CLINSEEKAGENT: AUTOMATING MULTIMODAL EVIDENCE SEEKING FOR AGENTIC CLINICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) and agentic systems have shown promise for clinical decision support, but existing works largely assume that evidence has already been curated and handed to the model. Real-world clinical workflows instead require agents to actively seek, iteratively plan, and synthesize multimodal evidence from heterogeneous sources. In this paper, we introduce **ClinSeekAgent**, an automated agentic framework for dynamic multimodal evidence seeking that shifts the paradigm from passive evidence consumption to active evidence acquisition. Given a clinical task and access to raw patient data, ClinSeekAgent queries longitudinal EHR records and invokes imaging analysis tools when relevant; refines its hypotheses as new information emerges; and integrates the collected evidence into grounded clinical decisions. We validate ClinSeekAgent in two roles: as an inference-time agent for frontier LLMs and as a training pipeline for open-source models. For inference-time validation, we construct **ClinSeek-Bench**, which covers 13 text-only EHR task categories and 6 multimodal EHR–chest X-ray task categories, and pairs each question with both a *Curated Input* of pre-selected evidence and the raw patient data for automated evidence seeking. Without any benchmark-specific curation, ClinSeekAgent significantly outperforms curated input for both Claude Opus 5 and GPT-5.6 Sol, improving overall F1 from 61.7 to 63.9 and from 57.8 to 59.3 on text-only EHR tasks, and from 66.5 to 68.8 and from 60.0 to 62.5 on multimodal tasks. The largest gains occur in text-based decision-making and multimodal phenotype prediction, and ablations show that they come from multi-round evidence seeking rather than from access to more of the patient record. For training-time validation, we fine-tune Qwen3.5-35B-A3B on ClinSeek-Agent trajectories generated by GPT-5.6 Sol. The resulting **ClinSeek-35B-A3B** reaches 35.6 average F1 on AgentEHR-Bench (**+16.1** over its base model), outperforming all evaluated open-source models, including ones more than ten times larger, as well as Claude Opus 5. We will release our model, data, and code to facilitate future research.

1 INTRODUCTION

Recent large language models (LLMs) and agentic systems have shown strong potential in medical question answering, diagnostic reasoning, and clinical decision support (Wu et al., 2025a; Kim et al., 2024; Fallahpour et al., 2025; Yao et al., 2022; Schmidgall et al., 2024; Zhang et al., 2023). However, many existing medical-agent settings remain overly simplistic, deviating from real-world clinical workflows. They often rely on general medical knowledge (Wu et al., 2025b; Zhao et al., 2025) or short organized patient vignettes, whereas real-world clinical decision support requires actively seeking patient-specific longitudinal information from raw Electronic Health Record (EHR) tables (Johnson et al., 2016; 2023) and visual clues from medical imaging (Johnson et al., 2019). Such a limitation is particularly salient for clinical decision support, where the key challenge is not only to reason over given evidence, but also to decide where to retrieve evidence from, what evidence to retrieve, and how different pieces of evidence can be integrated into a grounded decision.

A growing line of EHR-specific work has moved closer to this goal by adapting LLMs to structured patient records and multimodal clinical data (Liao et al., 2025; Bae et al., 2023; Elsharief et al., 2025; Vasilev et al., 2025). For example, recent EHR reasoning pipelines convert structured tables into textual contexts, retrieve task-related entities, and synthesize reasoning data from pre-extracted

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.

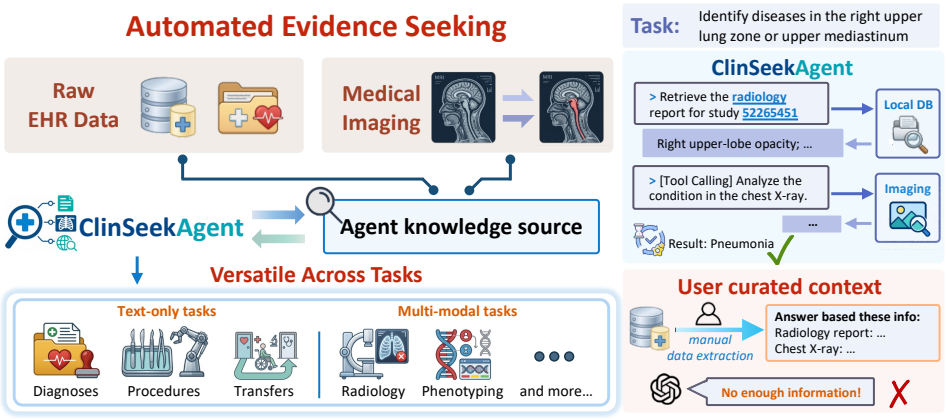

Figure 1: **ClinSeekAgent Overview.** ClinSeekAgent is an automated agentic evidence-seeking pipeline. It interacts with heterogeneous data sources to enable multimodal evidence seeking for clinical decision support. Compared with prior user-curated context settings, ClinSeekAgent is more flexible by acquiring richer information and knowledge from diverse tools.

patient information (Liao et al., 2025; Kweon et al., 2024). Multimodal clinical benchmarks also combine EHRs and medical images to support realistic prediction and question-answering tasks (Bae et al., 2023; Elsharief et al., 2025). These efforts are valuable, but they still largely *depend on a fixed evidence-packaging process* before inference: the relevant patient context is selected by benchmark construction, human priors, or task-specific rules. Recent studies of EHR agents have started to expose models to database tools (Liao et al., 2026; Jiang et al., 2025; Chen et al., 2025; Qian et al., 2026; Lee et al., 2025; Shi et al., 2024), but they remain limited in task scope, tool coverage, or modality support. As a result, there is a need for a general agentic framework that automates the evidence search process, rather than assuming that the evidence has already been surfaced.

To address this need, we introduce **ClinSeekAgent**, an automated agentic framework for dynamic multimodal evidence seeking in clinical reasoning (Sec. 2). As shown in Fig. 1, ClinSeekAgent differs from existing curated-evidence pipelines in that it does not passively consume a fixed evidence package prepared before inference. Instead, given a clinical query and access to heterogeneous clinical data sources, ClinSeekAgent actively gathers evidence through **(1)** raw EHR retrieval and **(2)** medical imaging tools (Sec. 2.2), iteratively refining its actions as new evidence emerges (Sec. 2.3). This enables the agent to recover longitudinal and multimodal patient evidence that fixed curated contexts may miss. For example, when asked to identify the diseases in the right upper lung zone or the upper mediastinum, ClinSeekAgent retrieves the radiology report of the study from the local EHR database, which notes a right upper-lobe opacity, and then analyzes the chest X-ray with an imaging tool to answer *pneumonia*, while the same model under the curated-context setting fails due to missing critical evidence (Fig. 1).

We validate ClinSeekAgent first as an inference-time pipeline through **ClinSeek-Bench**, an evaluation suite that reformulates existing EHR and multimodal clinical tasks into paired curated-context and agentic settings (Sec. 3.1). For each sample, the source benchmark (Liao et al., 2025; Elsharief et al., 2025; Bae et al., 2023) provides a task-specific evidence package that was originally used as input to the model. We preserve this original setting as *Curated Input*, where the model answers directly from the provided patient context. We then construct a paired *Automated Evidence-Seeking* setting by removing this context and providing only the patient identifier, raw data access, and ClinSeekAgent tools, requiring the model to retrieve and integrate the necessary evidence by itself. As a result, each sample in ClinSeek-Bench evaluates the same task and answer label under two modes: answering from pre-selected evidence, and autonomously seeking evidence from raw clinical data. ClinSeek-Bench includes **13** text-only EHR task categories derived from EHR-Bench (Liao et al., 2025), spanning risk prediction (mortality, readmission, length of stay, and escalation of care) and decision making (diagnosis, medication, procedure, laboratory and microbiology, vitals and measurements, imaging report, orders, encounter and flow, and next event), and **6** multimodal task categories adapted from EHRXQA (Bae et al., 2023) and MedMod (Elsharief et al., 2025) (chest X-ray finding presence, finding enumeration, and change comparison, clinical deterioration, mortality, and phenotype) (Tab. 5).

Our inference-time experiments show that ClinSeekAgent can improve over fixed curated inputs when paired with capable agentic models. On text-only EHR tasks, overall F1 rises from 61.7 to 63.9 for Claude Opus 5 and from 57.8 to 59.3 for GPT-5.6 Sol when moving from *Curated Input* to *ClinSeekAgent* (Sec. 3.2, Fig. 3). The largest gains occur in text-based decision-making tasks and multimodal phenotype prediction, where the agent must piece together evidence from longitudinal records and chest radiographs. On the multimodal benchmark, these models improve from 66.5 to 68.8 and from 60.0 to 62.5 overall F1, respectively (Fig. 3), suggesting that active evidence acquisition can recover clinical signals that fixed curated contexts may miss.

