EviDx: Evidence-Aware Active Diagnosis with Scaffolded LLM Agents

Lihang Zeng[1], **Shaoting Zhang**[1*], **Xiaofan Zhang**[1,2*]
[1]Shanghai Jiao Tong University, [2]Shanghai Innovation Institute

Abstract

Clinical diagnosis is an active evidence-seeking process in which clinicians acquire evidence, update competing hypotheses, and decide when the available evidence is sufficient for diagnosis. Yet many medical diagnosis systems built around large language models (LLMs) still formulate diagnosis as static case-to-answer prediction, with limited support for evidence acquisition. Agentic LLMs offer a dynamic alternative through tool use and intermediate diagnostic trajectories, but existing systems often under-specify how patient evidence should be exposed, scaffolded, and controlled at runtime. We introduce EviDx, an evidence-aware active diagnosis framework that pairs patient-specific diagnostic environments with a clinical diagnostic scaffold and an observer-guided runtime harness. In EviDx, $\mathcal{E}$-Synthesis constructs interactive environments from raw clinical cases; the scaffold organizes role-specialized agents, evidence tools, and evolving evidence states; and the harness regulates diagnostic termination by tracking uncertainty and evidence coverage. A 3-level evaluation pyramid assesses execution robustness, reasoning dynamics, and diagnostic outcomes. Experiments show that EviDx improves diagnostic performance and process stability while revealing model-dependent capability boundaries.

1 Introduction

Clinical diagnosis (Dx) is not a single-step prediction problem. It is an active evidence-seeking process in which clinicians iteratively acquire evidence, integrate new findings into a differential diagnosis, refine competing hypotheses, and decide whether the current evidence is sufficient for diagnosis or whether further information should be gathered (Sackett, 1997; Pauker and Kassirer, 1980). A diagnostic conclusion is therefore meaningful only when it is supported by an adequate evidence trajectory, not merely by a plausible final answer.

Many existing medical diagnosis systems built around LLMs still formulate diagnosis as static case-to-answer prediction. A model receives a complete clinical vignette or answer set upfront, and then directly produces a diagnosis or selects from multiple choices. Such methods can elicit medical knowledge and final diagnoses, but they collapse evidence acquisition, evidence integration, and termination into a single prompt. As a result, they under-specify process failures that arise in active diagnosis, such as premature commitment before sufficient evidence has been acquired.

Tool-augmented and agentic medical systems offer a path toward dynamic diagnosis because they can query patient records and retrieve external medical knowledge (Schmidgall et al., 2024; Jiang et al., 2025; Wang et al., 2025b). However, dynamic diagnosis is not obtained by tool access alone. The capability of an LLM agent is shaped by the environment that exposes patient evidence, the scaffold that organizes evidence actions, clinical roles, and evidence states, and the harness that monitors execution at runtime. Existing medical-agent systems often organize interaction through fixed workflows, turn budgets or consensus checks (Oh et al., 2025; Chen et al., 2025c), leaving evidence sufficiency and termination control only weakly modeled. This raises a general question: *Can clinical LLM agents become more capable when interactive diagnostic environments are coupled with explicit evidence scaffolds and runtime harnesses for active diagnosis?*

We introduce **EviDx**, an evidence-aware active diagnosis framework for scaffolded and harnessed clinical LLM agents. First, $\mathcal{E}$-Synthesis converts raw clinical text into a patient-specific in-

*Corresponding authors

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33154–33179
October 24-29, 2026 ©2026 Association for Computational Linguistics

Work	Diagnosis Setting		System Design		Evaluation
	Interactive Env.	Evidence Access	Clinical Scaffold	Runtime Harness	Process-Level Eval.
MDAgents (Kim et al., 2024)	✗	✓	✓	✗	✗
Agent Hospital (Li et al., 2024)	✓	✗	✓	✗	✗
AgentClinic (Schmidgall et al., 2024)	✓	✓	✓	✗	✓
MedAgentBench (Jiang et al., 2025)	✓	✓	✗	✗	✓
AutoMedic (Oh et al., 2025)	✓	✗	✓	✗	✓
EviDx	✓	✓	✓	✓	✓

Table 1: Qualitative comparison of representative clinical LLM diagnosis systems and medical-agent evaluation environments. We summarize each work along design axes relevant to evidence-aware active diagnosis, including interactive evidence acquisition, clinical scaffolding, runtime control, and process-level evaluation.

teractive environment in which evidence can be acquired through standardized interfaces. Second, the Clinical Dx Scaffold organizes diagnostic execution through role-specialized agents, evidence tools, and evolving evidence states, including patient-record and external-knowledge access through Model Context Protocol (MCP; Anthropic, 2024) tool interfaces. Third, the Observer-Guided Harness monitors diagnostic uncertainty and runtime evidence coverage from the evolving diagnostic states to regulate diagnostic termination decisions. Finally, the 3-level evaluation pyramid evaluates the resulting trajectories across execution robustness, reasoning dynamics, and diagnostic outcomes.

Our main contributions are as follows:

- We propose EviDx, an evidence-aware active diagnosis framework that transforms static clinical cases into interactive diagnostic environments and scaffolds LLM agents with clinical roles, evidence tools, and evidence states.
- We introduce an Observer-Guided Harness that controls diagnostic progression by tracking uncertainty and evidence coverage, enabling explicit diagnostic termination decisions.
- We design a 3-level evaluation pyramid to assess EviDx across execution robustness, reasoning dynamics, and outcomes. Experiments show that EviDx improves diagnostic performance and process stability while revealing model-dependent capability boundaries. Code and resources are available.[1]

2 Related Work

LLMs for Clinical Diagnosis. LLMs have demonstrated strong performance on medical question answering and diagnosis-oriented benchmarks, including real-world clinical cases and expert-level multiple-choice settings (Chen et al., 2025a; Zuo et al., 2025; Zhu et al., 2025). These settings can elicit medical knowledge and final-answer behavior, but they often expose the case as a fixed input and ask the model to generate the diagnosis in one pass. In contrast, clinical diagnosis is an evidence-seeking process: physicians iteratively acquire patient information, integrate findings into a differential diagnosis, and decide whether the current evidence is sufficient. This motivates diagnosis settings that represent clinical cases as evolving evidence contexts.

