ClinSeekAgent: Automating Multimodal Evidence Seeking for Agentic Clinical Reasoning

Juncheng Wu* **Letian Zhang*** **Yuhan Wang*** **Haoqin Tu** **Hardy Chen** **Zijun Wang** **Cihang Xie** **Yuyin Zhou**

*equal technical contribution

UC Santa Cruz

Project Page: `https://ucsc-vlaa.github.io/ClinSeekAgent/`

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 only a clinical query and access to raw data sources, ClinSeekAgent gathers evidence by querying medical knowledge bases, navigating raw EHRs, and invoking medical imaging tools; refines its hypotheses as new information emerges; and integrates the collected evidence into grounded clinical decisions. ClinSeekAgent serves both as an inference-time agent for frontier LLMs and as a training-time pipeline for distilling high-quality agent trajectories into compact open-source models. To validate its inference-time effectiveness, we construct **ClinSeek-Bench**, which pairs *Curated Input* reasoning from fixed pre-selected evidence with *Automated Evidence-Seeking* over raw clinical data. On text-only EHR tasks, ClinSeekAgent improves Claude Opus 4.6 from 60.0 to 63.2 overall F1 and MiniMax M2.5 from 43.1 to 47.3, with positive risk-prediction gains in 7 out of 9 evaluated host models. On multimodal tasks, ClinSeekAgent improves Claude Opus 4.6 from 47.5 to 62.6 (**+15.1**); all evaluated models improve across the three CXR-related task groups. We further validate ClinSeekAgent as a training pipeline by distilling agentic evidence-seeking trajectories into **ClinSeek-35B-A3B**, which achieves 34.0 average F1 on existing AgentEHR-Bench, improving over its Qwen3.5-35B-A3B baseline by **+11.9** points and approaching Claude Opus 4.6. We will fully 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 [1, 2, 3, 4, 5, 6]. However, many existing medical-agent settings remain overly simplistic, deviating from real-world clinical workflows. They often rely on general medical knowledge [7] or short organized patient vignettes, whereas real-world clinical decision support requires actively seeking evidence from various sources: general medical knowledge from external references [8], patient-specific longitudinal information from raw Electronic Health Record (EHR) tables [9, 10], and visual clues from medical imaging [11]. Such a limitation is particularly salient for clinical decision support, where the key challenge is not

Preprint.

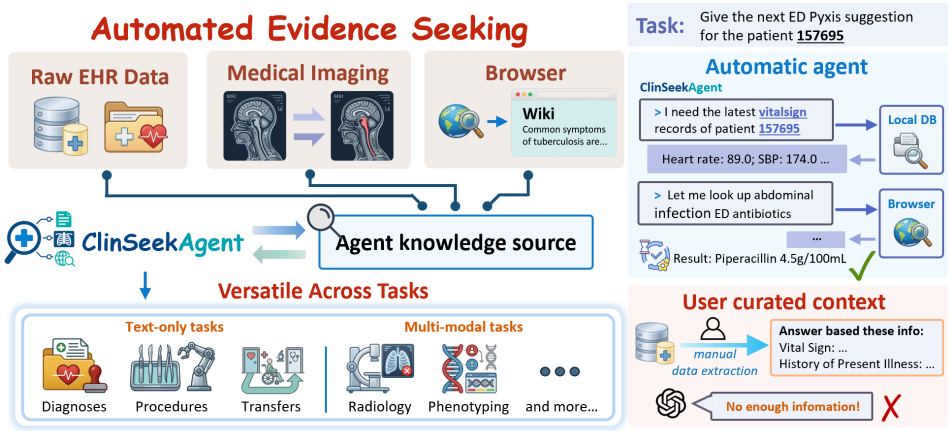


Figure 1: **ClinSeekAgent Overview.** ClinSeekAgent is an automated agentic evidence-seeking pipeline. It interacts with heterogeneous data sources to enable multimodal evidence seeking for clinical decision support. Compared with prior user-curated context settings, ClinSeekAgent is more flexible by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

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 [12, 13, 14, 15]. For example, recent EHR reasoning pipelines convert structured tables into textual contexts, retrieve task-related entities, and synthesize reasoning data from pre-extracted patient information [12, 16]. Multimodal clinical benchmarks also combine EHRs and medical images to support realistic prediction and question-answering tasks [13, 14]. 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 [17, 18, 19, 20, 21, 22], 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. 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)** web search, **(2)** raw EHR retrieval, and **(3)** medical imaging tools, iteratively refining its actions as new evidence emerges. This enables the agent to recover patient-specific, multimodal, and external medical signals that fixed curated contexts may miss. For example, when asked to provide the next ED Pyxis suggestion, ClinSeekAgent retrieves recent vital signs from the local EHR database, searches for relevant antibiotics for abdominal infection in the ED, and integrates these signals to correctly predict *piperacillin*, while the same model under the curated-context setting fails due to missing critical evidence.

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. For each sample, the source benchmark [12, 14, 13] 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 text-only EHR tasks derived from EHR-Bench [12], which covers **45** decision-making and risk-prediction tasks, and **6** multimodal task groups adapted from EHRXQA [13] and MedMod [14] (see Sec. 3).

Our inference-time experiments show that ClinSeekAgent can improve over fixed curated inputs when paired with capable agentic models. On text-only EHR tasks, Claude Opus 4.6 improves from 60.0 overall F1 under *Curated Input* to 63.2 under *Automated Evidence-Seeking*, and MiniMax M2.5 improves from 43.1 to 47.3 (Tab. 1). The gains are especially pronounced in risk prediction and multimodal clinical tasks, where relevant evidence is often sparse, longitudinal, or distributed across EHR tables and medical images. On the multimodal benchmark, ClinSeekAgent improves 5 out of 6 evaluated models, with Claude Opus 4.6 improving from 47.5 to 62.6 overall F1 (Tab. 2), suggesting that active evidence acquisition can recover clinical signals that fixed curated contexts may miss.

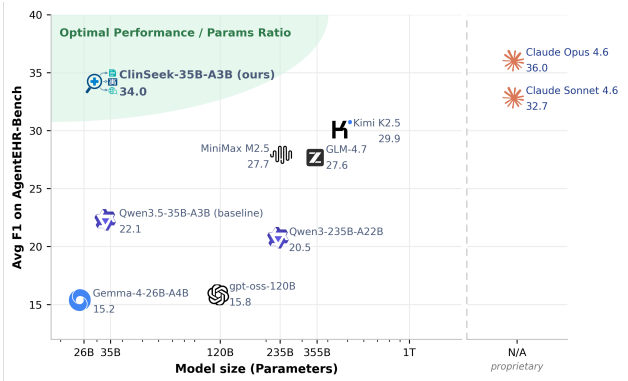


Figure 2: **Performance–model size comparison on AgentEHR-Bench.** ClinSeek-35B-A3B achieves strong performance among open-source models while maintaining a favorable parameter-efficiency tradeoff.

While these inference-time results demonstrate the effectiveness of ClinSeekAgent, they also suggest that automated evidence seeking depends on the agentic model's ability to plan and execute long-horizon tool use. Therefore, we further validate ClinSeekAgent as a training pipeline for open-source clinical agents. Using ClinSeekAgent, we collect high-quality clinical search trajectories from a strong teacher model and fine-tune Qwen3.5-35B-A3B [23], resulting in **ClinSeek-35B-A3B**. On the existing AgentEHR-Bench [17], ClinSeek-35B-A3B improves over its base model from 22.1 to 34.0 average F1, outperforming all evaluated open-source baselines and approaching Claude Opus 4.6 at 36.0 (Fig. 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.

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 20 tools across three complementary evidence sources: *EHR retrieval*, *web search*, and *medical image analysis*. Specifically, it provides 11 EHR tools for accessing patient-specific longitudinal records, including schema inspection, temporal

retrieval, SQL-based querying, and candidate-term grounding; 3 browser tools for acquiring external medical knowledge through web search; and 6 image tools for extracting visual evidence through DICOM preprocessing, chest X-ray classification, report generation, phrase grounding, and anatomical segmentation. The complete tool list are provided in Appendix C.

2.3 Agentic Evidence-Seeking Trajectories

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

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

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, web search, 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 benchmark, 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 includes both text-only and multimodal clinical tasks. For *text-only evaluation*, we use EHR-Bench from EHR-R1 [12], which contains 45 EHR analysis subtasks covering decision-making and risk-prediction scenarios. We randomly sample 40 examples from each subtask, resulting in 1,800 text-only examples. For *multimodal evaluation*, we adapt EHRXQA [13] and MedMod [14], both built on MIMIC-IV EHRs and 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 six task groups: CXR finding presence, CXR finding enumeration, CXR temporal change comparison, 24-hour decompensation prediction, in-hospital mortality prediction, and phenotype prediction.