Ablations show that iterative, multi-round evidence seeking drives the gains (Sec. 3.3). The variation across host models further suggests that this process depends on the agent's ability to plan and execute long-horizon tool use (Sec. B). Therefore, we further validate ClinSeekAgent as a training pipeline for open-source clinical agents (Sec. 4). Using ClinSeekAgent, we collect high-quality clinical search trajectories from a strong teacher model and fine-tune Qwen3.5-35B-A3B (Qwen Team, 2026), resulting in **ClinSeek-35B-A3B**. On the existing AgentEHR-Bench (Liao et al., 2026), ClinSeek-35B-A3B improves over its base model from 19.5 to 35.6 average F1, outperforms all evaluated open-source models including Qwen3.5-397B-A17B (28.1), and exceeds Claude Opus 5 (31.6); among proprietary models only GPT-5.6 Sol scores higher (38.8) (Fig. 2, Tab. 2). These results show that ClinSeekAgent is not only effective as an inference-time pipeline, but can also serve as a scalable training pipeline for distilling clinical evidence-seeking behavior into open-source models (Sec. 4.3).

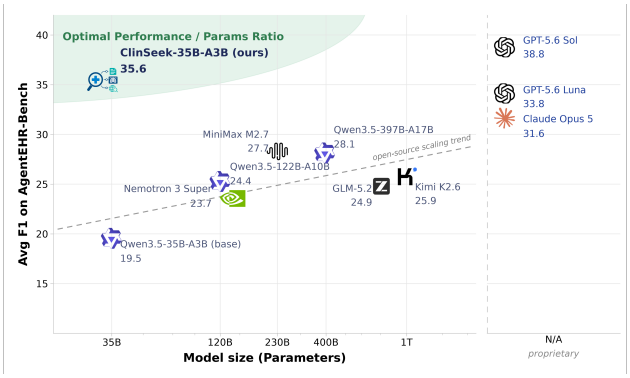

Figure 2: **Performance–model size comparison on AgentEHR-Bench.** ClinSeek-35B-A3B outperforms all open-source models, as well as Claude Opus 5.

2 ClinSeekAgent: Multimodal Evidence-Seeking Pipeline

2.1 Task Formulation and Interaction Protocol

Each clinical task instance is defined as:

$$x = (p, t, q, \mathcal{M}, \mathcal{Y}), \tag{1}$$

where p is the patient identifier, t is the reference timestamp or prediction time, q is the clinical task instruction, $\mathcal{M}$ denotes optional modality-specific metadata such as image paths, and $\mathcal{Y}$ denotes the answer schema or candidate label space when available. During inference, the model is not given the curated patient context used by the source benchmark. Instead, it receives x and access to the ClinSeekAgent tool space, and invokes tools to retrieve evidence needed for the task. At step k, the model π_θ observes the task instance and the previous interaction history

$$h_{k-1} = \{(a_1, o_1), \ldots, (a_{k-1}, o_{k-1})\}, \tag{2}$$

and either invokes another tool or terminates the answering process as its next action:

$$a_k \sim \pi_\theta(\cdot \mid x, h_{k-1}). \tag{3}$$

If a_k is a tool call, the environment returns an observation o_k; otherwise, the model outputs the final prediction $\hat{y}$ following the specified answer schema. For EHR-related tasks, the agent first loads the patient database with `ehr.load_ehr`, and all EHR queries are restricted to records available *before the reference timestamp* t.

2.2 Multi-Source Tool Space

ClinSeekAgent exposes a unified tool space with 22 tools across two complementary evidence sources: *EHR retrieval* and *medical image analysis*. Specifically, it provides 16 EHR tools for

accessing patient-specific longitudinal records, including schema inspection, temporal, keyword-, and value-based retrieval, SQL-based querying, and candidate-term grounding; and 6 image tools for extracting visual evidence through DICOM preprocessing, image visualization, chest X-ray classification, report generation, phrase grounding, and anatomical segmentation. The complete tool list is provided in Appendix F.

2.3 Agentic Evidence-Seeking Trajectories

ClinSeekAgent represents each run as an open-ended evidence-seeking trajectory:

$$\tau = \left(x,\ (a_k, o_k)_{k=1}^{K},\ \hat{y}\right)$$

where x is the task instance, a_k is a tool action, o_k is the corresponding tool observation, and $\hat{y}$ is the final answer. The trajectory records both the final prediction and the sequence of evidence-seeking decisions that produced it.

Unlike rule-based retrieval pipelines, ClinSeekAgent does not impose an ordering over evidence sources. Depending on the task, the model may begin with schema inspection, EHR querying, image analysis, or candidate retrieval, and may interleave these tools across multiple turns. Thus, ClinSeekAgent standardizes the environment and tool interface, while the evidence-seeking policy is induced by the agentic model.

3 Inference-time Validation: Curated Input vs Automated Evidence Seeking

3.1 ClinSeek-Bench Construction

We construct ClinSeek-Bench to validate ClinSeekAgent as an inference-time evidence-seeking pipeline. Each example is paired into two settings with the same task definition and answer label: *Curated Input*, where the model answers from the evidence package provided by the source EHR reasoning benchmark (Liao et al., 2025; Bae et al., 2023; Elsharief et al., 2025), and *Automated Evidence-Seeking*, where this context is removed and the model must retrieve evidence from raw clinical data using ClinSeekAgent tools.

Source Benchmarks. ClinSeek-Bench has two parts. ClinSeek-Bench-Text is built from EHR-Bench (Liao et al., 2025), which contains 45 EHR analysis subtasks over MIMIC-IV; we randomly sample 40 examples from each subtask, resulting in 1,800 examples. We group the subtasks into 13 task categories: 4 for risk prediction (mortality, readmission, length of stay, and escalation of care) and 9 for decision making (diagnosis, medication, procedure, laboratory and microbiology, vitals and measurements, imaging report, orders, encounter and flow, and next event). ClinSeek-Bench-MM is built from EHRXQA (Bae et al., 2023) and MedMod (Elsharief et al., 2025), both of which pair MIMIC-IV EHRs with MIMIC-CXR chest radiographs. After reconstructing the official examples and preserving their task definitions, splits, labels, and EHR-CXR pairing rules, we obtain 989 examples across 6 task categories: CXR finding presence, CXR finding enumeration, CXR change comparison, clinical deterioration, mortality, and phenotype. Tab. 5 describes each task category.

Curated Input Data Collection. We preserve the *original EHR reasoning benchmark inputs* as the *Curated Input* setting. These inputs reflect the evidence-packaging process of the source benchmarks, where task-relevant patient information is selected before inference. For EHR-Bench, the original setting uses rule-based templates to convert recent patient events into instruction-answer samples: models observe up to 100 events from the past 24 hours and predict either the next clinical event or a future risk outcome. For EHRXQA and MedMod, we keep the original task-specific EHR context, selected CXR studies, image identifiers, labels, and pairing rules from the official repositories.

Automated Evidence-Seeking Data Generation. We convert each curated example into an *Automated Evidence-Seeking* example by removing the curated context while keeping the same task instruction and answer label. The model is instead given the patient identifier, prediction-time cutoff, optional linked CXR identifiers, and access to ClinSeekAgent tools. For EHR-Bench, we use the timestamp of the last event in the original input as the reference cutoff, allowing the agent to access the patient's full raw EHR history before that time. For multimodal tasks, we preserve the original

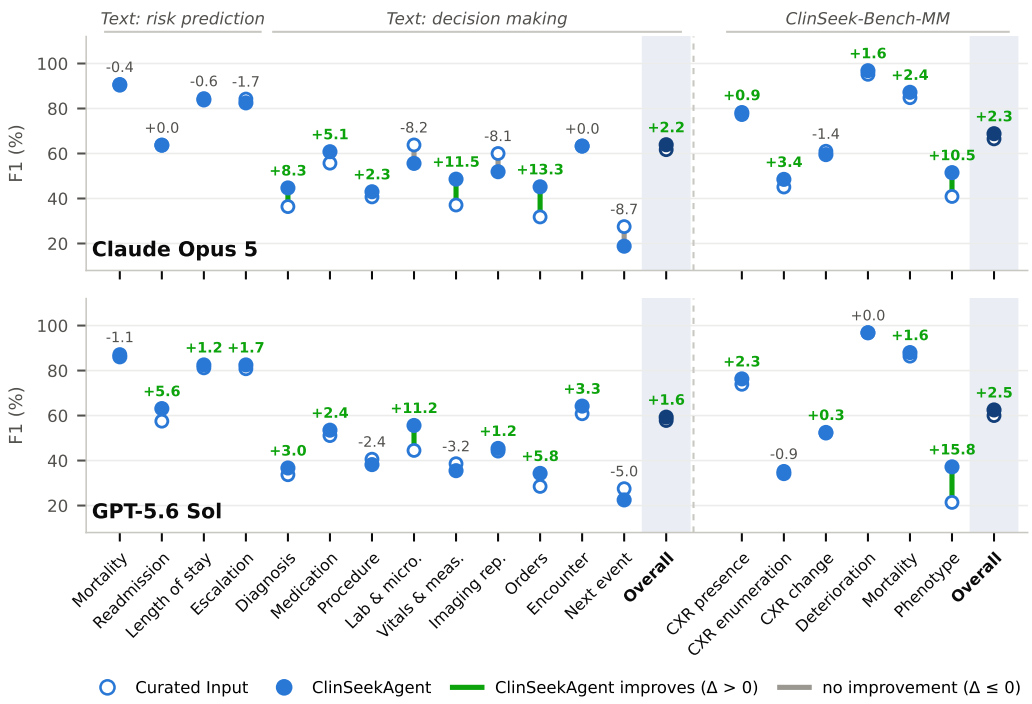

Figure 3: **ClinSeekAgent vs. Curated Input by task category** (sample-wise F1, %) for Claude Opus 5 (top) and GPT-5.6 Sol (bottom) on ClinSeek-Bench-Text (left) and ClinSeek-Bench-MM (right). Hollow markers are the curated input, filled markers are ClinSeekAgent, and the label above each pair gives their difference; green marks categories where ClinSeekAgent improves. Categories follow the task taxonomy in Sec. B; the shaded column is the overall score.

patient-level task, label, and valid EHR-CXR linkage, but require the agent to retrieve EHR evidence and invoke imaging tools when needed.