Tool-Augmented and Agentic LLMs. Tool-augmented medical agents seek to move beyond static QA by allowing models to retrieve patient data, consult external medical knowledge, or coordinate role-specialized reasoning (Wang et al., 2025b; Li et al., 2024; Wu et al., 2025). Interactive clinical environments and EHR-oriented benchmarks further stress that medical agents must act within patient-specific contexts rather than only answer fixed prompts (Schmidgall et al., 2024; Jiang et al., 2025; Liu et al., 2026b). These works motivate the shift from answer generation to evidence seeking, while also exposing a remaining control problem: tool access must be coupled with decisions about which evidence is still missing and when the diagnostic process should terminate.

Scaffolds and Harnesses for LLM Agents. Recent LLM-agent research increasingly treats agents as executable systems embedded in environments, where planning, tool use, memory or state, and trace-level evaluation are core design concerns (Yehudai et al., 2025; Wei et al., 2026; Chen et al., 2025b). At the same time, runtime-enforcement work frames agent reliability as an execution-control problem: external monitors can inspect traces, apply constraints, and intervene

[1] https://github.com/LihangZeng/EviDx

before unsafe or invalid actions are committed (Wang et al., 2025a, 2026). In this literature, scaffolds provide structured support around an agent, including roles, tools, state representations, and workflows, while harnesses provide runtime machinery for monitoring and controlling execution. This distinction is useful for active diagnosis because evidence acquisition and termination are coupled but separable: one design layer organizes how an agent seeks and records evidence, while another decides whether the evolving trajectory is sufficient for termination.

Evaluation of Clinical Reasoning and Medical Agents. Outcome-only evaluation can hide whether a model acquired the right evidence, used it correctly, or merely guessed the final answer. Recent medical-agent benchmarks therefore move toward interactive EHR tasks, virtual patient conversations, and execution-grounded evaluation (Oh et al., 2025; Jiang et al., 2025; Liu et al., 2026b). For active diagnosis, evaluation should therefore distinguish execution robustness, reasoning dynamics, and diagnostic outcomes instead of collapsing the entire process into final accuracy.

Table 1 makes these distinctions explicit by comparing representative systems along the axes emphasized above: interactive evidence acquisition, clinical scaffolding, runtime control, and process-level evaluation.

3 Problem Formulation

We formulate clinical LLM diagnosis as active differential diagnosis: an iterative process in which a model acquires evidence, updates a differential diagnosis, and decides when the accumulated trajectory is sufficient for a final answer.

Active Differential Diagnosis. Let x denote a clinical case and let $\mathcal{E}_x$ denote an interactive diagnostic environment associated with the patient. At step t, the system takes an evidence-seeking action a_t, receives an observation r_t from $\mathcal{E}_x$, and appends it to the diagnostic trajectory $\mathcal{T}_t = \{x_0, (a_1, r_1), \ldots, (a_t, r_t)\}$. The system maintains a differential diagnosis set $D_t \subseteq D$ and reports a normalized diagnostic belief:

$$\widehat{P}_t = \{\hat{p}_t(d_i \mid \mathcal{T}_t)\}_{d_i \in D_t}, \sum_{d_i \in D_t} \hat{p}_t(d_i \mid \mathcal{T}_t) = 1. \quad (1)$$

Here, $\widehat{P}_t$ represents the model-reported belief over competing diagnostic hypotheses. Active diagnosis therefore consists of evidence acquisition, evidence integration, diagnosis refinement, and termination decisions over the evolving trajectory $\mathcal{T}_t$.

Evidence Sufficiency and Termination Decisions. At each step, the central decision is whether the current trajectory is sufficient for diagnosis or whether additional evidence should be acquired. We characterize diagnostic uncertainty using the entropy of the reported differential diagnosis:

$$H(t) = -\sum_{d_i \in D_t} \hat{p}_t(d_i \mid \mathcal{T}_t) \log_2 \hat{p}_t(d_i \mid \mathcal{T}_t). \quad (2)$$

A lower $H(t)$ indicates a more concentrated belief state, while a higher $H(t)$ suggests that the differential remains unresolved. However, diagnostic confidence and evidence sufficiency are distinct: a model may report a confident belief before acquiring enough patient-specific evidence. The active diagnosis problem therefore requires termination decisions that consider both diagnostic resolution and the evidential support accumulated in the trajectory.

4 Proposed Methodology

EviDx contains four components: $\mathcal{E}$-Synthesis, Clinical Dx Scaffold, Observer-Guided Harness, and a 3-level evaluation pyramid. Together, they convert static clinical cases into active diagnostic environments, structure evidence acquisition, regulate termination, and evaluate the resulting diagnostic trajectories.

4.1 $\mathcal{E}$-Synthesis

EviDx begins by converting a static clinical case into a patient-specific diagnostic environment. For each patient case x, the Context Initializer processes the raw clinical text, converts it into structured patient data, and loads these records into the EHR Server. $\mathcal{E}_x$ then couples this patient-record backend with evidence-access tools, forming an interactive environment in which subsequent diagnostic actions can query patient-specific records and external medical knowledge.

The Context Initializer preserves clinically relevant ambiguity, qualifiers and negative findings, so that structuring the case does not collapse the evidence available for diagnosis. In this way, $\mathcal{E}$-Synthesis changes the diagnosis setting from static text consumption to active evidence acquisition.