Curated Input Data Collection. We preserve the *original 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 rather than only the curated 24-hour window. For multimodal tasks, we preserve the original patient-level task, label, and valid EHR-CXR linkage, but require the agent to retrieve EHR evidence and invoke imaging tools when needed. Across all tasks, we hide any information after the prediction cutoff to prevent temporal leakage.

3.2 Evaluation Setting

We evaluate ClinSeekAgent under the *Automated Evidence-Seeking* setting and compare it with the paired *Curated Input* setting defined in Sec. 3.1. We evaluate 12 strong proprietary and publicly available models, including Claude Opus 4.6 [24], Claude Sonnet 4.6 [25], GLM-4.7 [26], Qwen3.5-35B-A3B [23], Gemma-4-26B-A4B-it [27], MiniMax M2.5 [28], Kimi K2.5 [29], Qwen3-VL-235B [30], gpt-oss-120B [31], MedGemma-27B-it [32], EHR-R1-8B, and EHR-R1-72B [12]. Domain-specialized reasoning models such as EHR-R1 and MedGemma are evaluated only under *Curated Input*, while models without sufficient multimodal capability are excluded from multimodal tasks when appropriate. We report sample-wise F1(%) as the primary metric: F1 is computed for each example and then averaged within each task group, with the overall score averaged over the full benchmark. More inference details are provided in Appendix D.

3.3 Main Results: ClinSeekAgent Improves State-of-the-Art Agentic Models

We evaluate the ClinSeekAgent framework and the Curated Input baseline on the collected benchmarks, and report the performance of both methods as well as their differences in Tab. 1 and Tab. 2.

ClinSeekAgent improves text-only EHR tasks when paired with strong agentic models. As shown in Tab. 1, the strongest agentic models achieve better overall performance with the ClinSeekAgent pipeline than with the Curated Input baseline. Claude Opus 4.6 improves from 60.0 to 63.2, yielding a +3.2-point gain, while MiniMax M2.5 improves from 43.1 to 47.3, corresponding to a +4.2-point gain. These results suggest that when a model has sufficient tool-use and planning ability, ClinSeekAgent can effectively leverage patient-level retrieval to improve clinical prediction performance. On the other hand, weaker models show less pronounced or unstable gains from the pipeline. For example, Claude Sonnet 4.6 achieves only a near tie, with a modest **+0.9**-point improvement overall. Other models, including Qwen3.5-35B-A3B(+0.2), Kimi K2.5(-11.3), Qwen3-VL-235B(-9.8), etc., either perform comparably to or underperform the Curated Input baseline in the overall results.

ClinSeekAgent brings broader gains on multimodal tasks, with larger improvements for stronger agents. The advantage of ClinSeekAgent becomes more consistent in the multimodal benchmark. As reported in Tab. 2, ClinSeekAgent improves the overall performance of five out of the six evaluated models. The largest gains are observed for the strongest agentic models: Claude Opus 4.6 improves by **+15.1** points, and Claude Sonnet 4.6 improves by **+6.9** points. Strong open-source multimodal models also benefit from the pipeline, with Qwen3-VL-235B improving by **+5.9** points and Gemma-4-26B-A4B-it improving by **+6.6** points, even though neither model benefits from ClinSeekAgent on text-only EHR tasks. These results suggest that agentic access to patient information is especially valuable when clinical decisions require jointly integrating EHR context and multimodal evidence, where fixed curated inputs are less likely to cover all task-relevant information.

3.4 Advantage Analysis of ClinSeekAgent

We further analyze the advantages of ClinSeekAgent on both text-only and multimodal benchmarks.

Text-only: ClinSeekAgent shows substantial advantage on risk prediction. In Fig. 3, we show how much ClinSeekAgent pipeline wins over Curated Input baseline on text-only tasks. The heatmap shows that the advantage of ClinSeekAgent is concentrated in the risk-prediction group: 7 out of 9 evaluated models achieve a positive average gain on risk prediction when using ClinSeekAgent. At the subtask level, the improvements are particularly pronounced on long-horizon hospital-event prediction tasks. For Claude Opus 4.6, ClinSeekAgent substantially improves three tasks: *Mortality Hospital* by **+12.5** points, *LengthOfStay* by **+16.2** points, and *ED Hospitalization* by **+12.5** points. Similar patterns are observed for other strong and mid-sized models. Claude Sonnet 4.6 improves by **+30.0** points on *ED Hospitalization* and **+17.5** points on *LengthOfStay*.

This advantage is consistent with the nature of risk prediction tasks. Risk-prediction questions depend on sparse but decisive evidence distributed across the patient record, which is the primary advantage of our pipeline. ClinSeekAgent allows the agent to actively search for these signals and integrate them into the prediction. In contrast, a fixed Curated Input baseline cannot enumerate all such task-relevant signals in advance, especially when the relevant evidence varies across patients and subtasks.

Table 1: **Comparison between ClinSeekAgent and Curated Input baseline on text-based EHR tasks.** The strongest models achieve improvements over the baseline under the ClinSeekAgent framework, including Opus 4.6, Sonnet 4.6, and MiniMax M2.5, which we attribute to their strong agentic capabilities. The gains brought by our framework are most pronounced on risk-prediction tasks.

Model	Risk Prediction			Decision Making			Overall		
	ClinSeek	Curated Input	Δ	ClinSeek	Curated Input	Δ	ClinSeek	Curated Input	Δ
Closed-source models									
Claude Opus 4.6	90.7	81.0	**+9.7**	44.8	45.9	**-1.1**	63.2	60.0	**+3.2**
Claude Sonnet 4.6	90.0	77.5	**+12.5**	35.9	42.6	**-6.7**	57.5	56.6	**+0.9**
Open-source models									
EHR-R1-72B	–	67.1	–	–	45.2	–	–	53.9	–
GLM-4.7	75.1	70.4	**+4.7**	23.1	38.6	**-15.5**	43.9	51.3	**-7.4**
Qwen3.5-35B-A3B	84.4	73.6	**+10.8**	22.0	29.0	**-7.0**	47.0	46.8	**+0.1**
Gemma-4-26B-A4B-it	83.5	78.6	**+4.9**	17.3	27.8	**-10.5**	43.8	48.1	**-4.3**
MiniMax M2.5	86.7	68.4	**+18.3**	21.0	26.3	**-5.3**	47.3	43.1	**+4.2**
Kimi K2.5	65.0	79.9	**-14.9**	19.8	28.8	**-9.0**	37.9	49.2	**-11.3**
Qwen3-VL-235B	67.9	71.0	**-3.1**	19.1	33.4	**-14.3**	38.6	48.4	**-9.8**
gpt-oss-120b	75.4	74.0	**+1.4**	16.6	22.3	**-5.7**	40.1	43.0	**-2.9**
MedGemma-27B-it	–	65.0	–	–	25.2	–	–	41.1	–
EHR-R1-8B	–	64.0	–	–	23.4	–	–	39.7	–

Table 2: **Comparison between ClinSeekAgent and Curated Input baseline on multimodal EHR tasks.** We evaluate models with multimodal capabilities and find that our pipeline brings consistent improvements across most task groups and model families.

Model	Method	CXR: finding presence	CXR: finding enumeration	CXR: change comparison	Mortality (24 h)	Inpatient mortality	Phenotype (CCS groups)	Multimodal overall
Claude Opus 4.6	ClinSeekAgent	78.3	43.6	54.8	92.0	74.4	45.5	**62.6**
	Curated Input	55.2	31.6	38.0	93.6	69.6	11.5	**47.5**
	Δ	**+23.2**	**+12.0**	**+16.8**	**-1.6**	**+4.8**	**+34.0**	**+15.1**
Claude Sonnet 4.6	ClinSeekAgent	79.5	41.3	51.5	64.0	68.8	26.1	**54.9**
	Curated Input	64.8	29.7	34.7	90.4	70.4	13.8	**48.0**
	Δ	**+14.7**	**+11.6**	**+16.8**	**-26.4**	**-1.6**	**+12.3**	**+6.9**
Qwen3.5-35B-A3B	ClinSeekAgent	73.8	34.2	44.4	91.2	74.4	0.3	**51.7**
	Curated Input	59.1	34.1	30.7	90.4	81.6	0.5	**46.9**
	Δ	**+14.7**	**+0.2**	**+13.7**	**+0.8**	**-7.2**	**-0.2**	**+4.8**
Kimi K2.5	ClinSeekAgent	61.4	34.9	43.8	71.2	62.4	12.3	**46.9**
	Curated Input	56.3	24.7	35.0	91.2	87.2	12.4	**47.5**
	Δ	**+5.1**	**+10.2**	**+8.8**	**-20.0**	**-24.8**	**-0.1**	**-0.6**
Qwen3-VL-235B	ClinSeekAgent	70.4	35.7	47.8	79.2	61.6	6.0	**49.8**
	Curated Input	60.3	21.1	32.8	87.2	72.8	6.6	**43.9**
	Δ	**+10.1**	**+14.6**	**+15.0**	**-8.0**	**-11.2**	**-0.6**	**+5.9**
Gemma-4-26B-A4B-it	ClinSeekAgent	78.9	21.6	38.4	65.6	71.2	0.4	**44.9**
	Curated Input	56.9	21.4	25.4	79.2	60.0	0.0	**38.2**
	Δ	**+22.0**	**+0.2**	**+13.0**	**-13.6**	**+11.2**	**+0.4**	**+6.7**

Multimodal: compositional tool use bridges visual, EHR, and external evidence. Among the multimodal tasks in Tab. 2, the gains are most pronounced on CXR-related benchmarks, where ClinSeekAgent consistently improves performance over the Curated Input baseline across all evaluated models, including mid-sized models such as Qwen3.5-35B-A3B and Gemma-4-26B-A4B-it. On the Phenotype task, Claude Opus 4.6 also obtains a remarkable **+34.0**-point improvement.