Temporal Filtering. Both data types are filtered by the same strict temporal policy: the model can only see records that were available before the prediction cutoff. Because diagnosis and procedure codes are assigned when an encounter is completed, they are treated as unavailable until then, and the codes of the encounter being predicted are always hidden. More details are provided in Sec. G.

3.2 Main Results: ClinSeekAgent Improves Frontier Agentic Models

Experimental Settings. We compare two inference modes on the paired data of Sec. 3.1. In the Curated Input mode, the pre-prepared evidence package is rendered in the prompt and the model answers from it. ClinSeekAgent runs in the Automated Evidence-Seeking setting: given only the patient identifier and the prediction cutoff, the model retrieves evidence from the raw record with the tools of Sec. 2.2 for up to 200 rounds before calling `finish`. Both modes run with the same reasoning budget and one run per question; the remaining settings are listed in Sec. H. Since most tasks are multi-label prediction, we use sample-wise F1 (%) as the unified metric across all tasks; its computation is detailed in Sec. H.1. We evaluate two frontier models, Claude Opus 5 (Anthropic, 2026b) and GPT-5.6 Sol (OpenAI, 2026), and report both modes and their difference per task category in Fig. 3; results for Claude Opus 4.6 (Anthropic, 2026a), GLM-5.2 (Team, 2026), and Kimi K2.6 (Team et al., 2026) are given in Sec. B.

Automatic evidence seeking replaces hand-designed curation. As shown in Fig. 3, switching from Curated Input to ClinSeekAgent raises the overall F1 of Claude Opus 5 from 61.7 to 63.9 on ClinSeek-Bench-Text and from 66.5 to 68.8 on ClinSeek-Bench-MM, and that of GPT-5.6 Sol from 57.8 to 59.3 and from 60.0 to 62.5, respectively. The gains are broad rather than driven by a single task: ClinSeekAgent improves 23 of the 38 model–category pairs, including 13 of 19 categories for GPT-5.6 Sol and 10 of 19 for Claude Opus 5, with the largest gains reaching +13.3 (orders) and +11.5

Table 1: **Ablating the two ingredients of ClinSeekAgent** (overall F1, %). Rounds: model calls per question. In parentheses: that setting – ClinSeekAgent, red for a decrease and green otherwise, bold when the paired-bootstrap 95% CI excludes zero.

Setting	Record	Tools	Rounds	ClinSeek-Bench-Text Opus 5	ClinSeek-Bench-Text Sol	ClinSeek-Bench-MM Opus 5	ClinSeek-Bench-MM Sol
Curated Input	recent	–	1	61.7 (**–2.2**)	57.8 (**–1.6**)	66.5 (**–2.3**)	60.0 (**–2.5**)
Full-Record Input	full	–	1	62.6 (–1.3)	59.6 (+0.3)	66.6 (**–2.2**)	60.3 (**–2.3**)
Single-Pass Retrieval	full	✓	2	58.5 (**–5.4**)	57.6 (**–1.7**)	62.6 (**–6.1**)	61.2 (–1.4)
ClinSeekAgent	full	✓	5–10	63.9	59.3	68.8	62.5

(vitals and measurements) for Opus 5 and +15.8 (phenotype) and +11.2 (laboratory and microbiology) for Sol; the largest decreases (-8 to -9 points for Opus 5) fall in the three smallest categories, with 40–80 questions each.

ClinSeekAgent achieves this *without the benchmark-specific curation rules* that Curated Input relies on: the agent decides what to retrieve from the raw record. Models with stronger agentic capability benefit more from the pipeline: Claude Opus 4.6, GLM-5.2, and Kimi K2.6 improve by smaller margins, +5.1 (Opus 4.6) and +2.0 (Kimi K2.6) on ClinSeek-Bench-MM and +0.7 (GLM-5.2) on ClinSeek-Bench-Text, and only the gain of Opus 4.6 is significant (Sec. B).

Reasoning-intensive tasks benefit the most. The largest improvements appear on the decision-making tasks of ClinSeek-Bench-Text and on phenotype prediction in ClinSeek-Bench-MM. Decision-making tasks ask which clinical events come next, e.g., the diagnoses, medications, or orders, so the model must relate the findings in the record to the patient's course and enumerate a set of concrete items. Predicting the next diagnoses, medications, and orders improves for both models (+8.3/+5.1/+13.3 for Opus 5, +3.0/+2.4/+5.8 for Sol); in addition, Opus 5 gains +11.5 on vitals and measurements and Sol gains +11.2 on laboratory and microbiology. Phenotype prediction asks for the set of chronic and acute conditions of an ICU stay, which has to be assembled from the radiograph together with several EHR tables; it improves by +10.5 for Opus 5 and +15.8 for Sol. Risk prediction, in contrast, only asks for a binary label and is already close to saturation under Curated Input; ClinSeekAgent maintains this high level on both benchmarks (ClinSeekAgent vs. Curated Input: 81.7 vs. 82.2 for Opus 5 and 79.6 vs. 78.2 for Sol over the 720 risk-prediction questions of ClinSeek-Bench-Text, and 87–97 F1 on decompensation and mortality in ClinSeek-Bench-MM).

The improvements are stable across multiple runs. To check that the differences above are not run-to-run noise, we repeat both inference modes for GPT-5.6 Sol four more times on both benchmarks (Sec. C). Across the five runs, the overall F1 has a standard deviation of at most 0.4 points within each mode (ClinSeekAgent vs. Curated Input: 59.4$\pm$0.1 vs. 57.4$\pm$0.2 on ClinSeek-Bench-Text; 62.4$\pm$0.4 vs. 59.7$\pm$0.3 on ClinSeek-Bench-MM), and ClinSeekAgent beats Curated Input in every run, by +1.6 to +2.3 and +2.3 to +3.3 points. At the question level, the runs agree closely: Krippendorff's α of per-question F1 across runs is 0.87–0.91 for both modes (metric definitions in Sec. C.1).

3.3 Where Does the Gain Come From?

ClinSeekAgent differs from Curated Input in two ways: (1) the model can access the whole record instead of a curated window; and (2) it gathers evidence from the record over multiple rounds of tool calls and reasoning rather than in a single step. To ablate the two factors, we add two settings that provide extended record access in different ways but without the multi-round loop (Tab. 1). *Full-Record Input* is a variant of curated input: the curated evidence package is replaced by the patient's full pre-cutoff record, rendered into the prompt in the same way and truncated oldest-first when it does not fit in the 100k-token context window. *Single-Pass Retrieval* is a variant of ClinSeekAgent: the agent keeps the same tools and access to the whole record, but must issue all of its retrieval calls in one turn and answer from what comes back. All four settings share the temporal policy of Sec. 3.1.

Extended record access alone does not help. Neither way of extending record access recovers the gain of ClinSeekAgent. Full-Record Input gives the model the whole record instead of the curated window, yet its overall F1 exceeds that of Curated Input by only +0.8/+1.9 points on ClinSeek-Bench-Text and +0.1/+0.3 on ClinSeek-Bench-MM (Opus 5/Sol). The length of the record makes this concrete. Fig. 4 stratifies the ClinSeek-Bench-Text questions by the length of the patient's full pre-cutoff record: below 30k tokens ClinSeekAgent is on par with both prompt-based settings, and

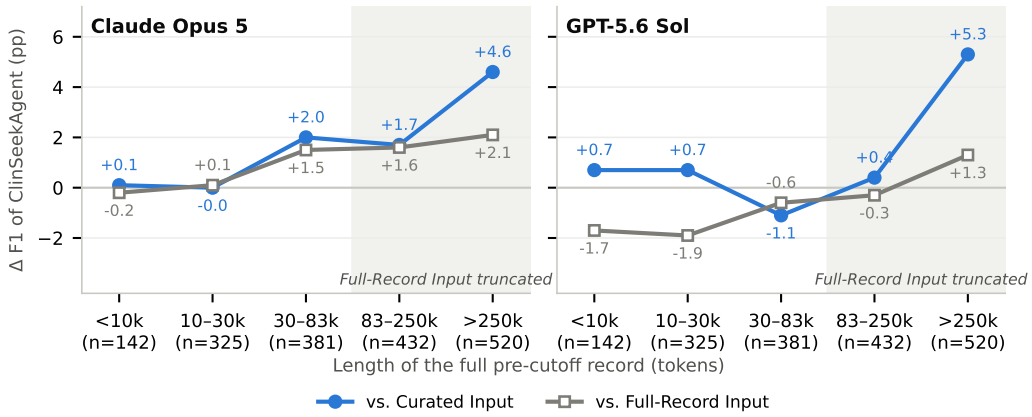

Figure 4: **Gain of ClinSeekAgent grows with the length of the patient record** (ClinSeek-Bench-Text, F1 points). Questions are binned by the token count of the full pre-cutoff record; in the shaded bins the record does not fit in the 100k-token context window of Full-Record Input and is truncated.