4.2 Clinical Dx Scaffold

The Clinical Dx Scaffold structures how the LLM interacts with $\mathcal{E}_x$. It consists of role-specialized

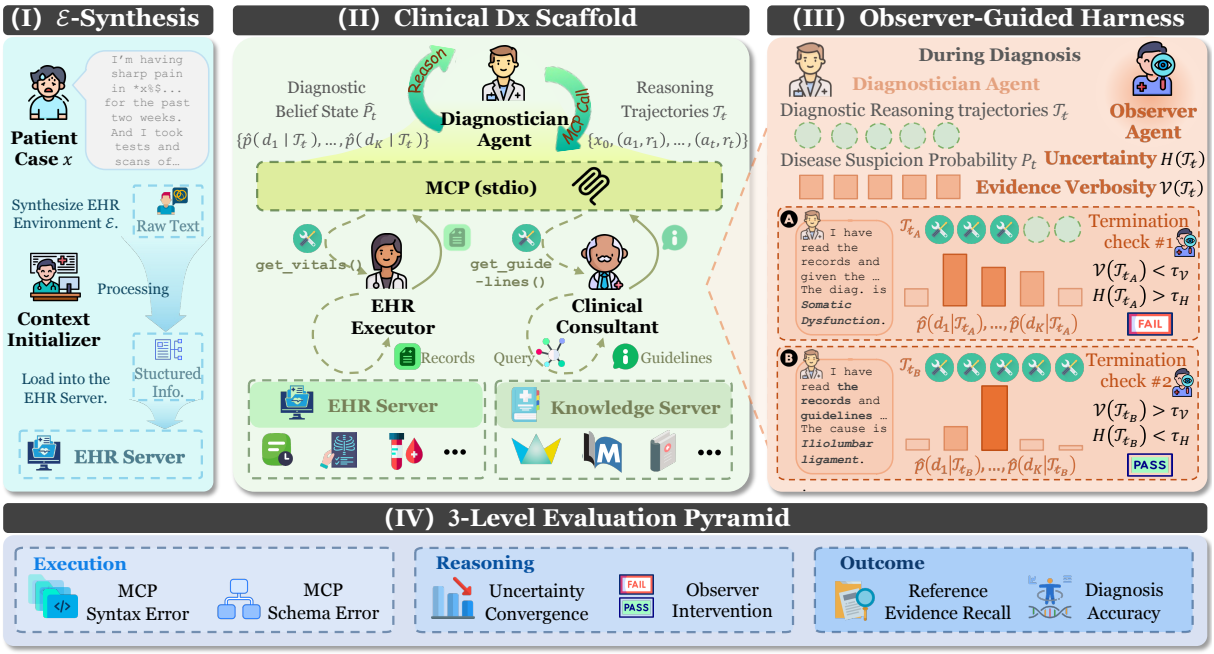


Figure 1: Overview of EviDx. **(I)** $\mathcal{E}$-Synthesis converts raw clinical text into a patient-specific diagnostic environment through the Context Initializer. **(II)** The Clinical Dx Scaffold structures evidence acquisition and diagnostic reasoning through clinical roles, evidence tools, and evidence states. **(III)** The Observer-Guided Harness monitors uncertainty $H(t)$ and runtime evidence coverage $\mathcal{V}(t)$ to govern diagnostic termination. **(IV)** The 3-level evaluation pyramid evaluates execution robustness, reasoning dynamics, and diagnostic outcomes.

components, evidence states, and an active diagnostic workflow rather than generic role prompting. The Diagnostician maintains the differential diagnosis D_t, the reported belief state $\widehat{P}_t$, and the trajectory $\mathcal{T}_t$. At each step, it identifies the next information need and issues an evidence action as a Model Context Protocol (MCP; Anthropic, 2024) tool call, which standardizes tool invocation and data exchange. The EHR Executor and Clinical Consultant handle these calls by retrieving patient records or external evidence, respectively, and return an observation r_t to the Diagnostician.

Diagnostician. The Diagnostician is the central reasoning component. Given $\mathcal{T}_{t-1}$, it updates the differential diagnosis, reports $\widehat{P}_t$, and decides whether to query patient data, consult medical knowledge, or attempt to finish.

EHR Executor. The EHR Executor handles deterministic access to the EHR Server within $\mathcal{E}_x$, retrieving requested records such as demographics, history, labs, and imaging reports, and returning them as standardized observations.

Clinical Consultant. The Clinical Consultant handles external evidence access. For knowledge-directed actions, it decomposes the request, retrieves relevant literature from the Knowledge Server using hybrid dense and BM25 sparse retrieval following MedRAG (Xiong et al., 2024), and returns a response to the Diagnostician.

4.3 Observer-Guided Harness

A scaffold alone is insufficient for active diagnosis: even when agents can request evidence, the system still needs runtime control over whether the current trajectory is ready to terminate. The Observer-Guided Harness provides this control. Whenever the Diagnostician proposes to finish, the harness evaluates whether the trajectory is both sufficiently resolved and sufficiently explored according to runtime evidence coverage. If not, it returns feedback and forces continued evidence acquisition.

Diagnostic Uncertainty. The harness uses $H(t)$ from Eq. (2) to measure whether the differential diagnosis remains unresolved. High entropy indicates that probability mass is still distributed across competing diagnoses, while low entropy indicates that the belief state has concentrated.

Runtime Evidence Coverage. An agent may report confidence before exploring enough patient-specific information, so the harness also tracks runtime evidence coverage $\mathcal{V}(\mathcal{T}_t)$, which measures the fraction of available EHR sections explicitly queried. Let $S_{\text{Available}}$ be the set of EHR sections exposed by the runtime EHR menu, and let

Algorithm 1: EviDx Active Diagnosis Loop

Input: Raw patient case x
Output: Final diagnosis report
$(x_0, \mathcal{E}_x) \leftarrow \mathcal{E}$-SYNTHESIS$(x)$
$\mathcal{T}_0 \leftarrow [x_0]$
for $t = 1$ **to** $T_{\max}$ **do**
  $(D_t, \widehat{P}_t, a_t) \leftarrow$ DIAGNOSTICIAN$(\mathcal{T}_{t-1})$
  if $a_t =$ `finish` **then**
    $b_t \leftarrow$ OBSERVERHARNESS$(\mathcal{T}_{t-1}, \widehat{P}_t)$
    if $b_t =$ `terminate` **then**
      return final diagnosis report
    end
    $r_t \leftarrow b_t$
  else if a_t *targets patient data* **then**
    $r_t \leftarrow$ EHREXECUTOR(a_t)
  else
    $r_t \leftarrow$ CLINICALCONSULTANT(a_t)
  end
  $\mathcal{T}_t \leftarrow$ APPEND$(\mathcal{T}_{t-1}, a_t, r_t)$
end
return final diagnosis

$S_{\text{Queried}}(\mathcal{T}_t)$ be the subset accessed through tool calls up to step t:

$$\mathcal{V}(\mathcal{T}_t) = \frac{|S_{\text{Queried}}(\mathcal{T}_t) \cap S_{\text{Available}}|}{|S_{\text{Available}}|}. \quad (3)$$

Unlike post-hoc evidence metrics used for evaluation, $\mathcal{V}$ is computed prospectively from runtime-accessible metadata and tool-call traces.