These gains come from the compositional tool use enabled by ClinSeekAgent. Compared with the Curated Input baseline, ClinSeekAgent can combine three complementary sources of evidence: **(a)** CXR classifier outputs with per-finding probabilities, providing structured visual evidence beyond the model's native image understanding. **(b)** SQL queries over ICU events for patient-specific temporal signals; and **(c)** browser search for task-specific medical definitions, such as the 25-phenotype Harutyunyan-2019 taxonomy. Together, these tools ground multimodal reasoning in image findings, structured EHR evidence, and benchmark-relevant clinical knowledge, explaining the remarkable improvements. In Fig. 4, we provide a concrete case comparison with the Curated Input baseline. Under the ClinSeekAgent framework, the model invokes a medical imaging expert to obtain professional CXR analysis and diagnosis, extracts sparse information over a long time span from raw EHR data, and uses the browser tool to acquire external knowledge. ClinSeekAgent achieves an F1 = 83.3 by comprehensively leveraging these tools. In contrast, the Curated Input setting fails

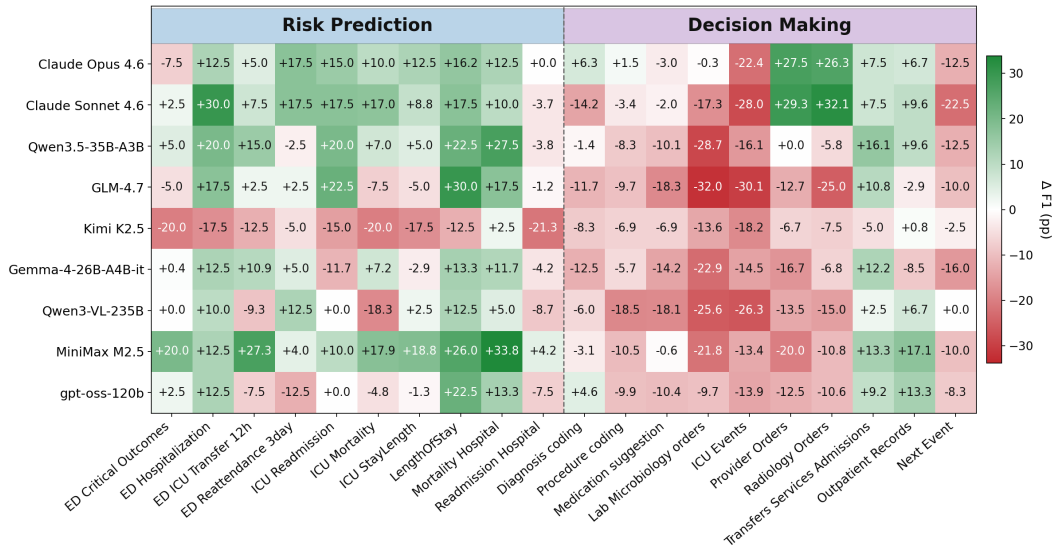


Figure 3: **Visualization of fine-grained text-based subtasks.** We categorize the tasks in EHR-Bench into fine-grained groups and report the performance gains brought by ClinSeekAgent pipelines. Green indicates an advantage over Curated Input baseline, while red indicates a disadvantage.

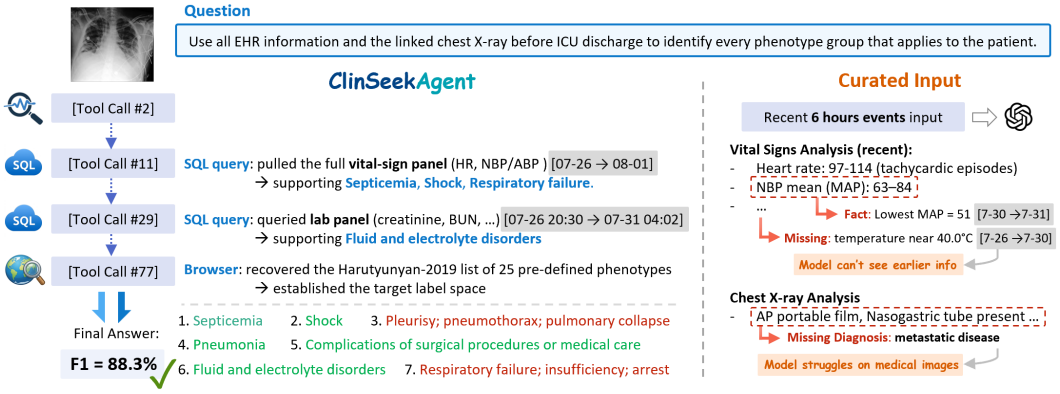


Figure 4: **Comparison between the ClinSeekAgent pipeline and the Curated Input baseline.**

to provide the correct answer due to the limited patient context and insufficient ability to analyze medical images.

3.5 Failure Analysis on Decision-Making Task

As shown in Fig. 3, the main weakness of ClinSeekAgent appears in the decision-making task group. Unlike risk prediction, where most models obtain positive gains, decision-making subtasks show less consistent improvements and often degrade under the ClinSeek pipeline. In Tab. 1, Qwen3.5-35B-A3B with ClinSeekAgent substantially outperforms the domain-tuned EHR-R1-72B reasoning-only model on risk prediction (84.4 vs. 67.1, **+17.3** points), but trails the domain expert by 23.2 points (22.0 vs. 45.2). This contrast shows that the paradigm gap is task-family-specific: ClinSeek-style retrieval is highly effective for risk prediction, but sometimes fails to find the critical information for decision making. In Sec. F.1, we provide a concrete example where our pipeline collects excessive irrelevant information but overlooks the key signals leading to the correct answer. In contrast, the Curated Input baseline identifies similar patterns in the historical context and makes the correct judgment.

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

We next validate ClinSeekAgent as a training pipeline for open-source EHR agents. While the previous section evaluates ClinSeekAgent as an inference-time evidence-seeking workflow, here we ask whether the same pipeline can generate supervision for transferring long-horizon clinical

Model	Diagnoses	Labs	Microbiology	Procedures	Transfers	Avg.
Closed-source models						
Claude Opus 4.6	**58.5**	**42.1**	**27.2**	**31.1**	20.9	**36.0**
Claude Sonnet 4.6	54.4	35.6	23.4	26.3	**23.7**	32.7
Open-source models						
Kimi K2.5	46.9	33.7	18.9	27.9	22.1	29.9
MiniMax-M2.5	51.5	29.0	19.0	22.0	17.0	27.7
GLM-4.7	46.4	28.6	16.6	23.7	**22.9**	27.6
Qwen3-235B-A22B	30.6	20.3	17.3	24.9	9.6	20.5
Tongyi DeepResearch 30B-A3B	25.8	14.9	8.8	17.9	13.2	16.1
gpt-oss-120b	27.3	12.8	12.4	19.1	7.6	15.8
Gemma-4-26B-A4B-it	17.9	18.5	19.7	11.2	8.8	15.2
OpenSeeker-30B	20.4	4.5	12.8	14.2	10.6	12.5
Qwen3.5-35B-A3B (base)	36.6	17.7	16.2	21.9	18.1	22.1
ClinSeek-35B-A3B (ours, SFT)	**55.4**	**38.5**	**27.6**	**31.7**	16.7	**34.0**
Ours − base	+18.8	+20.8	+11.4	+9.8	-1.4	**+11.9**
Ours − teacher	-3.1	-3.6	+0.4	+0.6	-4.2	-2.0

Table 3: **AgentEHR Benchmark five-task evaluation.** We report F1 scores (%). The best performer in each group is highlighted in bold.

search behavior to a smaller model. This experiment tests whether the student can learn not only final-answer prediction, but also the evidence-seeking process induced by ClinSeekAgent.

4.1 Experimental Settings

We use Claude Opus 4.6 as the teacher model to generate ClinSeekAgent trajectories from the training split of our text-based benchmark, and fine-tune Qwen3.5-35B-A3B with supervised fine-tuning. The training data are rendered in the native tool-call format with a maximum sequence length of 52K tokens. Full training details are provided in Appendix E.