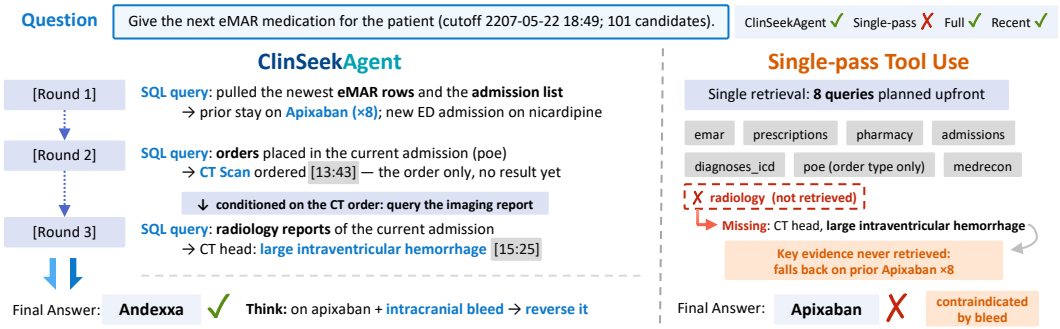

Figure 5: **Case study: ClinSeekAgent vs. Single-Pass Retrieval**.

its margin grows with record length to +4.6/+5.3 over Curated Input and +2.1/+1.3 over Full-Record Input on records above 250k tokens, where the prompt can no longer hold the record and a fixed curation rule has the most to miss. A case study (Fig. 6 in Sec. E) shows that having the record in the prompt is not the same as finding the deciding row: the unit to which the same patient was admitted after a comparable orthopedic visit two weeks earlier sits at 26% of the untruncated Full-Record prompt, yet the model answers with a string near the end of the prompt that is not a care unit in the record. ClinSeekAgent instead notices that the admission transfer is still pending, widens its time window, and retrieves the precedent.

Single-Pass Retrieval can access the entire record, but it is no better than Curated Input either: its overall F1 differs from that of Curated Input by –3.2/–0.1 on ClinSeek-Bench-Text and –3.9/+1.2 on ClinSeek-Bench-MM, with no significant gain. Reading a longer prompt or issuing all queries in one turn thus adds little over Curated Input.

Multi-round reasoning is the source of the gain. Single-Pass Retrieval and ClinSeekAgent differ only in whether retrieval is iterative, so their gap measures the effect of multi-round reasoning alone: +5.4/+1.7 on ClinSeek-Bench-Text and +6.1/+1.4 on ClinSeek-Bench-MM. Each round lets the agent condition its next query on what the previous one returned, which neither variant can do. Fig. 5 shows this on one question: the first round returns the patient's prior anticoagulant and a pending CT order, so the agent requests the CT report, finds a large intracranial bleed, and predicts the reversal agent. Single-Pass Retrieval had to plan all of its queries before seeing any result and thus without that pointer; it never retrieved the report and predicted the anticoagulant that the bleed contraindicates.

4 Training-time Validation: Teaching Open Models to Use ClinSeekAgent

Sec. 3.3 shows that the gain of ClinSeekAgent comes from multi-round evidence seeking, and Sec. 3.2 shows that this gain is consistent only for models with strong agentic capability. We therefore ask

Table 2: **Results on AgentEHR-Bench** (F1, %; 100 questions per task). All models run ClinSeek-Agent under the same temporal policy. Bold marks the best model in each group. Last two rows: difference of ClinSeek-35B-A3B from its base model and from its teacher, red for a decrease and green otherwise; the gains over the base model are all significant under a paired bootstrap (95% CI).

Model	Diagnoses	Labs	Microbiology	Procedures	Transfers	Avg.
Closed-source models						
GPT-5.6 Sol (teacher)	47.0	43.1	**31.1**	**45.5**	**27.3**	**38.8**
GPT-5.6 Luna	43.7	37.8	24.8	37.9	24.9	33.8
Claude Opus 4.6	**51.1**	42.1	25.4	28.7	18.6	33.2
Claude Sonnet 4.6	49.5	36.8	22.9	25.9	24.5	31.9
Claude Opus 5	48.0	**44.5**	23.2	20.5	22.0	31.6
Open-source models						
Qwen3.5-397B-A17B	38.4	30.9	21.2	26.2	23.7	28.1
MiniMax M2.7	39.4	22.3	22.1	32.0	22.9	27.7
Kimi K2.6	42.8	20.2	18.5	24.6	23.3	25.9
GLM-5.2	**43.8**	20.4	16.0	24.0	20.4	24.9
Qwen3.5-122B-A10B	32.6	23.7	21.6	23.8	20.5	24.4
Nemotron 3 Super	29.5	27.5	23.1	24.9	13.3	23.7
Qwen3.5-35B-A3B (base)	28.4	16.5	19.5	17.9	15.1	19.5
ClinSeek-35B-A3B (ours)	41.7	**40.4**	**27.1**	**41.3**	**27.4**	**35.6**
Ours – base	+13.3	+23.9	+7.6	+23.4	+12.3	+16.1
Ours – teacher	–5.3	–2.7	–4.0	–4.2	+0.1	–3.2

whether ClinSeekAgent can also serve as a training pipeline that transfers this capability to a smaller open-source model. Specifically, we use the trajectories that ClinSeekAgent produces with a strong teacher as supervision, and test whether the student learns the evidence-seeking process rather than only the final answers.

4.1 Experimental Settings

We use GPT-5.6 Sol as the teacher to run ClinSeekAgent on the 3,000 questions of the AgentEHR-Bench (Liao et al., 2026) training split, and fine-tune Qwen3.5-35B-A3B on the resulting trajectories with supervised fine-tuning. The teacher's intermediate reasoning is placed in the student's native thinking block, and the tool calls are rendered in its native tool-call format (Sec. I). We evaluate on its test split, using 100 questions for each of its five tasks; none of these questions appears in the training data. All models run ClinSeekAgent with the same tools, temporal policy, and inference settings as in Sec. 3.2; the base and fine-tuned models are served locally with identical decoding settings.

4.2 ClinSeek-35B-A3B Achieves Open-Source State-of-the-Art

Tab. 2 reports the results on AgentEHR-Bench. Training on ClinSeekAgent trajectories raises the average F1 of Qwen3.5-35B-A3B from 19.5 to 35.6 (**+16.1**). The model improves on all five tasks, by +7.6 on *Microbiology* up to +23.9 on *Labs* and +23.4 on *Procedures*, and every gain is significant under a paired bootstrap. ClinSeek-35B-A3B outperforms every open-source model we evaluate, including Qwen3.5-397B-A17B, which has about eleven times more parameters, by +7.5 points. It also exceeds Claude Opus 4.6 by +2.4 points and Claude Opus 5 by +4.0, and reaches 92% of its teacher GPT-5.6 Sol (35.6 vs. 38.8), the only closed-source model that scores higher; on *Transfers* it matches the teacher (27.4 vs. 27.3). These results show that ClinSeekAgent trajectories transfer long-horizon EHR evidence seeking to a smaller open-source model.

4.3 What Does ClinSeek-35B-A3B Learn?

To understand where the gain comes from, we compare the 500 trajectories of ClinSeek-35B-A3B with those of its base model (Sec. J). Both precision and recall improve (18.6 to 32.4 and 31.1 to 53.3). Two changes in behavior accompany this gain:

- **More complex SQL queries.** The student learns to write SQL with joins, aggregation, or time filters, which it uses in 53% of the questions compared with 0.8% for the base model; it typically first summarizes what the record contains and then drills into the current admission.
- **Reasoning grounded on the patient record.** The student surveys the record before narrowing down: it reads twice as many of the available tables (72% vs. 36%), issues a larger share of targeted follow-up queries (36% vs. 13%), reads imaging or discharge notes in 74% vs. 12% of the questions, and refers to the current admission in its reasoning in 80% vs. 39%.

The student thus learns the evidence-seeking process using ClinSeekAgent, not only its final answers.

5 Related Work

Medical Reasoning with Curated Evidence. Recent medical LLMs have shown strong performance in medical question answering and diagnostic reasoning (Tu et al., 2024; Ossowski et al., 2025; Huang et al., 2025b;a; Li et al., 2023; Shi et al., 2026; Wang et al., 2026), demonstrating that LLMs can encode medical knowledge and reason over clinical scenarios. These settings differ from real-world clinical decision support, where models must first identify and retrieve task-relevant evidence from longitudinal patient records, rather than only reason over provided patient vignettes (Jin et al., 2019; 2021), summarized clinical notes (Kweon et al., 2024), or task-specific patient contexts (Yu et al., 2025; Zuo et al., 2025). Recent EHR and multimodal clinical benchmarks move closer to real clinical data by grounding tasks in structured patient records, radiology reports, and medical images (Liao et al., 2025; Elsharief et al., 2025; Bae et al., 2023). However, these works still largely follow the curated-evidence paradigm: task-relevant records, reports, or multimodal inputs are selected before inference. In contrast, ClinSeekAgent focuses on automating this evidence-seeking step, allowing the agent to dynamically query raw EHR tables and invoke medical image analysis tools.