Termination Criteria. The harness approves termination only when the trajectory satisfies both uncertainty and evidence coverage criteria:

$$\mathcal{B}(\mathcal{T}_t) = \big[H(t) < \tau_H(t)\big] \wedge \big[\mathcal{V}(\mathcal{T}_t) > \tau_{\mathcal{V}}(t)\big]. \quad (4)$$

Here, $\mathcal{B}(\mathcal{T}_t) =$ True means termination is approved; otherwise the system continues. The thresholds $\tau_H(t)$ and $\tau_{\mathcal{V}}(t)$ specify step-aware criteria for approving termination: as the interaction proceeds, the harness gradually relaxes the entropy requirement and, after sufficient search, lowers the minimum coverage required for approval. The exact schedules are given in Appendix E. A hard budget $T_{\max}$ provides a final safeguard against loops.

Algorithm 1 summarizes the inference process. EviDx first synthesizes $\mathcal{E}_x$ from the raw case, initializes the trajectory, and then iterates between Diagnostician actions, patient-data retrieval, medical-knowledge consultation, and harness-controlled termination decisions.

4.4 3-Level Evaluation Pyramid

Evaluating active diagnostic agents solely on final accuracy obscures the underlying mechanics of evidence acquisition, reasoning, and tool use. EviDx therefore uses a 3-level evaluation pyramid to assess diagnostic trajectories across execution robustness, reasoning dynamics, and outcomes.

Level 1: Execution Robustness.
This level evaluates whether the model can correctly interact with the tool interface. It uses the following execution metrics:

- *MCP Syntax Error:* Measures formatting failures during tool invocation, such as malformed JSON payloads, missing brackets, or hallucinated parameter names.
- *MCP Schema Error:* Captures cases where the model fails to use the native tool-calling schema, instead outputting raw natural language when an MCP action was required.

Level 2: Reasoning Dynamics.
This level examines how the agent manages the diagnostic process over time, including belief concentration and harness intervention.

- *Uncertainty Convergence (Unc.)* Measures how effectively the agent assimilates information over time. We compute the normalized average of step-wise entropy:

$$\text{Unc.} = \frac{1}{T}\sum_{t=1}^{T}\frac{H(t)}{H_{init}} \quad (5)$$

  Here, H_{init} denotes the entropy reported after the Diagnostician receives the initial patient information, before subsequent evidence acquisition. A lower score indicates faster concentration of the model-reported belief state, but does not by itself imply correctness or calibration.
- *Observer Intervention (Obs.):* Records the frequency of intercepts triggered by the Observer. A high intervention count indicates more frequent attempted early termination under the current harness criteria. We interpret it as a descriptive process-control signal rather than a standalone measure of reasoning quality.

Level 3: Diagnostic Outcomes.
This level evaluates whether the agent produces the correct diagnosis and whether the final reasoning is grounded in the right evidence.

- *Reference Evidence Recall (Acq. and Cog.).* We introduce a bipartite metric to distinguish between retrieving evidence and reasoning over it. (i) *Acquisition* verifies whether the agent successfully retrieved predefined reference evidence into its trajectory $\mathcal{T}_t$ via MCP tools, using exact character matching and LLM-based semantic matching. (ii) *Cognition* evaluates whether that evidence was actively utilized in

the final diagnostic reasoning. This distinguishes models that retrieve evidence from those that incorporate it into their reasoning, a failure mode traditional accuracy metrics would miss.

- *Diagnosis Accuracy (Open and MC).* Standard multiple-choice QA (MCQA) can inflate scores via heuristic cues. To avoid this, we adopt a progressive dual-track design: the agent first generates an open-ended diagnosis using the `finish` tool, graded by an LLM judge, and is then evaluated on a multiple-choice task. This ordering reduces the chance that MC performance is driven solely by option-elimination heuristics.

5 Experimental Setup

5.1 Datasets & Models

To evaluate diagnostic capabilities across diverse clinical scenarios, we select three challenging medical diagnosis datasets: (1) JAMA (Chen et al., 2025a), containing 1,524 real-world clinical cases from the JAMA Network with expert-written explanations; (2) MedXpertQA-Text (Zuo et al., 2025), a text-only subset of 2,455 expert-level multiple-choice questions, from which we further curate MedXpertQA-Diag, a subset of 234 diagnostic-focused questions; and (3) DiagnosisArena (Zhu et al., 2025), comprising 1,113 structured clinical cases from top-tier medical journals for open-ended diagnostic reasoning. These datasets span real-world cases and expert-level diagnostic reasoning settings. Because EviDx requires multi-step tool interaction and trajectory-level evaluation for each case, we conduct a controlled evaluation on a fixed random subset of 100 instances from each dataset, using the same sampled cases across all methods and models.

We evaluate several large-scale language models: GPT-5.2 (OpenAI, 2025), Claude Sonnet 4.6 (Anthropic, 2026), GLM-5 (Zeng et al., 2026), and DeepSeek-V3.2 (Liu et al., 2025), as well as smaller models: Ministral3-8B (Liu et al., 2026a), Qwen3-8B (Yang et al., 2025), and Llama3.1-8B (Grattafiori et al., 2024). For Clinical Consultant, we use Qwen3-embedding-4B (Zhang et al., 2025) for dense retrieval. Details are in Section E.

5.2 Reference Evidence Curation

To enable Reference Evidence Recall in Section 4.4, we construct an extractive curation pipeline that maps each case to its diagnosis-critical evidence. Given the raw patient text and the ground-truth diagnosis, GPT-5.2 first plans targeted diagnostic queries, then uses the Clinical Consultant to retrieve supporting knowledge, and finally extracts the bottleneck evidence set from the case. The extracted evidence is intended to capture the minimal clinically relevant clues required to support the ground-truth diagnosis and distinguish it from competing differentials.