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

Table 3 reports the AgentEHR-Bench five-task evaluation results. ClinSeekAgent trajectory distillation improves the same Qwen3.5-35B-A3B base model from 22.1 to 34.0 average F1, yielding a **+11.9**-point gain. The improvement is especially strong on *Diagnoses* (**+18.8**), *Laboratory Events* (**+20.8**), *Microbiology Events* (**+11.4**), and *Procedures* (**+9.8**), with *Transfers* as the only task showing a slight drop (**-1.4**). The distilled model achieves the strongest open-source performance in our evaluation. ClinSeek-35B-A3B reaches 34.0 average F1, outperforming Kimi K2.5 by **+4.1** points, MiniMax-M2.5 by **+6.3**, and GLM-4.7 by **+6.4**. It also closes most of the gap to Claude Opus 4.6, reaching 94.4% of the teacher's performance (34.0 vs. 36.0) and surpassing Claude Sonnet 4.6 by **+1.3**. These results show that ClinSeekAgent-generated trajectories can transfer long-horizon EHR agentic capability into a smaller open-source model.

4.3 What Does the Student Learn?

We further analyze the tool-use behavior of ClinSeek-35B-A3B to understand what is learned beyond final-answer imitation. As shown in Figure 5, the distilled model does not substantially shorten the search process: the base model makes 33,043 tool calls on the same 500 AgentEHR-Bench questions, while ClinSeek-35B-A3B makes 31,446 calls. Instead, the main change is how the model allocates its tool budget. ClinSeek-35B-A3B learns a more diverse and flexible EHR retrieval policy. Most notably, its use of the free-form SQL tool `ehr.run_sql_query` increases from 649 to 3,932 calls, corresponding to a share increase from 2.0% to 12.5%. This shift suggests that ClinSeekAgent trajectories teach the student to treat the EHR as a programmable database, rather than relying only on fixed retrieval templates. Together with the stronger AgentEHR-Bench performance in Table 3, this indicates that ClinSeekAgent distillation transfers procedural evidence-seeking behavior, not merely final-answer patterns.

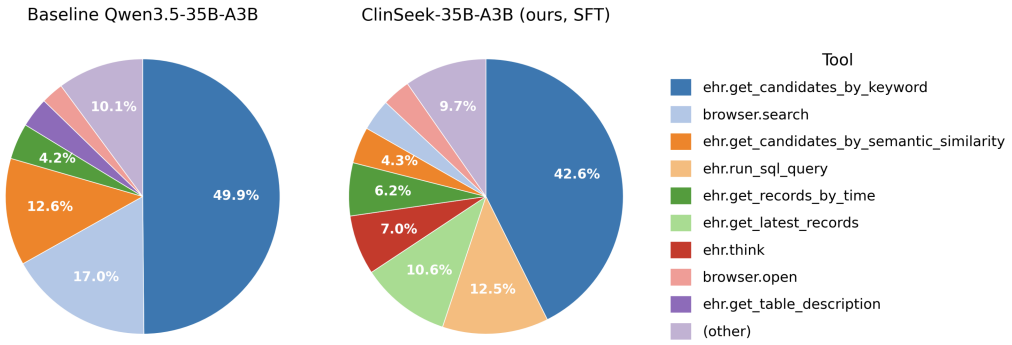


Figure 5: **Tool-call distribution before and after SFT training.**

5 Related Work

Medical Reasoning with Curated Evidence. Recent medical LLMs have shown strong performance in medical question answering and diagnostic reasoning [33, 34, 35, 36, 37, 38, 39], 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 [40, 41], summarized clinical notes [16], or task-specific patient contexts [42, 43]. Recent EHR and multimodal clinical benchmarks move closer to real clinical data by grounding tasks in structured patient records, radiology reports, and medical images [12, 14, 13]. 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, medical images, and external knowledge sources.

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 [2], while DeepMed [39] and Meissa [44] train medical agents to perform multi-step evidence search or interaction for medical reasoning [39, 44]. Closer to EHR-based decision support, AgentEHR [17], MedAgentBench [18], and FHIR-AgentBench [21] 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 [5]. These works demonstrate the promise of agentic clinical AI, but their evidence-seeking processes are typically limited to either 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, medical image analysis tools, and external knowledge sources, 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 allows an agentic model to proactively gather, refine, and synthesize evidence from diverse sources rather than merely relying on user-curated inputs. To evaluate ClinSeekAgent as an inference-time pipeline, we reformulate text-only and multimodal clinical tasks into an agentic setting and show that ClinSeekAgent improves strong agentic models, especially when evidence is longitudinal, sparse, or distributed across modalities. To evaluate ClinSeekAgent as a training pipeline, we distill long-horizon evidence-seeking trajectories into an open-source student model, achieving open-source state-of-the-art performance on AgentEHR-Bench while improving tool-use behavior. 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.

References

[1] 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*, 2025.

[2] 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.

[3] 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.

[4] 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.

[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.

[6] 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.

[7] 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*, 2025.

[8] 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*, pages 4442–4457, 2025.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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*, pages 354–371. World Scientific, 2025.

[20] 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.

[21] 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.

[22] 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*, pages 22315–22339, 2024.

[23] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

[24] Anthropic. Claude opus 4.6. `https://www.anthropic.com/news/claude-opus-4-6`, 2026.

[25] Anthropic. Claude sonnet 4.6. `https://www.anthropic.com/news/claude-sonnet-4-6`, 2026.

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

[27] Google DeepMind. Welcome gemma 4: Frontier multimodal intelligence on device. `https://huggingface.co/blog/gemma4`, 2026.

[28] MiniMax. Forge: Scalable agent rl framework and algorithm. `https://huggingface.co/blog/MiniMax-AI/forge-scalable-agent-rl-framework-and-algorithm`, 2026.

[29] 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.

[30] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[31] Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

[32] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[33] 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.

[34] 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.

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

[36] 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*, 2025.

[37] 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.

[38] 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.

[39] 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.

[40] 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)*, pages 2567–2577, 2019.

[41] 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.

[42] 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.

[43] 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.

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

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. However, we observe that trajectories produced by the teacher model, Claude Opus 4.6, are not always tool-efficient. Some trajectories 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 Uncertainty Estimation

We report uncertainty estimates for all F1-acc results using per-sample scores. For each model, task, and evaluation setting, we compute the mean per-sample F1-acc and report a two-sided 95% Student-t confidence interval over evaluation samples. All values are reported in percentage points as mean $\pm$ CI radius. The sample size N denotes the number of evaluated questions in each cell. For pooled results, the "Overall" row in the text-only EHR table pools all text-only samples, the "Overall" column in the multimodal table pools all multimodal task groups, and the AgentEHR "Avg." column pools the five evaluated subtasks. These confidence intervals quantify uncertainty of the estimated mean F1 over evaluation samples, but they are not paired significance tests between methods.

Tab. 4 reports uncertainty estimates for the text-only EHR tasks. The overall estimates are relatively stable because they pool $N = 1800$ samples, with CI radii around two points. The results remain consistent with our main finding: ClinSeekAgent improves strong agentic models such as Claude Opus 4.6 (60.0 ± 2.11 to 63.2 ± 2.09) and MiniMax M2.5 (43.1 ± 2.17 to 47.3 ± 2.24), while gains are more task- and model-dependent for weaker agents.

Tab. 5 reports confidence intervals for multimodal tasks. The pooled overall results use $N = 989$ samples, with CI radii around three points. ClinSeekAgent improves five of six evaluated models overall, including Claude Opus 4.6 (47.5 ± 2.89 to 62.6 ± 2.65), Claude Sonnet 4.6 (48.0 ± 2.88 to 54.9 ± 2.79), and Qwen3-VL-235B (43.9 ± 2.95 to 49.8 ± 2.91). This supports our conclusion that agentic evidence seeking is especially useful when information is distributed across EHR and imaging sources.

Tab. 6 reports confidence intervals for the AgentEHR five-task evaluation. ClinSeek-35B-A3B improves over the Qwen3.5-35B-A3B base model from 22.1 ± 2.00 to 34.0 ± 1.98 over $N = 500$ samples, exceeds the strongest evaluated open-source peer Kimi K2.5 (29.9 ± 1.93), and approaches the Claude Opus 4.6 teacher (36.0 ± 2.05). These results further support ClinSeekAgent as an effective training pipeline for open-source EHR agents.

C 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, browser tools for external medical knowledge search, and image tools for extracting visual evidence from medical images. Tab. 7 summarizes the tool names and their functions.

Table 4: **Confidence intervals for text-based EHR tasks.** Each cell reports mean F1-acc in percentage points with the 95% CI radius computed over per-sample scores. Delta columns are omitted for compactness.