Agentic Evidence Seeking over Clinical Data. Recent medical agent systems have begun to move beyond single-pass reasoning by introducing tool use, search, and multi-agent collaboration into clinical tasks. MDAgents adaptively organizes multiple LLM agents for medical decision making (Kim et al., 2024), while DeepMed (Wang et al., 2026) and Meissa (Chen et al., 2026) train medical agents to perform multi-step evidence search or interaction for medical reasoning. Closer to EHR-based decision support, AgentEHR (Liao et al., 2026), MedAgentBench (Jiang et al., 2025), and FHIR-AgentBench (Lee et al., 2025) evaluate agents in interactive clinical record environments, requiring models to retrieve patient information and reason over structured records. AgentClinic further studies tool-using agents in simulated multimodal clinical environments (Schmidgall et al., 2024). These works demonstrate the promise of agentic clinical AI, but their evidence-seeking processes are typically limited to medical knowledge search, multi-agent discussion, EHR-only interaction, or simulated clinical tools. In contrast, ClinSeekAgent provides a unified multimodal evidence-seeking pipeline over raw EHR tables and medical image analysis tools, and further validates this pipeline both at inference time and through trajectory-based training of open-source agents.

6 Conclusion

In this paper, we introduce ClinSeekAgent, an automated agentic framework for dynamic multimodal evidence seeking in clinical decision support, which lets an agentic model gather evidence from raw EHR records and chest X-ray analysis tools instead of relying on curated inputs. As an inference-time pipeline, ClinSeekAgent improves two frontier models over curated inputs on both text-only and multimodal clinical tasks without any benchmark-specific curation, with the largest gains on reasoning-intensive tasks. Ablations show that these gains come from multi-round evidence seeking rather than from access to more of the patient record, and that they grow with the length of the record. As a training pipeline, ClinSeekAgent trajectories teach an open-source model, ClinSeek-35B-A3B, to write more complex queries and to ground its reasoning in the patient record, outperforming all evaluated open-source models on AgentEHR-Bench. Our results suggest that moving from passive evidence consumption to active evidence acquisition is a promising direction for building more flexible, grounded, and capable clinical AI agents.

AI USE STATEMENT

Generative AI plays two distinct roles in this work. First, large language models are part of the systems and pipelines we study, as described in the paper: frontier LLMs act as the agent of ClinSeekAgent and as the model under Curated Input (Sec. 3); GPT-5.6 Sol generates the ClinSeekAgent trajectories used as synthetic supervision for training ClinSeek-35B-A3B (Sec. 4); and Claude Sonnet 4.6 serves as an answer judge for single-label multimodal questions (Sec. H.1). Second, the authors used an AI coding assistant (Claude, through Claude Code) for research support. It was used to implement and debug experiment code, including the agent runners, tool servers, data-processing and scoring scripts; to write the scripts that compute the reported statistics (paired bootstrap confidence intervals, multi-run consistency, record-length stratification, and trajectory behavior analysis); and to draft and edit parts of the paper. We have not used generative AI tools to propose the research idea or hypotheses of this work, to create benchmark labels, or to develop theoretical models or mathematical claims; translation, surveys, interviews, and transcription are not applicable to this work. We reviewed all AI-assisted work. Every number reported in the paper is computed by scripts from the raw run outputs and was cross-checked against those outputs by the authors; each case study was verified against the original trajectories; AI-written code was reviewed and tested by the authors; and AI-drafted text was revised by the authors, with explanatory claims restricted to those supported by our experiments. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Data. All benchmarks are built from public, credentialed-access datasets (MIMIC-IV and MIMIC-CXR on PhysioNet) and existing benchmarks (EHR-Bench, EHRXQA, MedMod, and AgentEHR-Bench). Because the patient data cannot be redistributed, we will release the scripts that rebuild ClinSeek-Bench from these sources, together with the question identifiers of every sampled example. Sec. 3.1 describes how the benchmark is constructed, Tab. 5 lists its task categories, and Sec. G specifies the temporal policy that determines which records the model may see.

Inference and evaluation. The tool space of ClinSeekAgent is described in Sec. 2.2 and listed in Sec. F. Sec. H gives the metric and its computation, the inference settings of every model (Tab. 10), and how runs that end without an answer are handled. Statistical significance is assessed with a paired bootstrap over questions using a fixed random seed. Because of API cost, run-to-run consistency is verified on one model, GPT-5.6 Sol, with five independent runs per inference mode (Sec. C). Closed-source models are accessed through provider APIs, whose model versions may change over time; we will release the raw outputs of all reported runs so that every table and figure can be recomputed without re-querying these models.

Training. Sec. 4.1 and Sec. I describe the training data and the full fine-tuning configuration of ClinSeek-35B-A3B, which is evaluated with the same inference settings as its base model. We will release the model weights, the training trajectories, and the code for the agent, training, scoring, and analysis.

REFERENCES

Anthropic. Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, 2026a.

Anthropic. System card: Claude opus 5. https://www.anthropic.com/claude-opus-5-system-card, 2026b.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36:3867–3880, 2023.

Eric Chen, Sam Postelnik, Kameron Black, Yixing Jiang, and Jonathan H Chen. Medagentbench v2: Improving medical llm agent design. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pp. 354–371. World Scientific, 2025.

Yixiong Chen, Xinyi Bai, Yue Pan, Zongwei Zhou, and Alan Yuille. Meissa: Multi-modal medical agentic intelligence. *arXiv preprint arXiv:2603.09018*, 2026.

Shaza Elsharief, Saeed Shurrab, Baraa Al Jorf, L Julián Lechuga López, and Farah E Shamout. Medmod: Multimodal benchmark for medical prediction tasks with electronic health records and chest x-ray scans. *Proceedings of Machine Learning Research*, 287:1–23, 2025.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*, 2025a.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlthinker: Simple baselines for multimodal medical reasoning. *arXiv preprint arXiv:2508.02669*, 2025b.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi: 10.1056/AIdbp2500144. URL `https://ai.nejm.org/doi/full/10.1056/AIdbp2500144`.

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. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37: 79410–79452, 2024.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611, 2024.

Gyubok Lee, Elea Bach, Eric Yang, Tom Pollard, Alistair Johnson, Edward Choi, Jong Ha Lee, et al. Fhir-agentbench: Benchmarking llm agents for realistic interoperable ehr question answering. *arXiv preprint arXiv:2509.19319*, 2025.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, et al. Ehr-r1: A reasoning-enhanced foundational language model for electronic health record analysis. *arXiv preprint arXiv:2510.25628*, 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. *arXiv preprint arXiv:2601.13918*, 2026.

OpenAI. Gpt-5.6 system card. `https://deploymentsafety.openai.com/gpt-5-6`, 2026.

Timothy Ossowski, Sheng Zhang, Qianchu Liu, Guanghui Qin, Reuben Tan, Tristan Naumann, Junjie Hu, and Hoifung Poon. Octomed: Data recipes for state-of-the-art multimodal medical reasoning. *arXiv preprint arXiv:2511.23269*, 2025.

Lingfei Qian, Mauro Giuffre, Yan Wang, Huan He, Qianqian Xie, Xuguang Ai, Xeuqing Peng, Fan Ma, Ruey-Ling Weng, Donald Wright, et al. Ehrnavigator: A multi-agent system for patient-level clinical question answering over heterogeneous electronic health records. *arXiv preprint arXiv:2601.10020*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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.

Baorong Shi, Bo Cui, Boyuan Jiang, Deli Yu, Fang Qian, Haihua Yang, Huichao Wang, Jiale Chen, Jianfei Pan, Jieqiong Cao, et al. Medxiaohe: A comprehensive recipe for building medical mllms. *arXiv preprint arXiv:2602.12705*, 2026.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei 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.

GLM-4.7 Team. Glm-4.7: Advancing the coding capability. `https://z.ai/blog/glm-4.7`, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Kiril Vasilev, Alexandre Misrahi, Eeshaan Jain, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Michael Moor, and Charlotte Bunne. Mtbbench: A multimodal sequential clinical decision-making benchmark in oncology. *arXiv preprint arXiv:2511.20490*, 2025.

Zihan Wang, Hao Wang, Shi Feng, Xiaocui Yang, Daling Wang, Yiqun Zhang, Jinghao Lin, Haihua Yang, and Xiaozhong Ji. Deepmed: Building a medical deepresearch agent via multi-hop med-search data and turn-controlled agentic training & inference. *arXiv preprint arXiv:2601.18496*, 2026.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025a.

Juncheng Wu, Sheng Liu, Haoqin Tu, Hang Yu, Xiaoke Huang, James Zou, Cihang Xie, and Yuyin Zhou. Knowledge or reasoning? a close look at how llms think across domains. *arXiv preprint arXiv:2506.02126*, 2025b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, et al. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. Huatuogpt, towards taming language model to be a doctor, 2023. URL `https://arxiv.org/abs/2305.15075`.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pp. 4442–4457, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

TECHNICAL APPENDIX

A LIMITATIONS AND DISCUSSION

While ClinSeekAgent demonstrates promising results as both an inference-time and training-time pipeline, several limitations remain. First, the current multimodal evaluation tasks are still relatively simple in many cases. Although they involve both EHR and imaging evidence, many examples can be solved with a small number of tool calls or with limited cross-modal interaction. This does not fully stress-test the long-horizon multimodal evidence-seeking capability that ClinSeekAgent is designed for. Future benchmarks should include more challenging clinical scenarios where the agent must iteratively combine raw EHR retrieval, medical image analysis, external knowledge, and temporal reasoning over extended patient histories.