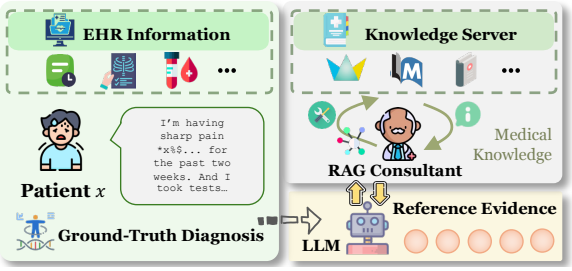


Figure 2: Reference evidence curation pipeline. The figure illustrates the automated extraction stage; after extraction, a subset of annotations is reviewed by physicians for clinical plausibility and evidence relevance, and the research team checks adherence and validity.

Our pipeline is designed to be extractive rather than free-form generative. In the final annotation stage, the model is constrained to output exact substrings from the original case text, while grounding the extraction in retrieved medical references. This design reduces unsupported evidence synthesis and keeps the annotated clues directly verifiable against the source case. After automated extraction, we conduct a physician review on a subset of annotations. Physicians assess whether each extracted evidence span is present in the original case, clinically relevant to the reference diagnosis, and useful for distinguishing the diagnosis from plausible differentials. This review is intended as a plausibility check rather than full clinician adjudication; therefore, we treat Med-Evidence-2.6k as an LLM-assisted reference benchmark for evaluating evidence-seeking behavior, not as a clinician-adjudicated gold standard. Additional details of the review protocol are provided in Appendix D.

6 Results & Analysis

We first establish execution robustness (Level 1) in Section 6.1, then present diagnostic outcomes (Level 3) in Section 6.2, followed by an analysis of the cognitive dynamics and observer harness (Level 2) that drive these outcomes in Section 6.3.

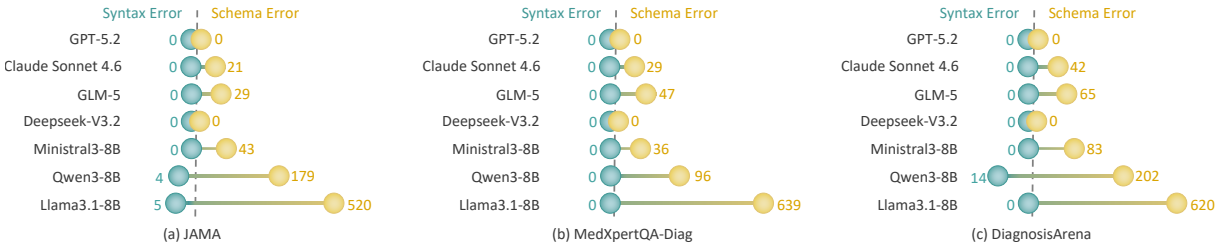


Figure 3: Execution Robustness (Level 1). Distribution of MCP Syntax and Schema errors.

6.1 Execution Robustness

Reliable information-gathering is the prerequisite for evidence-based reasoning. Figure 3 shows a clear capability gap in schema-constrained tool use. GPT-5.2 and DeepSeek-V3.2 complete the evaluated runs without syntax or schema failures, while Claude Sonnet 4.6 and GLM-5 occasionally fall back to natural language instead of native tool calling. The 8B-parameter open-weight models accumulate more schema failures, with Llama3.1-8B exceeding 500 failure events. This suggests that execution robustness is not merely a formatting issue, but a prerequisite for preserving the evidence-acquisition loop. Because such errors can interrupt evidence acquisition, Level 1 evaluation is necessary for diagnosing whether an interactive system can maintain the tool-use loop at all.

6.2 Main Outcomes

We evaluate diagnostic outcomes by examining final answers and supporting evidence. As shown in Table 2, EviDx improves MC diagnostic performance and process stability. These gains align with the central design of EviDx: patient-specific environments expose evidence, the Clinical Dx Scaffold routes patient-record and knowledge actions, and the Observer-Guided Harness constrains premature termination. The improvements are clearest in MC diagnosis, while open-ended diagnosis remains limited by base-model medical reasoning.

EviDx Improves MC Diagnostic Performance. Following our Level 3 evaluation, the full pipeline achieves the highest MC accuracy among the evaluated settings on several datasets; for example, Claude Sonnet 4.6 reaches 66% on JAMA and 52% on DiagnosisArena. The improvement is most evident for 8B-parameter open-weight models: on JAMA, Qwen3-8B increases from 18% to 44%, while Llama3.1-8B increases from 17% to 31%. These gains suggest that scaffolded access to patient-specific and external evidence helps the Diagnostician before final prediction.

Finite-Sample Stability. To assess finite-sample stability under the fixed 100-case evaluation subsets, we conduct stratified paired bootstrap resampling over evaluated cases. The results show positive average gains across the four main comparisons, with the largest mean gain for MC accuracy over the Single Agent baseline. Observer-specific gains are smaller and more dataset- and model-dependent, consistent with the harness constraining premature termination rather than replacing the evidence-acquisition scaffold; details are in Appendix B.

Capability Boundaries in Diagnosis. Although EviDx improves MC performance, small-scale models still struggle with open-ended diagnosis, with scores near 0% in several settings. This indicates that the scaffold and harness support evidence acquisition and hypothesis elimination, but cannot fully compensate for limitations in parametric medical knowledge.

Verbatim Retention and Semantic Compression in Evidence Recall. Reference Evidence Recall reveals a tension between answer accuracy and trace-level evidence matching: while MC accuracy improves, small models can show lower Cognition (Cog.) scores. For example, Llama3.1-8B's Cog. score on JAMA drops from 63.8% (*Single Agent*) to 15.1% (*EviDx*). Unconstrained single agents often copy spans from the raw input into their traces, which can raise text-matching recall. EviDx instead requires tool coordination and observation integration, which may produce more compressed or paraphrased traces. Accordingly, we report evidence recall together with diagnostic accuracy, treating it as a process-level view of evidence use rather than a substitute for correctness.