Model	Task Group	N	ClinSeek	Curated Input
Claude Opus 4.6	Risk Prediction	720	90.7 ± 2.13	81.0 ± 2.87
Claude Opus 4.6	Decision Making	1080	44.8 ± 2.67	45.9 ± 2.64
Claude Opus 4.6	Overall	1800	63.2 ± 2.09	60.0 ± 2.11
Claude Sonnet 4.6	Risk Prediction	720	90.0 ± 2.20	77.5 ± 3.06
Claude Sonnet 4.6	Decision Making	1080	35.9 ± 2.58	42.6 ± 2.63
Claude Sonnet 4.6	Overall	1800	57.5 ± 2.16	56.6 ± 2.15
GLM-4.7	Risk Prediction	720	75.1 ± 3.16	70.4 ± 3.34
GLM-4.7	Decision Making	1080	23.1 ± 2.32	38.6 ± 2.57
GLM-4.7	Overall	1800	43.9 ± 2.22	51.3 ± 2.16
Qwen3.5-35B-A3B	Risk Prediction	720	84.4 ± 2.65	73.6 ± 3.23
Qwen3.5-35B-A3B	Decision Making	1080	22.0 ± 2.29	29.0 ± 2.44
Qwen3.5-35B-A3B	Overall	1800	47.0 ± 2.24	46.8 ± 2.20
Gemma-4-26B-A4B-it	Risk Prediction	720	83.5 ± 2.72	78.6 ± 2.80
Gemma-4-26B-A4B-it	Decision Making	1080	17.3 ± 2.12	27.8 ± 1.97
Gemma-4-26B-A4B-it	Overall	1800	43.8 ± 2.25	48.1 ± 1.99
MiniMax M2.5	Risk Prediction	720	86.7 ± 2.49	68.4 ± 3.30
MiniMax M2.5	Decision Making	1080	21.0 ± 2.25	26.3 ± 2.40
MiniMax M2.5	Overall	1800	47.3 ± 2.24	43.1 ± 2.17
Kimi K2.5	Risk Prediction	720	65.0 ± 3.49	79.9 ± 2.94
Kimi K2.5	Decision Making	1080	19.8 ± 2.19	28.8 ± 2.42
Kimi K2.5	Overall	1800	37.9 ± 2.17	49.2 ± 2.20
Qwen3-VL-235B	Risk Prediction	720	67.9 ± 3.41	71.0 ± 3.32
Qwen3-VL-235B	Decision Making	1080	19.1 ± 2.17	33.4 ± 2.49
Qwen3-VL-235B	Overall	1800	38.6 ± 2.18	48.4 ± 2.17
gpt-oss-120b	Risk Prediction	720	75.4 ± 3.15	74.0 ± 3.19
gpt-oss-120b	Decision Making	1080	16.6 ± 2.05	22.3 ± 2.22
gpt-oss-120b	Overall	1800	40.1 ± 2.21	43.0 ± 2.18

D Evaluation and Inference Settings

We use sample-wise F1 as the primary metric. For each example, we compute F1 between the normalized prediction and the ground-truth answer, and then average scores within each task group; overall scores are averaged over all evaluated examples. All models are evaluated with one run per question. For agentic evaluation, the agent interacts with the available tools until it calls the `finish` tool or reaches the maximum interaction budget. Closed-source models are evaluated through AWS Bedrock or provider APIs, while open-source models are served with vLLM using an OpenAI-compatible API. For multimodal evaluation, CXR images are resized so that the longest edge is at most 1568 pixels, and image-tool outputs are returned through the same tool-calling interface as EHR and web-search results. See Tab. 8 for detailed settings.

E Training Settings for ClinSeek-35B-A3B

Table 9 summarizes the training configuration used for ClinSeek-35B-A3B.

Table 5: **Confidence intervals for multimodal EHR tasks.** Each cell reports mean F1-acc in percentage points with the 95% CI radius; task-specific sample sizes are shown in the column headers.

Model	**Method**	CXR finding presence ($N = 177$)	CXR finding enumeration ($N = 220$)	CXR change comparison ($N = 222$)	Mortality 24 h ($N = 125$)	Inpatient mortality ($N = 125$)	Phenotype CCS ($N = 120$)	Overall ($N = 989$)
Claude Opus 4.6	ClinSeek	78.3 ± 6.10	43.6 ± 5.03	54.8 ± 6.26	92.0 ± 4.82	74.4 ± 7.76	45.5 ± 3.50	62.6 ± 2.65
Claude Opus 4.6	Curated Input	55.2 ± 7.38	31.6 ± 4.74	38.0 ± 6.12	93.6 ± 4.35	69.6 ± 8.18	11.5 ± 2.48	47.5 ± 2.89
Claude Sonnet 4.6	ClinSeek	79.5 ± 5.99	41.3 ± 4.90	51.5 ± 6.35	64.0 ± 8.53	68.8 ± 8.24	26.1 ± 3.59	54.9 ± 2.79
Claude Sonnet 4.6	Curated Input	64.8 ± 7.09	29.7 ± 4.61	34.7 ± 6.03	90.4 ± 5.24	70.4 ± 8.11	13.8 ± 2.49	48.0 ± 2.88
Qwen3.5-35B-A3B	ClinSeek	73.8 ± 6.52	34.2 ± 5.07	44.4 ± 6.50	91.2 ± 5.04	74.4 ± 7.76	0.3 ± 0.55	51.7 ± 2.99
Qwen3.5-35B-A3B	Curated Input	59.1 ± 7.29	34.1 ± 4.78	30.7 ± 5.85	90.4 ± 5.24	81.6 ± 6.89	0.5 ± 0.46	46.9 ± 2.95
Kimi K2.5	ClinSeek	61.4 ± 7.22	34.9 ± 4.91	43.8 ± 6.30	71.2 ± 8.05	62.4 ± 8.61	12.3 ± 2.82	46.9 ± 2.89
Kimi K2.5	Curated Input	56.3 ± 7.36	24.7 ± 4.32	35.0 ± 6.01	91.2 ± 5.04	87.2 ± 5.94	12.4 ± 2.74	47.5 ± 2.90
Qwen3-VL-235B	ClinSeek	70.4 ± 6.77	35.7 ± 4.88	47.8 ± 6.27	79.2 ± 7.21	61.6 ± 8.64	6.0 ± 1.79	49.8 ± 2.91
Qwen3-VL-235B	Curated Input	60.3 ± 7.26	21.1 ± 4.34	32.8 ± 6.05	87.2 ± 5.94	72.8 ± 7.91	6.6 ± 1.94	43.9 ± 2.95
Gemma-4-26B-A4B-it	ClinSeek	78.9 ± 6.05	21.6 ± 5.20	38.4 ± 6.41	65.6 ± 8.44	71.2 ± 8.05	0.4 ± 0.83	44.9 ± 3.07
Gemma-4-26B-A4B-it	Curated Input	56.9 ± 7.35	21.4 ± 4.44	25.4 ± 5.75	79.2 ± 7.21	60.0 ± 8.71	0.0 ± 0.00	38.2 ± 2.95

Table 6: **Confidence intervals for AgentEHR five-task evaluation.** Each cell reports mean F1 score in percentage points with the 95% CI radius; Avg. pools the five subtasks.

Model	Diagnoses ($N = 100$)	Labs ($N = 100$)	Microbiology ($N = 100$)	Procedures ($N = 100$)	Transfers ($N = 100$)	Avg. ($N = 500$)
Claude Opus 4.6	58.5 ± 3.19	42.1 ± 3.96	27.2 ± 4.77	31.1 ± 3.16	20.9 ± 3.80	36.0 ± 2.05
Claude Sonnet 4.6	54.4 ± 2.99	35.6 ± 3.44	23.4 ± 3.95	26.3 ± 2.78	23.7 ± 3.81	32.7 ± 1.83
Kimi K2.5	46.9 ± 3.62	33.7 ± 4.04	18.9 ± 4.53	27.9 ± 3.76	22.1 ± 3.46	29.9 ± 1.93
MiniMax-M2.5	51.5 ± 3.69	29.0 ± 4.19	19.0 ± 3.85	22.0 ± 5.17	17.0 ± 3.80	27.7 ± 2.15
GLM-4.7	46.4 ± 3.39	28.6 ± 4.01	16.6 ± 3.87	23.7 ± 3.74	22.9 ± 4.06	27.6 ± 1.91
Qwen3-235B-A22B	30.6 ± 4.04	20.3 ± 3.37	17.3 ± 4.40	24.9 ± 5.51	9.6 ± 3.41	20.5 ± 1.96
gpt-oss-120b	27.3 ± 4.20	12.8 ± 3.26	12.4 ± 3.60	19.1 ± 5.36	7.6 ± 2.89	15.8 ± 1.84
Tongyi DeepResearch 30B-A3B	25.8 ± 4.55	14.9 ± 3.61	8.8 ± 3.12	17.9 ± 5.52	13.2 ± 4.79	16.1 ± 2.00
Gemma-4-26B-A4B-it	17.9 ± 4.47	18.5 ± 4.46	19.7 ± 5.25	11.2 ± 4.79	8.8 ± 3.59	15.2 ± 2.04
OpenSeeker-30B	20.4 ± 4.82	4.5 ± 2.22	12.8 ± 4.63	14.2 ± 5.57	10.6 ± 3.68	12.5 ± 1.97
Qwen3.5-35B-A3B (base)	36.6 ± 4.56	17.7 ± 3.84	16.2 ± 4.27	21.9 ± 4.33	18.1 ± 4.32	22.1 ± 2.00
ClinSeek-35B-A3B (ours, SFT)	55.4 ± 3.26	38.5 ± 3.57	27.6 ± 4.59	31.7 ± 3.17	16.7 ± 3.72	34.0 ± 1.98

Table 7: **ClinSeekAgent tool space.** ClinSeekAgent provides tools for patient-specific EHR retrieval, external medical knowledge search, and medical image analysis.

