with progressive bone loss and required ordering anti-resorptive therapy and a follow-up DXA scan. The baseline placed the alendronate order and created a DXA `ServiceRequest`, but recorded the request with `intent` set to `plan` rather than `order`. Its final summary nonetheless reported:

> Both orders have been successfully placed.

The follow-up DXA checkpoint rejected the request ("Intent 'plan' should be 'order' "), so the reported completion was not supported by the record. This checkpoint failed in two of the three baseline runs and passed in all three CLINCH runs, where completion is checked against runtime-recorded actions and results rather than the agent's own account (Section 3.5).

Documented work without a corresponding action. Verification also applies to the note itself. In the medication task of Appendix D.5, final verification found that patient counseling had been described in the note but not represented by a communication action. The feedback reopened that item; CLINCH created the communication record and updated the note so that documentation and execution records agreed.

D.4 Patient-Specific Assessment of Drug-Related Pulmonary Toxicity

The task asked the agent to assess new chest CT findings in a patient receiving trastuzumab deruxtecan (Enhertu). The agent had to relate the findings to the treatment timeline, respiratory symptoms, prior imaging, and vital signs; assess possible interstitial lung disease (ILD); recommend whether to continue or hold therapy; arrange pulmonary evaluation and follow-up testing; and produce a clinical assessment note.

The DeepSeek-V4-Pro baseline run passed 5 of 7 checkpoints, whereas the CLINCH run passed all 7. Both runs retrieved the relevant imaging and clinical history. In the two trajectories, the most salient difference was how the evidence was combined to assess the pulmonary finding and how that assessment was carried into the final note.

The baseline retrieved the April 28 chest CT, which described mild linear and ground-glass opacities, predominantly in the left lower lobe and also involving the lingula and right middle lobe. The radiologist favored atelectasis or an inflammatory/infectious process, while the chronic right-sided radiation changes were stable. The run also retrieved the relevant patient context: the chronic dry cough began on November 5, Enhertu was started on November 12, the patient denied dyspnea,

oxygen saturation was 99%, and earlier CT imaging showed no ILD. The final summary retained the temporal relationship:

> Key observation: The cough began approximately 1 week BEFORE Enhertu initiation (11/05/2022 vs 11/12/2022), initially in the context of a URI.

Although the baseline distinguished Grade 1 from Grade 2 disease, it gave the pre-existing cough too much weight when applying that distinction to this patient:

> Given that the cough is chronic, persistent for months, and interferes with sleep (nocturnal awakenings documented 02/03/2023), this is most consistent with GRADE 2 ILD/pneumonitis.

This interpretation used an established symptom as the principal support for a symptomatic grade of a newly suspected drug-related lung injury, while giving less weight to the symptom's onset before treatment, the absence of dyspnea or hypoxemia, and the mild, nonspecific nature of the new imaging findings. The resulting Grade 2 assessment was carried into the treatment recommendation, including immediate corticosteroid therapy. The note also expressed the worsening pathway inconsistently, with conflicting instructions about whether corticosteroids should be held or initiated. These inconsistencies corresponded to failures in the clinical-reasoning and documentation checkpoints.

After retrieving the core patient evidence, CLINCH submitted a decision-specific medical query during execution to assess whether to hold, modify, or continue Enhertu. The query included the patient context and a candidate plan that already proposed holding treatment pending further evaluation. The retrieval response summarized the relevant grading and management guidance:

> Grade 1 is defined as asymptomatic with radiographic findings only. It is recommended to temporarily discontinue T-DXd (S5).

The agent then combined this guidance with patient-specific evidence and recorded a Grade 1 interpretation in the workspace. The workspace preserved the cough timeline, stable oxygenation, comparison between the prior and current CT, alternative explanations, and unresolved conditions. The decision and associated work were represented as:

> **Decision:** HOLD Enhertu pending pulmonary workup.
> **Action:**
> 1. [planned] STAT pulmonology referral (target 1–2 wks)
> 2. [planned] Order PFTs with DLCO
> 3. [planned] Order infectious workup
> 4. [planned] Repeat CT chest in 4–6 wks

The state also retained the patient-specific interpretation that "Cough predates Enhertu 1 wk, weakening attribution." The workspace thereby linked the evidence to the Grade 1 assessment, the temporary treatment hold, the planned work, and the conditions for later treatment decisions. Retrieved guidance became part of a maintained decision state available to subsequent actions and documentation.

The planned actions were then recorded in the EHR. CLINCH created the pulmonary referral, pulmonary-function request, infectious workup, and follow-up CT request, and changed the structured Enhertu medication status to `on-hold`; the medication update was confirmed by read-back. The task outline subsequently recorded the decision node as executed together with its action results and follow-up conditions. The final note preserved the Grade 1 assessment, the unresolved differential, and the distinction between the current hold, the follow-up workup, and escalation if symptoms changed. In this trajectory, patient evidence, decision-specific clinical guidance, maintained task state, execution records, and final documentation remained connected across the task.