Second, our current training pipeline relies primarily on supervised fine-tuning over teacher-generated trajectories. Teacher trajectories may contain redundant or low-value tool calls, which can pollute the context window and teach the student suboptimal evidence-seeking behavior. Improving the quality of teacher trajectories through refinement, filtering, or compression is therefore an important direction for future work. In addition, post-SFT reinforcement learning could further improve the model's generalization, efficiency, and robustness by directly optimizing successful and concise clinical evidence seeking rather than merely imitating teacher behavior. We are actively working on these directions to build more challenging evaluations and more efficient training pipelines for clinical evidence-seeking agents.

B RESULTS FOR MORE MODELS

Tab. 3 and 4 report ClinSeekAgent and Curated Input for Claude Opus 4.6, GLM-5.2, and Kimi K2.6 under the same protocol as Sec. 3.2. On ClinSeek-Bench-Text all three stay within two points of Curated Input overall; on ClinSeek-Bench-MM both evaluated models improve, Claude Opus 4.6 mostly on phenotype prediction and Kimi K2.6 mostly on CXR finding presence and change comparison. Tab. 5 lists the task categories used in Fig. 3.

C MULTI-RUN CONSISTENCY

All main-text results use one run per question. To quantify run-to-run variation, we repeat both inference modes for GPT-5.6 Sol four additional times on ClinSeek-Bench-Text and ClinSeek-Bench-MM with identical settings, giving five independent runs per mode and benchmark. Tab. 6 lists

Table 3: **ClinSeekAgent vs. Curated Input on ClinSeek-Bench-Text for more models** (sample-wise F1, %). Agent = ClinSeekAgent, Curated = Curated Input, Δ = Agent – Curated on the same questions; bold marks a gain whose paired-bootstrap 95% CI excludes zero. Risk Prediction: 18 tasks, N=720; Decision Making: 27 tasks, N=1080.

Model	Risk Prediction			Decision Making			Overall		
	Agent	Curated	Δ	Agent	Curated	Δ	Agent	Curated	Δ
Claude Opus 4.6	81.2	81.2	0.0	40.1	43.0	-2.9	56.6	58.3	-1.7
GLM-5.2	79.0	75.6	**+3.4**	33.5	34.6	-1.1	51.7	51.0	+0.7
Kimi K2.6	79.6	76.4	+3.2	30.1	33.5	-3.3	49.9	50.6	-0.7

Table 4: **ClinSeekAgent vs. Curated Input on ClinSeek-Bench-MM for more models** (sample-wise F1, %; GLM-5.2 was not evaluated on the multimodal benchmark). Categories follow Tab. 5. Δ = ClinSeekAgent – Curated Input; bold marks a gain whose paired-bootstrap 95% CI excludes zero.

Model	Setting	CXR presence	CXR enumeration	CXR change	Clinical deterioration	Mortality	Phenotype	Overall
		N=215	N=182	N=222	N=125	N=125	N=120	N=989
Claude Opus 4.6	ClinSeekAgent	71.9	40.5	54.2	92.0	77.6	46.8	62.4
	Curated Input	76.6	35.3	50.1	95.2	68.8	17.2	57.2
	Δ	-4.7	**+5.2**	+4.1	-3.2	+8.8	**+29.6**	**+5.1**
Kimi K2.6	ClinSeekAgent	78.3	32.8	52.0	89.6	77.6	15.8	57.8
	Curated Input	72.3	34.0	46.3	95.2	78.4	11.9	55.7
	Δ	**+6.0**	-1.2	**+5.7**	-5.6	-0.8	**+3.9**	+2.0

the overall F1 of each run: the standard deviation across runs is at most 0.4 points for either mode, ClinSeekAgent exceeds Curated Input in all ten runs, and the paired 95% confidence interval of the run-averaged difference excludes zero on both benchmarks. Tab. 7 reports the agreement between runs at the question level with the metrics defined below.

C.1 Consistency Metrics

Let $s_{ik} \in [0,1]$ be the F1 of question i in run k, with n questions and m=5 runs per mode and benchmark.

- **Identical F1.** The share of questions whose F1 is the same in all m runs, $\frac{1}{n}\sum_i \mathbb{1}[s_{i1} = \cdots = s_{im}]$.
- **Pairwise agreement.** For each of the $\binom{m}{2}$ pairs of runs (k,l), the share of questions with $s_{ik} = s_{il}$, averaged over pairs.
- **Krippendorff's α.** Treating the runs as m raters of each question, $\alpha = 1 - D_o/D_e$ with the interval metric, where $D_o = \frac{1}{n\binom{m}{2}}\sum_i \sum_{k<l}(s_{ik} - s_{il})^2$ is the observed disagreement within questions and D_e is the same quantity over all $\binom{nm}{2}$ pairs of scores regardless of question. $\alpha = 1$ means perfect agreement and $\alpha = 0$ agreement at chance level.
- **Identical answer.** The share of questions whose final answer set (after the normalization used for scoring) is the same in all m runs. On ClinSeek-Bench-MM this is lower than the identical-F1 share because multi-label answers are free text whose wording varies between runs even when the mapped labels, and hence F1, do not.
- **Run-averaged paired CI.** Each question's score is first averaged over the m runs of a mode, $\bar{s}_i = \frac{1}{m}\sum_k s_{ik}$; the difference $\bar{s}_i^{\text{agent}} - \bar{s}_i^{\text{curated}}$ is then resampled over questions (4,000 bootstrap samples) to obtain the 95% interval reported in Tab. 6.

Table 5: **Task categories of ClinSeek-Bench.** Each ClinSeek-Bench-Text task is mapped to one category; ClinSeek-Bench-MM categories follow the source task groups.

Category	Tasks
ClinSeek-Bench-Text, risk prediction	
Mortality	Predict whether the patient dies during hospitalization (from an ED visit or an inpatient stay), or within 1, 2, 3, 7, or 14 days in the ICU
Readmission	Predict whether the patient returns to the ED within 72 hours, is readmitted to the ICU during the same hospitalization, or is readmitted to the hospital within 30 or 60 days
Length of stay	Predict whether the ICU stay exceeds 7 or 14 days, or the hospital stay exceeds 3 or 7 days
Escalation of care	Predict whether an ED patient is hospitalized, is transferred to the ICU within 12 hours, or has a critical outcome (in-hospital death or ICU transfer within 12 hours)
ClinSeek-Bench-Text, decision making	
Diagnosis	Predict the next diagnoses of a hospital admission or ED visit, as ICD titles or CCS categories
Medication	Predict the next administered medication, prescriptions (as drug names or ATC classes), ED medication reconciliation (as drug names or ATC classes), ED dispensing-cabinet (Pyxis) items, and ICU input and ingredient items
Procedure	Predict the next procedures, as ICD titles or CCS categories, and the next ICU procedure items
Laboratory & microbiology	Predict the next laboratory tests and microbiology tests
Vitals & measurements	Predict the next ICU charted items, datetime items, and output items, and the next outpatient measurements (e.g., blood pressure, weight)
Imaging report	Predict the next radiology examination
Orders	Predict the type of the next provider order
Encounter & flow	Predict the type of the next admission, the next clinical service, and the care unit of the next transfer
Next event	Predict which EHR table the next event is recorded in
ClinSeek-Bench-MM	
CXR finding presence	Answer whether a finding is present in a chest X-ray study (EHRXQA)
CXR finding enumeration	List the findings of a chest X-ray study (EHRXQA and MedMod)
CXR change comparison	Answer how findings change between two chest X-ray studies of the patient (EHRXQA)
Clinical deterioration	Predict whether an ICU patient dies within the next 24 hours (MedMod)
Mortality	Predict whether an ICU patient dies during the current ICU stay (MedMod)
Phenotype	Predict the set of acute and chronic conditions of an ICU stay (MedMod)

Table 6: **Five independent runs of GPT-5.6 Sol under both inference modes** (overall F1, %). Run 1 is the run reported in Fig. 3. The 95% CI is the paired-bootstrap interval of the run-averaged per-question difference (Sec. C.1).

	ClinSeek-Bench-Text			**ClinSeek-Bench-MM**		
	ClinSeekAgent	Curated Input	Δ	ClinSeekAgent	Curated Input	Δ
Run 1	59.3	57.8	+1.6	62.5	60.0	+2.5
Run 2	59.3	57.3	+2.0	62.9	59.5	+3.3
Run 3	59.6	57.4	+2.2	62.6	59.9	+2.7
Run 4	59.3	57.3	+2.0	61.9	59.4	+2.6
Run 5	59.5	57.3	+2.3	62.2	59.9	+2.3
Mean $\pm$ SD	59.4$\pm$0.1	57.4$\pm$0.2	+2.0$\pm$0.3	62.4$\pm$0.4	59.7$\pm$0.3	+2.7$\pm$0.4
95% CI of Δ			[0.9, 3.1]			[1.5, 4.0]

D GAIN BY RECORD LENGTH

Tab. 8 complements Fig. 4: it stratifies the ClinSeek-Bench-Text questions by the size of the patient's full pre-cutoff record (in tokens) and reports the difference between ClinSeekAgent and Full-Record Input, and between ClinSeekAgent and Curated Input, within each stratum. Records in the two

Table 7: **Question-level agreement between the five runs of GPT-5.6 Sol.** Metrics are defined in Sec. C.1.