Source	Tool	Function
EHR	`ehr.load_ehr`	Load the patient-specific EHR database at the reference timestamp.
EHR	`ehr.get_table_description`	Retrieve table description and column information from database schema.
EHR	`ehr.get_table_names`	Retrieve available EHR and candidate tables.
EHR	`ehr.get_column_names`	Inspect the schema of a specified table.
EHR	`ehr.get_records_by_time`	Retrieve table records within a specified time range.
EHR	`ehr.run_sql_query`	Execute SQL for filtering, joining, aggregation, or trend analysis.
EHR	`ehr.get_candidates_by_semantic_similarity`	Retrieve candidate medical terms from dictionary tables.
EHR	`ehr.get_candidates_by_keyword`	Search diagnosis codes by keyword.
EHR	`ehr.get_latest_records`	Finds the latest timestamp and returns all records with that timestamp.
EHR	`ehr.think`	Record intermediate reasoning process
EHR	`ehr.finish`	Submit the final answer list
Web	`browser.search`	Search external medical knowledge sources.
Web	`browser.open`	Open and inspect retrieved pages or URLs.
Web	`browser.find`	Find exact terms or passages within an opened page.
Image	`image.dicom_processor`	Convert DICOM images to PNG and extract metadata.
Image	`image.image_visualizer`	Render images for inspection.
Image	`image.chest_xray_classifier`	Predict probabilities for chest X-ray pathologies.
Image	`image.chest_xray_report_generator`	Generate structured chest X-ray findings and impression.
Image	`image.xray_phrase_grounding`	Ground a specified radiographic finding in the image.
Image	`image.chest_xray_segmentation`	Segment anatomical structures in chest radiographs.

Table 8: **Default inference settings.** We use the same settings across models whenever supported by the corresponding backend.

Setting	Value
Temperature	1.0
Maximum output tokens	8192
Maximum agent rounds	200
Maximum concurrency	6
Maximum tool-result length	100,000 characters
Image maximum edge	1568 pixels
Stopping criterion	`finish` tool call or maximum-round limit
Primary metric	Mean sample-wise F1

Table 9: SFT configuration for ClinSeek-35B-A3B. The model is fine-tuned on long-horizon ClinSeekAgent trajectories rendered in native tool-call format with a 52K-token maximum sequence length.

Component	**Configuration**
Base model	Qwen3.5-35B-A3B
Teacher model	Claude Opus 4.6
Training objective	Supervised fine-tuning on ClinSeekAgent trajectories
Training data format	Native tool-call format with `<tool_call>` / `<tool_response>`
Training / validation size	7,204 / 147 examples after length filtering
Maximum sequence length	52,000 tokens
Dropped examples	18.3% due to length filtering
Training epochs	3
Global batch size	32
Micro batch size	1 per GPU
Optimizer	Megatron optimizer with CPU offload
Learning rate	2×10^{-5}
Minimum learning rate	2×10^{-6}
Learning rate schedule	Cosine decay with 10 warmup steps
Weight decay	0.1
Gradient clipping	1.0
Precision	bfloat16
Backend	Megatron + mbridge
Hardware	8$\times$ H200 GPUs
Tensor parallelism	2
Pipeline parallelism	1
Expert parallelism	8
Expert tensor parallelism	1
Context parallelism	1
Parameter / gradient / optimizer offload	Enabled
Random seed	42

F More Case Study

F.1 Failure mode analysis

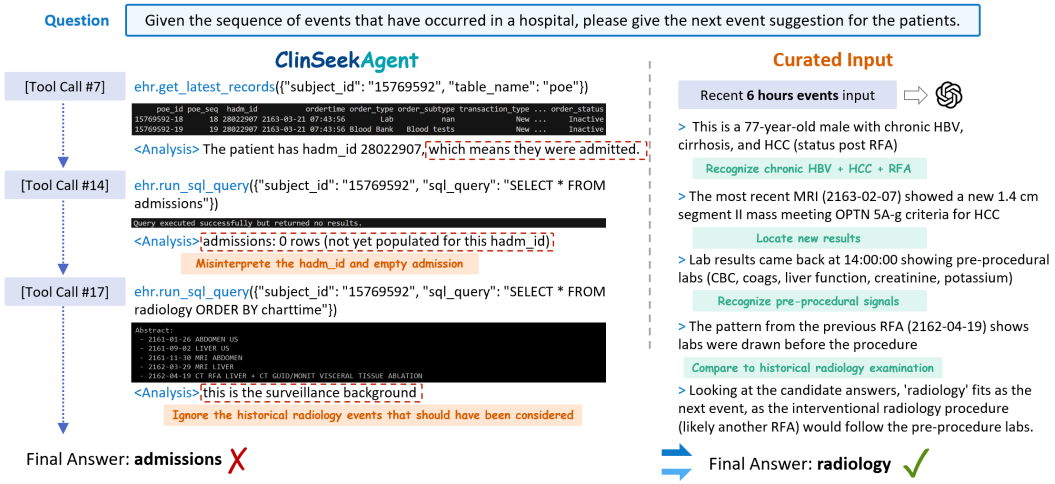
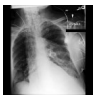


Figure 6: **Comparison between the ClinSeekAgent pipeline and the Curated Input baseline.** Our pipeline fails to locate critical patient information on a decision-making prediction task.

F.2 More successful cases



Question

Use all EHR information and the linked chest X-ray before the cutoff to decide whether the patient will die within the next 24 hours. Return only the final answer text.

Step 1 — CXR Classifier (Call #1)

Raw Input
{"image_path": ".../s59513062/5e78393d-...319dc8a9.jpg"}

Raw Output (probabilities 0–1, higher = more likely)
Lung Opacity: 0.700 Cardiomegaly: 0.659 Enlarged Cardiomediastinum: 0.620
Atelectasis: 0.610 Effusion: 0.587 Infiltration: 0.555
Consolidation: 0.541 Edema: 0.514 Pneumonia: 0.366
Pneumothorax: 0.505

Model's Conclusion
Multiple significant pathologies (cardiomegaly + bilateral opacity + probable effusion), but no high-probability acute catastrophe (pneumonia 0.37, pneumothorax 0.50 ≈ baseline). Findings consistent with chronic cardiopulmonary disease, not unequivocally terminal.

Step 2 — CXR Report Generator (Call #2)

Raw Input
{"image_path": ".../s59513062/5e78393d-...319dc8a9.jpg"}

Raw Output (key excerpts)
FINDINGS: "no focal consolidation, pleural effusion or pneumothorax identified"
IMPRESSION: "persistent retrocardiac consolidation may represent
pneumonia or aspiration. Likely a layering left effusion.
Right lung grossly clear. No pneumothorax."

Model's Conclusion
Report is internally inconsistent (FINDINGS denies what IMPRESSION affirms). Used only as soft prior for "retrocardiac process + possibly small left effusion + right lung clear". Notably, the right lung being clear contradicts a fulminant bilateral ARDS picture — a point ClinSeek later leans on.

Step 3 — Full-Day Events for May 14 (Call #7)

Raw Input
{"subject_id": "16773288", "table_name": "events",
"start_time": "2134-05-14 00:00:00",
"end_time": "2134-05-14 16:14:23"}

Raw Output (~99 KB, truncated)
- Chest tube #1 / R pleural #2 outputs steadily draining
- Foley output present hourly (preserved renal perfusion)
- Vitals across the day (extracted by model in Think #8):
  - HR: 73–85, sinus + occasional PVCs
  - SBP: 121 → 92 (10:00 dip) → recovered to 105 (16:00)
  - MAP: dipped to 56 at 10:00, recovered
  - SpO_2: 92–99 (nadir 92 at 10:00 on NC)

Model's Conclusion (Think #8)
"BP dropping trend … then recovered to 105 at 4 pm." → Hemodynamics are V-shaped, not monotonically decompensating. Renal output preserved. This is a struggling but stabilizing trajectory, not terminal hypoperfusion.