D.5 Diabetes Medication De-intensification After HbA1c Improvement

The second task required reassessment of diabetes therapy for a 72-year-old man whose HbA1c had fallen from 10.55% to 5.1%. The patient was receiving Janumet together with separate metformin

and glipizide. The task required an individualized glycemic target, an assessment of hypoglycemia risk, medication adjustment, follow-up testing, and a management note consistent with the actions taken.

This case uses GPT-5.6-Luna for both configurations.

Each of three independent baseline runs passed 5 of 6 checkpoints, with different omissions across runs. One note did not state a numeric individualized HbA1c target; another did not obtain or use eGFR; and a third treated a truncated result as evidence that renal-function data were unavailable. All three CLINCH runs passed all 6 checkpoints.

CLINCH brought the current and prior HbA1c values, age, retinopathy, renal function, and medication exposure into the same patient state:

> Hemoglobin A1c was 10.55% on 2021-02-13 and 5.10% on 2022-01-08, a 5.45 percentage-point decrease.

The agent then connected the laboratory trend to the treatment decision. It selected an individualized goal below 8.0% because the patient's age, retinopathy, and sulfonylurea exposure increased the consequences of hypoglycemia:

> The current hemoglobin A1c of 5.1% is below the individualized goal of less than 8.0%, selected because age 72, diabetic retinopathy, and sulfonylurea exposure increase the consequences of hypoglycemia.

The goal was carried into the medication plan. CLINCH stopped glipizide and the duplicate metformin order, continued Janumet as the treatment backbone, and ordered repeat HbA1c monitoring. The two medication changes were recorded in the EHR and confirmed by read-back.

This task also illustrates how long record returns are used during clinical work. Broad laboratory, medication, and history queries can contain relevant entries beyond the initial visible excerpt. CLINCH retained an index of the complete returned set and used targeted read-back to recover the relevant renal-function information. The workspace then kept the HbA1c trajectory, medication doses, renal function, comorbidities, executed medication changes, and follow-up plan available for later decisions and documentation.

After the initial note was prepared, final verification identified that counseling had been described in the document but had not yet been represented by a corresponding communication action. The feedback reopened that unfinished task item. CLINCH created the patient communication record, covering glucose monitoring, recognition and treatment of hypoglycemia, meal guidance, the stopped medications, and follow-up. It then updated the note so that the documented counseling, medication changes, follow-up order, and execution records agreed.

The final document stated that glipizide and the duplicate metformin prescription had been stopped, that Janumet should continue, and that glucose monitoring and repeat HbA1c assessment should follow. The trajectory connects individualized medical guidance, recovery of decision-relevant evidence, patient-state maintenance, medication execution, verification, and delivery. Across the three CLINCH runs, all task checkpoints were satisfied.

E Limitations and Future Directions

Dependence on the base model's clinical reasoning. CLINCH shapes how an agent gathers evidence, forms decisions, and completes execution; it does not change how the model interprets that evidence. The remaining failures therefore concentrate in clinical reasoning (Figure 3), and across the three base models the gain in task success follows their clinical-reasoning ability (Section 4.2). Even with the relevant records and guidance available, an agent can misattribute a symptom, misapply a clinical rule, or select an inappropriate treatment. Such an error is consistent with the patient record and with the actions that follow from it, so checks against recorded operations do not reveal it. Correcting it requires identifying the faulty premise, reassessing the evidence, revising the decision, and updating the actions and documentation that depend on it. Additional retrieval or review helps only when it exposes the mistaken interpretation; repeating the same reasoning leaves the error intact.

Inference cost. Maintaining workspace state, providing knowledge at decision points, and verifying completion require model calls beyond those of the baseline agent, so CLINCH uses more inference tokens (Tables 1 and 3; Appendix B). This cost follows from making evidence, decisions, and execution explicit and checkable, and additional token usage alone does not reproduce the resulting gains (Table 2). Much of the cost can be targeted rather than removed: review and knowledge support could be triggered by the uncertainty or downstream consequences of a decision, and auxiliary steps such as compression and consistency checking could be delegated to smaller models.

Evaluation in simulated EHR environments. PhysicianBench and MedAgentBench provide reproducible EHR environments with scored clinical tasks, which makes controlled comparison across models and configurations possible. Clinical deployment involves more: real records are larger and noisier, workflows span multiple clinicians, and any action in a live system requires privacy safeguards and clinician oversight. Our results therefore measure task completion under controlled conditions rather than clinical benefit.

Future directions. These limitations point to three directions. Because CLINCH does not depend on a particular base model, the model can be replaced as stronger or clinically specialized models become available; given that gains grow with clinical-reasoning ability, such models should convert more of the harness's execution support into task success. The harness can also inform training. Its workspace and verification records show where a trajectory fails, including which evidence was used, which decision was made, and which requirement remained unmet. These records can supervise post-training within the harness and domain-adaptive training on clinical evidence interpretation, conditional decision-making, and revision of earlier judgments, so that the model learns both clinical reasoning and how to act on the workspace and review feedback it will receive. Finally, evaluation in hospital settings, first through retrospective replay on de-identified records with clinician review of agent decisions and then through clinician-in-the-loop studies, would test whether gains in task completion carry over to clinical practice.