Metric	ClinSeekAgent	Curated Input
ClinSeek-Bench-Text		
Identical F1 in all five runs (%)	80.9	84.5
Pairwise identical F1, mean over run pairs (%)	90.4	92.2
Krippendorff's α of per-question F1	0.865	0.890
Identical answer set in all five runs (%)	70.8	74.9
ClinSeek-Bench-MM		
Identical F1 in all five runs (%)	70.5	73.4
Pairwise identical F1, mean over run pairs (%)	79.6	82.9
Krippendorff's α of per-question F1	0.882	0.907
Identical answer set in all five runs (%)	29.7	41.6

Table 8: **Gain of ClinSeekAgent by record length on ClinSeek-Bench-Text.** Questions are binned by the token count of the patient's full pre-cutoff record; † marks bins whose records do not fit in the 100k-token context window of Full-Record Input (after reserving room for the model's reasoning and output) and are truncated. Cells give the F1 difference of ClinSeekAgent over the named setting; bold: paired-bootstrap 95% CI excludes zero.

Model	**Comparison**	<30k	30–83k	83–250k†	>250k†
All questions (N per bin)		467	381	432	520
Claude Opus 5	vs. Full-Record Input	0.0	+1.5	+1.6	+2.1
	vs. Curated Input	0.0	+2.0	+1.7	**+4.6**
GPT-5.6 Sol	vs. Full-Record Input	–1.8	–0.6	–0.3	+1.3
	vs. Curated Input	+0.7	–1.1	+0.4	**+5.3**
Decision-making questions (N per bin)		320	245	234	281
Claude Opus 5	vs. Full-Record Input	+0.9	+2.3	+2.6	+3.6
	vs. Curated Input	+0.3	+3.9	+3.6	**+8.5**
GPT-5.6 Sol	vs. Full-Record Input	–1.1	–1.3	–2.7	+3.4
	vs. Curated Input	+1.3	–1.3	–0.9	**+6.9**

longest strata do not fit in the 100k-token context window of Full-Record Input and are truncated oldest-first. For both models the margin of ClinSeekAgent over Full-Record Input increases with record length, from about zero below 30k tokens to +2.1 (Opus 5) and +1.3 (Sol) above 250k tokens, and the margin over Curated Input is significant in the longest stratum (+4.6 and +5.3).

E Case Study: ClinSeekAgent vs. Full-Record Input

Fig. 6 shows the case discussed in Sec. 3.3, in which the deciding row is present in the Full-Record prompt but the prompt-based model does not use it.

F ClinSeekAgent Tool Space

ClinSeekAgent provides a unified tool interface for multi-source clinical evidence seeking. The tool space contains EHR tools for patient-specific longitudinal retrieval and image tools for extracting visual evidence from medical images. Tab. 9 summarizes the tool names and their functions.

G Temporal Policy

Each question carries a prediction time t. Because the patient database contains the complete longitudinal record, we construct a question-specific view before answering each question and expose

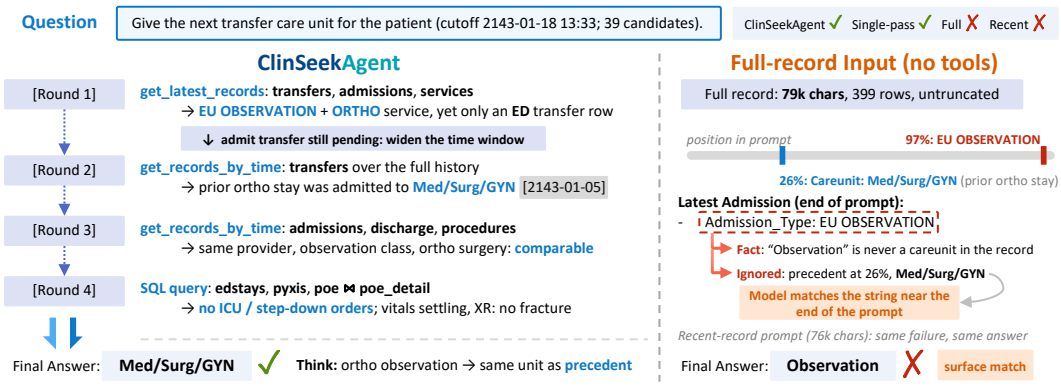

Figure 6: **Case study: ClinSeekAgent vs. Full-Record Input** (GPT-5.6 Sol, next transfer care unit). Left: the rounds of ClinSeekAgent and what each returned; right: the prompt of Full-Record Input.

only that view to the model in both inference modes. The view is constructed table by table according to the following rules.

1. **Availability time.** Each record is exposed according to the time at which its content would have become available in the EHR, rather than according to the clinical event date alone. A row is retained only if its availability time is before the prediction cutoff t. For structured tables, availability is determined by the corresponding charting or entry timestamp. For fields that are populated or updated after a row is created, values that would not yet have been available at t are withheld; retaining a row therefore does not imply that all of its fields are visible. For clinical notes, the document must have been charted and stored by t. For radiology reports, the report must have been finalized by t. Documents completed after the cutoff are excluded even if they describe an earlier encounter, procedure, or image.

2. **Diagnosis and procedure codes.** In MIMIC-IV, diagnosis and procedure codes are assigned when an encounter is completed and do not carry a separate time of entry. We therefore treat the codes of an encounter as available only 24 hours after its discharge or ED departure. For tasks that predict diagnosis or procedure codes, we additionally hide all such codes from the encounter being predicted, regardless of t.

3. **Readmission tasks.** These tasks predict whether the patient returns after the index admission, and the prediction cutoff is the discharge time of the index admission. The discharge summary is included only if it had been charted and stored by the cutoff; summaries completed after the cutoff are excluded. Other clinical notes follow the same document-availability rule described above.

4. **Multimodal questions.** Image and report availability are handled separately. A chest radiograph linked to the question is visible to the image tools only if it was acquired before the prediction cutoff t. Its associated radiology report is exposed only if the report itself had been finalized by t. Thus, availability of a chest radiograph does not imply availability of its report: reports completed after the cutoff are withheld even when the corresponding image was already available. The EHR tables follow the same temporal availability rules above.

All EHR retrieval tools, including `ehr.run_sql_query`, operate only on the same question-specific filtered view, so the model cannot retrieve rows or fields excluded by the temporal policy. The Curated Input prompts are rendered from this same filtered view, ensuring that both inference modes are subject to the same temporal availability constraints; they differ in how the available evidence is accessed and surfaced to the model.

Table 9: **ClinSeekAgent tool space.** EHR tools retrieve patient-specific evidence and image tools extract visual evidence from chest radiographs; ClinSeek-Bench-Text uses the 16 EHR tools and ClinSeek-Bench-MM all 22 tools.

Source	Tool	Function
EHR	`ehr.load_ehr`	Load the patient-specific EHR database at the prediction time; required first call.
EHR	`ehr.get_table_names`	List the available EHR tables and candidate (vocabulary) tables.
EHR	`ehr.get_column_names`	List the columns of a table.
EHR	`ehr.get_table_description`	Return the description and column information of a table.
EHR	`ehr.get_unique_values`	List the unique values of a categorical column.
EHR	`ehr.get_records_by_time`	Retrieve the records of a table within a time range.
EHR	`ehr.get_event_counts_by_time`	Count the events of every table within a time range.
EHR	`ehr.get_latest_records`	Return all records charted at the latest timestamp of a table.
EHR	`ehr.get_records_by_keyword`	Search the text columns of a table for a keyword.
EHR	`ehr.get_records_by_value`	Retrieve the records whose column exactly matches a value.
EHR	`ehr.run_sql_query`	Execute a SQL query over the patient's tables (filtering, joins, aggregation, trends).
EHR	`ehr.get_candidates_by_keyword`	Search a candidate table for a keyword.
EHR	`ehr.get_candidates_by_fuzzy_matching`	Retrieve candidate terms by fuzzy string matching.
EHR	`ehr.get_candidates_by_semantic_similarity`	Retrieve candidate terms by embedding similarity.
EHR	`ehr.think`	Record intermediate reasoning and plan the next action.
EHR	`ehr.finish`	Submit the final answer list.
Image	`image.dicom_processor`	Convert a DICOM file to PNG and extract its metadata.
Image	`image.image_visualizer`	Render an image for inspection.
Image	`image.chest_xray_classifier`	Predict probabilities for 18 chest X-ray pathologies.
Image	`image.chest_xray_report_generator`	Generate a structured chest X-ray report (findings and impression).
Image	`image.xray_phrase_grounding`	Localize a specified finding in a frontal chest X-ray.
Image	`image.chest_xray_segmentation`	Segment anatomical structures in a chest radiograph.

H Evaluation and Inference Settings

H.1 Metric

Let $\hat{Y}$ be the set of normalized labels the model returns for a question and Y the gold set. We compute $P = |\hat{Y} \cap Y|/|\hat{Y}|$, $R = |\hat{Y} \cap Y|/|Y|$, and $\mathrm{F1} = 2PR/(P+R)$ (0 when $\hat{Y}$ is empty), and average F1 over the questions of a task; task-group and overall scores average over all questions in the group or benchmark. When $|Y| = 1$ and the model returns one answer, F1 is 1 if the answer matches and 0 otherwise, i.e., accuracy; this covers all risk-prediction tasks and the single-label multimodal questions.