Step 4 — Prior-Day (May 13) Vital Baseline (Call #9)

Raw Input
{"sql_query": "SELECT * FROM events
WHERE label IN ('heart rate','arterial blood pressure systolic',
'arterial blood pressure mean','respiratory rate',
'o2 saturation pulseoxymetry','richmond-ras scale')
AND charttime >= '2134-05-13 00:00:00'
AND charttime < '2134-05-14 00:00:00'
ORDER BY charttime, label"}

Raw Output (truncated)
2134-05-13 00:00 HR 62 SBP 115 MAP 69 RR 23 SpO_2 93
2134-05-13 01:00 HR 60 SBP 120 MAP 76 RR 23 ...
... (stable through the day) ...

Model's Conclusion
24-hour-prior baseline shows stable hemodynamics and oxygenation on minimal support. The 5/14 deterioration is therefore acute-on-chronic, not a slow death spiral — the kind of acute event ICU teams routinely catch and stabilize.

Figure 7: **A case of Medmod Decompensation.** Page 1.

Step 5 — ABG / Lab Series via linksto IS NULL (Call #13)

Raw Input

```
{"sql_query": "SELECT * FROM events WHERE linksto IS NULL
        ORDER BY charttime DESC LIMIT 30"}
```

Raw Output (truncated)

2134-05-14 16:05 pH 7.21 pCO_2 90 pO_2 142 BE +4 Ca^{2+} 1.13 K^+ 4.5
2134-05-14 15:32 pH 7.20 pCO_2 91 pO_2 134 ...
2134-05-14 13:33 pH 7.25 pCO_2 78 pO_2 79
2134-05-14 13:02 pH 7.30 pCO_2 75 pO_2 58 Lactate 1.2 Glucose 118

Model's Conclusion (Think #28)

ABG trend is worsening respiratory acidosis (pH 7.30 → 7.21). However:
- Lactate 1.2 = normal → no tissue hypoperfusion / shock.
- pO_2 142 on FiO_2 100% → P/F = 142 (moderate ARDS, not unsurvivable).
- BE +4 with severe acidemia → mixed disorder, but metabolic side compensating.

This is isolated respiratory failure, not multi-organ failure — a critical distinction.

Step 6— Code Status / Comfort Measures (Call #31)

Raw Input

```
{"sql_query": "SELECT * FROM events
  WHERE label LIKE '%code status%' OR label LIKE '%comfort%'
    OR label LIKE '%goals of care%' OR label LIKE '%cmo%'
    OR label LIKE '%DNR%' OR label LIKE '%DNI%'
  ORDER BY charttime DESC LIMIT 10"}
```

Raw Output (truncated)

2134-05-14 16:00 status and comfort = "Asleep"
2134-05-14 14:00 status and comfort = "Asleep"
2134-05-09 ... status and comfort = "Calm"
(no DNR/DNI/CMO entries found)

Model's Conclusion

No DNR / DNI / Comfort-Measures-Only orders. "Asleep" reflects sedation under mechanical ventilation, not terminal unresponsiveness. The team is treating with full code intent → goals of care are still curative.

Step 7 — Ventilator-Setting Escalation (Call #37)

Raw Input

```
{"sql_query": "SELECT * FROM events
  WHERE (label LIKE '%fio2%' OR label LIKE '%tidal%'
    OR label LIKE '%peep%' OR label LIKE '%minute%'
    OR label LIKE '%pip%'  OR label LIKE '%plateau%'
    OR label LIKE '%set rr%'
    OR label LIKE '%respiratory rate (set)%')
  AND charttime >= '2134-05-14 13:00:00'
  ORDER BY charttime"}
```

Raw Output (extracted by model)

	13:30	14:00	14:30	16:00	
PEEP set	8	10	10	12	
Vt set	—	450	—	600	
RR set	—	16	—	20	
Minute Vol	4.3	7.6	10.1	14.7	(L/min)
FiO_2	100	100	100	100	(Step 8b)

Model's Conclusion (Think #39)

Settings are aggressively escalated, not de-escalated. The 16:05 ABG was drawn only 5 mia lagged menutes after the largest jump (Vt 600, RR 20, MV 14.7). Insufficient time for CO_2 to wash out → the bad ABG reflects asurement, not a treatment ceiling. Active resuscitation, not withdrawal.

Step 8 — Final Synthesis & Answer (ehr.think #39 tail → ehr.finish Call #40)

Final Synthesis (excerpt)

▌"On balance, this patient is very sick but still has a chance of surviving 24 hours ... (1) Hemodynamics are stable; (2) Lactate is normal; (3) Vent settings were just escalated and may not have had time to take effect; (4) CVICU has advanced capabilities. Final answer: no."

ehr.finish Call

```
{"response": ["no"]}
```

Result: Finish ✅ matches gold ["no"].

Figure 7: **A case of Medmod Decompensation.** Page 2.

Question

Use all EHR information and the linked chest X-ray before ICU discharge to identify every phenotype group that applies to the patient.

Step 1 — CXR Classifier (Call #1)

Raw Input
{"image_path": ".../s54270131/3062c495-...d113a650.jpg"}

Raw Output (probabilities, top entries)
Lung Opacity: 0.829 Effusion: 0.715
Atelectasis: 0.686 Infiltration: 0.636
Cardiomegaly: 0.634 Consolidation: 0.620
Edema: 0.617 Enlarged Cardiomediastinum: 0.576
Pneumonia: 0.546 Lung Lesion: 0.557

Model's Conclusion
Strong multi-finding image: bilateral opacities + effusion + atelectasis + consolidation. Each ≥ 0.55, suggesting a combined respiratory picture rather than a single pathology — points toward Pleurisy/effusion/collapse + Pneumonia + Respiratory failure phenotype cluster, not a single label.

Step 2 — CXR Report Generator (Call #2)

Raw Input
{"image_path": ".../s54270131/3062c495-...d113a650.jpg"}

Raw Output (key excerpts)
FINDINGS:
"right internal jugular central venous catheter unchanged in position.
No pneumothorax. Small left pleural effusion.
Numerous bilateral pulmonary nodules, consistent with known
metastatic disease."

IMPRESSION:
1. RIJ central line tip in distal SVC.
2. Lung volumes remain low.
3. Diffuse reticular pattern bilaterally — may represent edema or infection.
4. Stable small bilateral pleural effusions.

Model's Conclusion
- Bilateral pulmonary metastases → patient has known malignancy (priors metastatic).
- Diffuse reticular pattern → edema vs. infection — corroborates pneumonia phenotype if labs/vitals support it.
- Bilateral effusions + low lung volumes → atelectasis + pleural effusion phenotypes.
- Central line documented → prior procedural complications likely.

Step 3 — Comprehensive Vitals / Labs Sweep (Call #11)

Raw Input
{"sql_query":
"SELECT DISTINCT label, value, valuenum, charttime FROM events
WHERE label IN ('heart rate','respiratory rate',
'non invasive blood pressure systolic',
'non invasive blood pressure mean',
'temperature celsius','o2 saturation pulseoxymetry',
'wbc','platelet count','hemoglobin','creatinine (serum)',
'lactic acid','ph (arterial)','arterial o2 pressure',
'arterial co2 pressure','anion gap','albumin', ...)
ORDER BY charttime DESC LIMIT 200"}

Raw Output (truncated, ~15 KB)
2166-08-01 02:00 HR 112 SBP 94 MAP 63 RR 21 SpO_2 95
2166-08-01 01:00 HR 101 SBP 120 MAP 84 RR 24 SpO_2 99
... (recent values trending toward stability) ...
... earlier rows: HR up to 130s, MAP into the 50s, fevers logged ...

Model's Conclusion (Think #10 → #28)
The sepsis triad is present: tachycardia (HR > 100 sustained), low MAP episodes, and fever spikes (later isolated separately). Patient meets clinical SIRS/sepsis criteria; Septicemia + Shock become primary working hypotheses. Notes that dataset doesn't expose vasopressor administration directly — agent flags this as a gap to chase.

Step 4 — Vancomycin Trough Levels (Call #27) — Treatment as Evidence

Raw Input
{"sql_query":
"SELECT label, value, valuenum, charttime FROM events
WHERE label IN ('vancomycin (trough)','vancomycin (random)',
'tobramycin (random)')
AND charttime >= '2166-07-26 00:00:00'
ORDER BY charttime DESC"}

Raw Output
vancomycin (trough) 10.8 2166-07-29 22:53
vancomycin (trough) 11.3 2166-07-28 18:59
vancomycin (trough) 9.2 2166-07-27 18:25

Model's Conclusion
Three consecutive day troughs in the 9.2–11.3 µg/mL therapeutic range — patient is on scheduled IV vancomycin for ≥ 3 days. This is treatment-side proof of an active bacterial infection, locking in Septicemia even without an explicit blood-culture row. (Reasoning baseline never sees this — antibiotic levels aren't in the latest-80-row snapshot.)

Figure 8: **A case of Medmod Phenotyping.** Page 1.