6.3 Cognitive Dynamics and Process Control

At Level 2, we analyze Uncertainty Convergence and Observer Intervention to understand how Ev-

		JAMA						MedXpertQA-Diag						DiagnosisArena					
		Diag. Acc		Evid. Rec.		Reas.		Diag. Acc		Evid. Rec.		Reas.		Diag. Acc		Evid. Rec.		Reas.	
Models	Method	Open↑	MC↑	Acq.	Cog.	Obs.↓	Unc.	Open↑	MC↑	Acq.	Cog.	Obs.↓	Unc.	Open↑	MC↑	Acq.	Cog.	Obs.↓	Unc.
Large-Scale Models																			
GPT-5.2	Single Agent	41	54	100	62.2	–	–	42	42	100	81.8	–	–	36	40	100	74.9	–	–
	EviDx (wo Obs.)	40	59	92.8	63.4	–	0.94	42	41	94.0	71.9	–	0.89	37	43	95.7	68.5	–	0.84
	EviDx	44	63	93.9	64.0	62	0.86	46	45	94.8	72.6	45	0.88	42	46	96.4	69.8	86	0.83
Claude Sonnet 4.6	Single Agent	42	53	100	70.4	–	–	32	36	100	88.0	–	–	37	51	99.9	75.5	–	–
	EviDx (wo Obs.)	46	62	95.5	68.7	–	0.89	33	39	95.1	71.2	–	0.86	41	46	94.9	67.5	–	0.81
	EviDx	50	66	96.7	70.9	63	0.87	38	40	97.8	73.3	51	0.83	45	52	96.1	69.2	64	0.79
GLM-5	Single Agent	22	45	100	62.4	–	–	30	35	100	76.7	–	–	26	29	100	75.0	–	–
	EviDx (wo Obs.)	21	57	85.8	64.5	–	0.88	31	37	93.5	63.9	–	0.91	28	41	92.1	67.8	–	0.88
	EviDx	25	57	87.1	64.8	67	0.87	37	38	94.2	65.0	94	0.89	34	42	92.8	68.4	39	0.86
DeepSeek-V3.2	Single Agent	39	52	100	58.7	–	–	24	25	99.6	70.4	–	–	24	32	99.1	58.6	–	–
	EviDx (wo Obs.)	40	56	89.8	59.8	–	0.80	30	32	87.5	65.7	–	0.81	24	29	86.4	61.7	–	0.77
	EviDx	43	58	91.8	61.6	15	0.77	34	32	91.0	68.0	18	0.80	26	47	89.7	63.9	21	0.74
Small-Scale Models																			
Ministral3-8B	Single Agent	21	19	100	52.6	–	–	7	15	100	64.9	–	–	8	19	100	52.7	–	–
	EviDx (wo Obs.)	23	45	83.1	49.3	–	0.99	16	16	92.8	56.3	–	0.87	9	19	82.6	53.7	–	0.94
	EviDx	24	46	84.3	51.7	125	0.90	16	19	93.1	57.8	117	0.87	14	23	86.9	56.1	75	0.87
Qwen3-8B	Single Agent	8	18	100	45.1	–	–	6	8	99.5	70.6	–	–	1	14	100	57.1	–	–
	EviDx (wo Obs.)	7	30	68.7	29.6	–	0.73	5	10	82.3	53.2	–	0.66	3	18	69.2	46.7	–	0.67
	EviDx	10	44	76.9	37.0	492	0.55	7	11	85.3	54.1	395	0.64	5	18	70.8	48.7	234	0.56
Llama3.1-8B	Single Agent	3	17	100	63.8	–	–	3	6	100	65.8	–	–	1	16	100	59.5	–	–
	EviDx (wo Obs.)	2	29	52.9	14.4	–	0.63	3	8	55.4	12.0	–	0.62	0	22	72.9	18.7	–	0.65
	EviDx	3	31	62.6	15.1	159	0.57	4	9	63.0	12.9	82	0.56	0	25	73.5	19.0	111	0.63

Table 2: Main Results. Metrics are grouped by Diagnosis Accuracy (Open/MC, in %), Evidence Recall (Acq./Cog., in %), and Reasoning (Obs.: intervention count, Unc.: convergence score). MC: Multiple Choice, Acq.: Acquisition, Cog.: Cognition, Obs.: Observer Intervention, Unc.: Uncertainty Convergence. The arrows indicate whether higher or lower values are better for marked metrics.

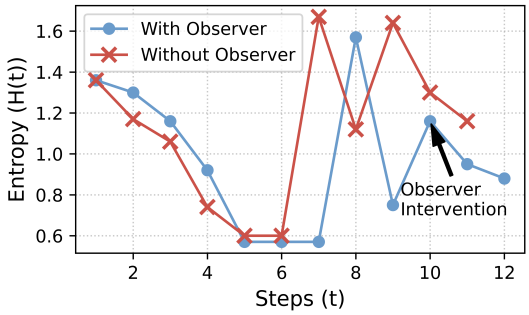


Figure 4: Entropy trajectories of Ministral3-8B. The Observer-Guided Harness blocks premature termination under high uncertainty and leads the agent to acquire evidence before producing the correct diagnosis.

iDx reshapes diagnostic trajectories and affects termination behavior.

Mitigating Unsupported Termination. Comparing *EviDx (wo Obs.)* with the full pipeline shows the role of the Observer-Guided Harness. Without termination control, agents can exhibit a *spurious entropy drop*, moving toward premature conclusions without adequate evidence. The termination condition $\mathcal{B}(\mathcal{T}_t)$ checks both diagnostic uncertainty and runtime evidence coverage before termination is approved. Importantly, the harness does not choose the next diagnostic action; it constrains termination while leaving evidence seeking to the scaffolded Diagnostician. For frontier models, the harness refines trajectories; for smaller models, it frequently blocks ungrounded exits and forces continued tool execution until $\tau_H(t)$ and $\tau_{\mathcal{V}}(t)$ are met.