On ClinSeek-Bench-Text, labels are matched after whitespace and case normalization; tasks with a fixed candidate list (e.g., admission type, care unit) require the answer to be one of the candidates, and medication tasks with ATC labels are matched on the ATC name.

On ClinSeek-Bench-MM, single-label questions are judged by Claude Sonnet 4.6 for semantic agreement with the gold label (polarity for yes/no questions, value for numeric questions, and

Table 10: **Inference settings.** The same settings are used for every model and evaluation setting whenever the backend supports them.

Setting	Value
Reasoning	extended thinking, 4,096-token budget (Claude); reasoning effort `high` (GPT-5.6); native (GLM-5.2, Kimi K2.6)
Temperature	0 where accepted; not settable with reasoning on Claude and GPT-5.6
Maximum output tokens per call	8,192
Maximum agent rounds	200
Runs per question	1
Tool set	16 EHR tools (text); 16 EHR + 6 image tools (multimodal)
Maximum tool-result length	100,000 characters
Curated Input context window	100k tokens
Images per question (multimodal)	≤ 4, longest edge 1,568 pixels
Stopping criterion	`finish` tool call or round limit
Metric	mean sample-wise F1; paired-bootstrap 95% CI

name for label questions); multi-label questions have each predicted item mapped to the task's gold vocabulary by sentence-embedding cosine similarity (`all-MiniLM-L6-v2`, threshold 0.55) before set F1 is computed. Re-judging an unchanged run moves its overall score by about 0.4 points. All comparisons between settings are paired on questions: we report the mean per-question difference Δ with a paired-bootstrap 95% confidence interval (4,000 resamples) and mark Δ in bold when the interval excludes zero.

H.2 Inference Settings

All models are accessed through one OpenAI-compatible inference gateway in front of the providers' deployments; ClinSeek-35B-A3B and its base model are served locally with vLLM behind the same interface. Tab. 10 lists the settings shared across models and inference modes. Reasoning is enabled for every model: extended thinking with a 4,096-token budget for Claude models, high reasoning effort for GPT-5.6, and the native reasoning mode of GLM-5.2 and Kimi K2.6. The Claude and GPT-5.6 APIs do not accept a sampling temperature when reasoning is enabled; temperature 0 is used wherever it is accepted. Every question is run once.

ClinSeekAgent. The agent sees the task instruction, the patient identifier, and the prediction time, and interacts with 16 EHR tools (schema exploration, time-, keyword-, and value-based retrieval, SQL, candidate lookup, `think`, and `finish`); multimodal questions add six image tools (visualizer, DICOM processor, CXR classifier, report generator, phrase grounding, segmentation), for 22 tools in total. Each tool result is truncated to 100,000 characters, and a question ends when the agent calls `finish` or after 200 rounds. GLM-5.2 and Kimi K2.6 frequently exhaust the output budget inside their reasoning trace before emitting an answer, in both settings. For these two models we append one answer-only round to every such question: the conversation so far is replayed and the model is asked to call `finish` from the evidence already gathered, with no further retrieval.

Curated Input. The evidence package of the source benchmark is re-rendered from the policy-filtered record into a single prompt inside a 100k-token window. Multimodal prompts additionally attach up to four CXR images (longest edge 1,568 pixels) and the linked report text (up to 4,000 characters). The model is asked to end its reply with an `<answer>` block, which is parsed into the same answer format as a `finish` call so that both settings are scored by the same code.

I Training Settings for ClinSeek-35B-A3B

Training data. GPT-5.6 Sol (reasoning effort high) runs ClinSeekAgent with the 16 EHR tools and the strict temporal policy on the 3,000 questions of the AgentEHR-Bench training split, which cover six tasks (diagnoses, procedures, laboratory tests, microbiology tests, prescriptions, and transfers). We keep the 2,947 trajectories that end with an answer and drop those longer than 52K tokens, leaving 2,681, which we split at random into 2,628 for training and 53 for validation. The evaluation questions of Tab. 2 come from the test split and do not appear in the training data. During data generation the teacher is asked to record its reasoning with a `think` tool before each action; for training, this reasoning is moved into the student's native thinking block of the following turn, so the student reasons in its thinking channel instead of calling the tool. The loss is computed on assistant tokens only (reasoning, tool calls, and the final answer).

Configuration. Tab. 11 lists the training configuration. We train for three epochs and report the checkpoint at the end of the second epoch, which has the lowest validation loss; validation loss rises in the third epoch.

Table 11: SFT configuration for ClinSeek-35B-A3B.

Component	Configuration
Base model	Qwen3.5-35B-A3B
Teacher model	GPT-5.6 Sol
Training objective	Full-parameter supervised fine-tuning on ClinSeekAgent trajectories
Training data format	Native thinking block and tool-call format
Training / validation size	2,628 / 53 trajectories
Maximum sequence length	52,000 tokens
Epochs	3 (checkpoint after epoch 2 reported)
Global / micro batch size	32 / 1 per GPU
Optimizer	Adam (β_1=0.9, β_2=0.999); optimizer states offloaded to CPU
Learning rate	2×10^{-5}, cosine decay to 2×10^{-6}, 10 warmup steps
Weight decay	0.1
Gradient clipping	1.0
MoE auxiliary loss	0.01
MoE z-loss	0.001
Precision	bfloat16; fp32 optimizer states
Backend	verl + Megatron + mbridge
Hardware	2× H200 GPUs
Parallelism	Tensor 2, expert 2, pipeline 1, context 1
Activation recomputation	Full
Random seed	42

J Behavior Analysis of ClinSeek-35B-A3B

This section gives the quantitative results behind Sec. 4.3. We compare the 500 AgentEHR-Bench trajectories of ClinSeek-35B-A3B with those of Qwen3.5-35B-A3B; both models are served in the same way with identical inference settings. Confidence intervals are obtained by resampling questions with replacement 4,000 times. The gain in average F1 (+16.1, 95% CI [14.0, 18.3]) comes from both precision (18.6 to 32.4) and recall (31.1 to 53.3). Each AgentEHR-Bench task also has a fixed label vocabulary (e.g., CCS diagnosis categories or laboratory test names) in which answers must be expressed; the agent searches it with the candidate lookup tools of Tab. 9 or with SQL, and we call such a search a *label lookup*.

More complex SQL queries. We consider only SQL queries over the patient record (471 for the base model, 2,964 for ClinSeek-35B-A3B), not those over the label vocabulary. A query counts as *complex* if it contains a join, a GROUP BY, an aggregate function (COUNT, MIN, MAX, AVG, SUM), or a comparison on a time column; question-level rates are the fraction of the 500 questions in which the pattern appears at least once (Tab. 12). A typical pattern of ClinSeek-35B-A3B first summarizes the orders of the patient and then joins the orders of the current admission with their details:

```
-- 1) Summarize the patient's orders by type and time span
SELECT order_type, order_subtype, COUNT(*) AS n,
       MIN(ordertime), MAX(ordertime)
FROM poe
GROUP BY order_type, order_subtype
ORDER BY n DESC;

-- 2) Join the orders of the current admission with their details
SELECT p.ordertime, p.order_type, p.order_subtype,
       d.field_name, d.field_value
FROM poe p
LEFT JOIN poe_detail d ON p.poe_id = d.poe_id
WHERE p.hadm_id = 28384443
ORDER BY p.ordertime;
```

In contrast, 80% of the SQL queries of the base model select all columns of a table without further structure.

Table 12: **SQL usage on AgentEHR-Bench.** Question-level rates: fraction of questions in which the pattern appears at least once.

Metric	Base	ClinSeek-35B-A3B
SQL queries per question	0.94	5.93
Complex queries among all SQL queries	1.3%	23.9%
Questions with a join	0%	41%
Questions with GROUP BY	0%	13%
Questions with an aggregate function	0%	18%
Questions with a time filter	1%	20%
Questions with any complex query	0.8%	53.4%

Reasoning grounded on the patient record. Tab. 13 summarizes how the two models explore the record. *Tables read* is, per trajectory, the fraction of the patient's available record tables (15.8 on average) that appear in at least one data-reading call, averaged over trajectories; calls that only inspect a table's schema are not counted. *Targeted queries* is, per trajectory, the fraction of data-reading calls after the first one that look up specific values: a keyword or value lookup, or an SQL query whose WHERE clause filters on more than the patient identifier; trajectories without such later calls are excluded (463 for the base model, 472 for ClinSeek-35B-A3B remain). The conclusion is unchanged under alternative definitions: excluding SQL over the label vocabulary (12.2% vs. 35.1%), counting the first data-reading round (10.0% vs. 23.8%), or also counting time-window reads of at most two days as targeted (22.2% vs. 50.6%). *Refers to the current admission* counts the questions whose reasoning text mentions the current admission, encounter, hospitalization, stay, or episode (or "this admission", "index admission", "same-day", "today").

Table 13: **How the two models explore the patient record on AgentEHR-Bench.**

Metric	Base	ClinSeek-35B-A3B
Tables read	36%	72%
Tables read in parallel in the first data-reading round	1.0	5.2
Targeted queries	13.2%	35.8%
Data-reading calls before the first label lookup	4.4	15.7
Questions that read imaging or discharge notes	12%	74%
Questions whose reasoning refers to the current admission	39%	80%