Step 5 — Fever Isolation Query (Call #30)

Raw Input

```
{"sql_query":
 "SELECT label, value, valuenum, charttime FROM events
  WHERE label IN ('temperature celsius','temperature fahrenheit')
  AND charttime >= '2166-07-26 00:00:00'
  AND (valuenum > 100.4
      OR (label = 'temperature celsius' AND valuenum > 38.0))
  ORDER BY charttime"}
```

Raw Output (truncated)

```
2166-07-26 20:00  T 101.8 °F
2166-07-26 23:25  T 40.0  °C  ← peak
2166-07-27 00:00  T 39.3  °C
2166-07-27 20:00  T 103.8 °F
2166-07-27 22:00  T 102.8 °F
2166-07-28 20:00  T 101.3 °F
... (defervescence by 7/30)
```

Model's Conclusion

Peak temperature 40 °C / 104 °F on admission day, sustained febrile course for ~48 h. Combined with leukocytosis and vancomycin → unambiguous Septicemia. Defervescence by 7/30 corresponds with antibiotic effect, which is consistent with treated-but-real sepsis (not chronic carrier state).

Step 6 — Sustained Hypotension Query (Call #43) — Shock Phenotype

Raw Input

```
{"sql_query":
 "SELECT label, value, charttime FROM events
  WHERE label LIKE '%blood pressure%'
  AND charttime >= '2166-07-26 00:00:00'
  AND valuenum < 65 AND label LIKE '%mean%'
  ORDER BY charttime DESC LIMIT 30"}
```

Raw Output (truncated)

```
non invasive blood pressure mean  63  2166-08-01 02:00
non invasive blood pressure mean  59  2166-07-31 20:00
non invasive blood pressure mean  55  2166-07-31 09:00
non invasive blood pressure mean  52  2166-07-31 04:00
non invasive blood pressure mean  51  2166-07-31 01:23
non invasive blood pressure mean  52  2166-07-30 01:30
... (≥ 20 hits ≤ 65 across multiple days)
```

Model's Conclusion

MAP < 65 sustained, dropping as low as 51. In the context of confirmed septicemia, this is the textbook hemodynamic definition of septic shock — even though no chartevents row names a vasopressor (a known data-coverage limit). Agent decides to label Shock based on the persistent MAP pattern + sepsis context.

Step 7 — HCUP CCS Phenotype Vocabulary Recovery via Browser (Call #77) — The Distinguishing Move

Raw Input

```
{"loc": 514, "num_lines": 50, "cursor": 18}
```

(after 6 prior browser.search/browser.find calls navigating to the Harutyunyan-2019 ACL Anthology PDF for the MIMIC-III phenotype taxonomy)

Raw Output

```
... Table 3: The 25 pre-defined phenotypes in MIMIC-III ...
- Acute and unspecified renal failure
... [continues to 25 names] ...
```

Model's Conclusion

This is the decisive vocabulary step. The phenotyping benchmark is closed-set with 25 specific HCUP CCS labels — using natural-language descriptions like "sepsis" or "atelectasis" would score 0.
… Compare to the reasoning baseline, which lacked this taxonomy and emitted MIMIC-CXR-style labels (Atelectasis / Pleural Effusion / Support Devices) → all three score 0 against the gold key. This single call is the single largest delta between paradigms.

Final Answer (Call #101 — ehr.finish)

```
{"response": [
  "Septicemia (except in labor)",
  "Shock",
  "Pleurisy; pneumothorax; pulmonary collapse",
  "Pneumonia (except that caused by tuberculosis or sexually transmitted disease)",
  "Respiratory failure; insufficiency; arrest (adult)",
  "Fluid and electrolyte disorders",
  "Complications of surgical procedures or medical care"
]}
```

Versus gold (5 labels): all 5 recovered ✅; 2 over-predictions (Pleurisy/collapse, Respiratory failure) — both clinically defensible from the CXR + low lung volumes + bilateral effusions.

Figure 8: **A case of Medmod Phenotyping.** Page 2.

Question

Given the sequence of events that have occurred in a hospital, please predict whether the patient's hospital stay will exceed 3 days.

Step 1 — Admissions Latest Record (Call #2)

Raw Input

{"subject_id": "12954732", "table_name": "admissions"}

Raw Output

hadm_id admittime dischtime admission_type service
25254375 2138-11-17 22:09:00 2138-11-21 12:38:00 URGENT OBS
discharge_location: HOME insurance: Medicaid race: BLACK/AFRICAN AMERICAN
hospital_expire_flag: 0

Model's Conclusion

Admission was URGENT to OBS service (not elective scheduled C-section).

Step 2 — Transfers Trajectory (Call #10)

Raw Input

{"subject_id": "12954732", "table_name": "transfers",
"start_time": "2138-11-17 00:00:00",
"end_time": "2138-11-18 21:49:39"}

Raw Output

eventtype careunit intime outtime
admit Labor & Delivery 2138-11-17 22:10:50 2138-11-17 23:15:33
transfer Labor & Delivery 2138-11-17 23:15:33 2138-11-18 03:30:04
transfer Labor & Delivery 2138-11-18 03:30:04 2138-11-18 05:32:42
transfer Obstetrics Postpartum 2138-11-18 05:32:42 2138-11-21 12:46:01

Model's Conclusion

Patient cycled through three L&D rooms over 7+ hours before delivery (23:15 → 03:30 → 05:32 transition) — a pattern consistent with prolonged labor, not a precipitous delivery.

Step 3 — Labs: Anemia Signal (Call #11)

Raw Input

{"subject_id": "12954732", "table_name": "labevents",
"start_time": "2138-11-17 00:00:00",
"end_time": "2138-11-18 21:49:39"}

Raw Output (truncated)

charttime label value ref_range flag
2138-11-17 22:05:00 Hematocrit 30.7 % 34.0 – 45.0 abnormal (STAT)
2138-11-17 22:05:00 Hemoglobin 9.9 g/dL 11.2 – 15.7 abnormal (STAT)
2138-11-17 22:05:00 MCH 27.8 pg 26 – 32 ok
2138-11-17 22:05:00 MCHC 32.2 g/dL 32 – 37 ok
2138-11-17 22:05:00 MCV 86 fL 82 – 98 ok

Model's Conclusion

Pre-delivery anemia (Hgb 9.9, Hct 30.7%, both flagged abnormal). MCV/MCHC normal → normocytic, consistent with dilutional / pregnancy-physiologic but borderline-low enough to require postpartum monitoring after blood loss. This is exactly the kind of "complicating condition" that bumps a delivery into DRG-with-CC territory and justifies a longer stay. Reasoning baseline never saw this row.

Figure 9: **A case of Length of Stay.** Page 1.

Step 4 — Prescriptions Horizon (Call #13)

Raw Input

{"subject_id": "12954732", "table_name": "prescriptions",
"start_time": "2138-11-17 00:00:00",
"end_time": "2138-11-18 21:49:39"}

Raw Output (key rows)

drug	starttime	stoptime	
Penicillin G Potassium	2138-11-18 06:00	2138-11-19 08:00	← GBS prophylaxis
Iso-Osmotic Dextrose	2138-11-18 06:00	2138-11-19 08:00	
Ibuprofen	2138-11-18 ...	2138-11-21 17:00	← horizon
Acetaminophen	2138-11-18 ...	2138-11-21 17:00	← horizon
Docusate	2138-11-18 ...	2138-11-21 17:00	← horizon
Dermoplast (topical)	2138-11-18 ...	2138-11-21 17:00	← horizon

Model's Conclusion

Two findings:

1. Penicillin G = GBS-positive mother on intrapartum prophylaxis → confirms a known indication for prolonged postpartum observation of the newborn (and by extension the dyad).
2. Multiple postpartum med stop times = 2138-11-21 17:00. Pharmacy team has already written orders extending >72 h past cutoff. Stop-time horizons are a clinically robust planning signal — providers don't write 4-day orders for a patient they expect to discharge in 24h. This is the cleanest non-leaky evidence for predicting LOS > 3 days.

Step 5 — DRG Codes (Call #22) — The Distinguishing Move

Raw Input

{"sql_query": "SELECT * FROM drgcodes WHERE hadm_id = 25254375"}

Raw Output

drg_type	drg_code	description	severity
APR	560	VAGINAL DELIVERY	1.0
HCFA	806	VAGINAL DELIVERY W/O STERILIZATION/D&C W CC	—

Model's Conclusion

HCFA 806 = "Vaginal Delivery WITH Complications/Comorbidities" (the "W CC" suffix). DRG codes with-CC are explicitly defined for cases requiring extended resource use. This single billing-level row encodes the answer: the institution itself classified this admission as complicated for reimbursement purposes.

Step 6 — Final Synthesis & Answer (ehr.think → ehr.finish Call #24)

Synthesis (excerpt)

> "Multiple factors point to a stay exceeding 3 days: anemia requiring monitoring, CC designation in DRG, medication orders extending to 11/21, and the patient's history of longer postpartum stays … The stay will likely exceed 3 days."

ehr.finish Call

{"response": ["yes"]}

Result: yes ✅ matches gold.

Figure 9: **A case of Length of Stay.** Page 2.