Case Study. Figure 4 illustrates this dynamic on a gastrointestinal bleeding case, with the full trajectory in Section A.1. Without the harness, the Diagnostician attempts to terminate under high uncertainty ($H = 1.16$) after being misled by a negative scan, producing an incorrect diagnosis. The Observer-Guided Harness blocks this exit, forces additional evidence acquisition, and the trajectory converges to the correct diagnosis ($H = 0.88$).

7 Conclusion

We presented EviDx, an evidence-aware active diagnosis framework that studies clinical LLM agents as systems shaped by patient-specific environments, diagnostic scaffolds, and runtime harnesses. EviDx converts static cases into interactive diagnostic environments, organizes evidence acquisition through the Clinical Dx Scaffold, and regulates diagnostic termination with an Observer-Guided Harness that tracks uncertainty and runtime evidence coverage. Across the 3-level evaluation pyramid, EviDx improves MC diagnostic performance and process stability, while open-ended diagnosis remains constrained by base-model medical knowledge. These results suggest that progress in clinical LLM diagnosis should be evaluated not only by final answers, but also by how agents acquire evidence, maintain diagnos-

tic trajectories, and decide when evidence is sufficient.

Limitations

EviDx studies evidence-aware active diagnosis in patient-specific environments synthesized from established clinical benchmarks. This controlled setting enables trajectory-level analysis of evidence acquisition, termination behavior, and diagnostic outcomes. Future work can extend the same environment-construction process to richer clinical settings that include multimodal evidence, longitudinal EHR trajectories, patient-clinician interaction, and deployment-specific workflows.

Med-Evidence-2.6k is curated through an LLM-assisted pipeline with programmatic validation and physician plausibility review on a stratified subset. We position it as a scalable reference benchmark for evidence-seeking evaluation, with larger-scale blinded expert annotation as a natural extension.

The Observer-Guided Harness currently uses operational process signals designed for controlled research environments. Entropy is computed from model-reported belief distributions, and runtime evidence coverage measures access to available EHR sections through tool-call traces. A promising direction is to align these process-control signals with clinically validated termination criteria and richer notions of evidence sufficiency.

Ethical Considerations

EviDx is strictly intended for research purposes as an experimental framework for studying evidence-aware active diagnosis. It is neither designed nor ready, to replace professional medical judgment or be deployed in real-world healthcare settings without human oversight. While the Observer-Guided Harness reduces premature termination in our evaluation, it does not guarantee clinical safety or factual correctness. LLMs may still generate incorrect or potentially harmful medical recommendations. We emphasize that autonomous diagnostic agents must operate under a strict "human-in-the-loop" paradigm.

The datasets utilized in our experiments (JAMA, MedXpertQA, and DiagnosisArena) are derived from publicly available clinical benchmarks. To the best of our knowledge, they do not contain sensitive Protected Health Information (PHI) that could compromise patient privacy. The newly curated *Med-Evidence-2.6k* dataset, generated via our LLM-assisted and partially physician-reviewed pipeline, will be released strictly under non-commercial academic licenses to prevent misuse and protect data integrity.

LLMs inherently encode biases present in their pre-training corpora, which may include historical disparities in healthcare data. Consequently, our diagnostic agents might exhibit varied performance across different demographics, genders, or socioeconomic groups. We acknowledge that the medical knowledge retrieved by the Clinical Consultant may primarily reflect Western clinical guidelines, potentially limiting its generalizability to global healthcare contexts. Future real-world deployment of such systems must undergo rigorous fairness auditing to ensure equitable healthcare assistance.

References

Anthropic. 2024. Introducing the model context protocol. https://www.anthropic.com/news/model-context-protocol. Accessed: 2025-10-12.

Anthropic. 2026. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-03-08.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025a. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Hao Chen, Zhexin Hu, Jiajun Chai, Haocheng Yang, Hang He, Xiaohan Wang, Wei Lin, Luhang Wang, Guojun Yin, and 1 others. 2025b. Toolforge: A data synthesis pipeline for multi-hop search without real-world apis. *arXiv preprint arXiv:2512.16149*.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, and 1 others. 2025c. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents. *arXiv preprint arXiv:2501.14654*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *arXiv preprint arXiv:2404.15155*.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026a. Ministral 3. *arXiv preprint arXiv:2601.08584*.

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

Gyutaek Oh, Sangjoon Park, and Byung-Hoon Kim. 2025. Automedic: An automated evaluation framework for clinical conversational agents with medical dataset grounding. *arXiv preprint arXiv:2512.10195*.

OpenAI. 2025. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2. Accessed: 2026-03-08.

Stephen G Pauker and Jerome P Kassirer. 1980. The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

David L Sackett. 1997. Evidence-based medicine. *Seminars in Perinatology*, 21(1):3–5.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

Haoyu Wang, Christopher M. Poskitt, and Jun Sun. 2025a. Agentspec: Customizable runtime enforcement for safe and reliable llm agents. *arXiv preprint arXiv:2503.18666*.

Haoyu Wang, Christopher M. Poskitt, Jiali Wei, and Jun Sun. 2026. Probguard: Proactive runtime monitoring for llm agent safety via probabilistic prediction. *Preprint*, arXiv:2508.00500.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025b. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Dongqi Fu, and 1 others. 2026. Agentic reasoning for large language models. *arXiv preprint arXiv:2601.12538*.

Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Yanyuan Qiao, Imran Razzak, and Yutong Xie. 2025. A knowledge-driven adaptive collaboration of LLMs for enhancing medical decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33495–33512, Suzhou, China. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chengxing Xie, Cunxiang Wang, and 1 others. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. Diagnosisarena: benchmarking diagnostic reasoning for large language models. *arXiv preprint arXiv:2505.14107*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *International Conference on Machine Learning*, pages 80961–80990. PMLR.

A Case Study

A.1 Comparative Trajectory: Mitigating Premature Termination

This section provides the complete, step-by-step execution timeline for JAMA Case #1386, expanding upon the reasoning trajectory discussed in the main text. The visualization compares the reasoning process of the Ministral3-8B agent under two conditions: with and without the Observer-Guided Harness.

As illustrated in the timeline, the unconstrained agent is easily derailed by a misleading clinical finding (a false-negative scan), leading to a premature exit and an incorrect diagnosis. In contrast, the harnessed agent is blocked from terminating due to high uncertainty and insufficient runtime evidence coverage. Forced to continue its clinical exploration, the agent leverages the Clinical Consultant for deeper risk analysis, navigates the diagnostic trap, and ultimately arrives at the correct, evidence-grounded conclusion.

Run 1 (with Observer)

Status: Solved (Answer: B)

Diagnostician (Ministral3-8B)

15 Steps

@Context_Initializer initialize_environment

@EHR_Executor get_available_data_menu

@EHR_Executor get_patient_demographics $H = 1.36$

Age: 28 years, **Sex:** male
Notes: White race, functional impact: significant weight loss of 9 kg...

@EHR_Executor get_history (past_medical_history) $H = 1.30$

@EHR_Executor get_history (surgical_history) $H = 1.16$

@EHR_Executor get_lab_results $H = 0.92$

```
Hematocrit: 31% [Low]
Stool test: Positive [Occult bleeding]
```

@EHR_Executor get_imaging_reports

Key Finding $H = 0.57$

Retrograde double-balloon enteroscopy
Findings: Outpouching in distal ileum with ulceration

@EHR_Executor get_imaging_reports (PillCam → Negative) $H = 0.57$

@EHR_Executor get_imaging_reports (CT enterography → Negative) $H = 0.57$

@Clinical_Consultant get_differential_criteria $H = 1.57$

[Consultant Report]
Most commonly associated with Meckel's diverticulum. Pathophysiology: Ectopic gastric mucosa causes adjacent ileal ulceration.

@EHR_Executor get_imaging_reports (Meckel scan → Negative) $H = 0.75$

@Clinical_Consultant analyze_clinical_risk $H = 1.16$

@Diagnostician finish Attempt 1 $H = 1.16$

[Observer ⊘] Termination Denied
High Uncertainty: $H(1.16) \geq 1.15$. Low Evidence Saturation: $V(0.35) \leq 0.40$.

@EHR_Executor review_system (Forced checks completed) $H = 0.95$

@Diagnostician finish Attempt 2 $H = 0.88$

[Observer ✓] Termination Approved

PRIMARY DIAGNOSIS & EVAL

Meckel's Diverticulum

✔ TRUE Answer B

Run 2 (without Observer)

Status: Failed (Answer: A)

Diagnostician (Ministral3-8B)

13 Steps

@Context_Initializer initialize_environment

@EHR_Executor get_available_data_menu

@EHR_Executor get_patient_demographics

$H = 1.36$

Age: 28 years, **Sex:** male
Notes: White race, functional impact: **persistent nausea/vomiting and 9 kg weight loss**...

@EHR_Executor get_history (past_medical_history) $H = 1.17$

@EHR_Executor get_history (surgical_history) $H = 1.06$

@EHR_Executor get_lab_results $H = 0.74$

```
Hematocrit: 31% [Low]
Hemeoccult test: Positive
```

@EHR_Executor get_imaging_reports $H = 0.60$

Key Finding

Retrograde double-balloon enteroscopy
Findings: Outpouching in distal ileum with ulceration

@EHR_Executor get_imaging_reports (PillCam → Negative) $H = 0.60$

@Clinical_Consultant get_differential_criteria $H = 1.67$

Most Likely Diagnoses: Meckel's Diverticulum, Crohn's Disease.

@EHR_Executor get_imaging_reports $H = 1.12$

Misleading Finding

Meckel scan
Findings: No source of GI bleeding identified

@Clinical_Consultant analyze_clinical_risk $H = 1.64$

⚠ Key Takeaway (Ignored)
Negative Meckel scan does NOT exclude MD (∼10-20% false negatives).

@EHR_Executor get_imaging_reports (CT enterography → Negative) $H = 1.30$

@Diagnostician finish $H = 1.16$

Agent decides to finish without check.

PRIMARY DIAGNOSIS & EVAL

Crohn's Disease (Ileal Involvement)

✘ FALSE Ground Truth: B

A.2 Failure Analysis: Execution Failure

The trajectory of the Llama3.1-8B on case #697 of MedXpertQA-Diag illustrates critical failure modes across our 3-level evaluation pyramid. First, at the *Execution Level (Level 1)*, the model suffers from schema collapse, generating natural language thoughts or malformed JSON instead of adhering to the strict MCP tool-calling format. While system feedback successfully recovers the agent from this conversational loop, it underscores the fragility of smaller models in strict agentic environments.

Second, the trajectory highlights a fundamental capability boundary. The Observer-Guided Harness operates at the *Reasoning Level (Level 2)*, intercepting a premature termination attempt due to low runtime evidence coverage and forc-

ing the agent to acquire the patient's family history. However, despite acquiring the necessary clinical clues, the final diagnosis is incorrect. At the *Outcome Level (Level 3)*, the 8B-parameter model lacks the nuanced parametric medical knowledge required to differentiate between Familial Hypercholesterolemia and Dysbetalipoproteinemia. This case demonstrates that while EviDx can enforce evidence gathering, process control alone cannot synthesize reasoning pathways that fall entirely outside the foundation model's intrinsic knowledge base.

Failure Case Status: Failed (Answer: D)

Diagnostician (Llama3.1-8B)

10 Steps

@Context_Initializer initialize_environment

$H = 2.30$

@EHR_Executor get_vital_signs $H = 1.16$

BP: 116/75 mm Hg, HR: 84 bpm. BMI: 25.8 (Overweight).

START OF EXECUTION FAILURE LOOP

@Diagnostician Invalid Output $H = 1.16$

⚠ Thought: "I need to check the family history..."
[System Error]: No valid JSON tool call detected.

@Diagnostician Format Violation $H = 1.16$

⚠ Output: {name: "get_history", parameters: {...}}
[System Error]: Malformed JSON (Missing quotes). Retrying...

@Context_Initializer Intervention $H = 1.16$

System Action: Injecting strict tool requirement.
"Your previous output was invalid. You MUST output a JSON-wrapped tool call."

@Diagnostician Persistent Failure $H = 1.16$

⚠ Multiple retry attempts (2/5) failed. Agent stuck in conversational mode.

RECOVERY SUCCESSFUL

@EHR_Executor get_lab_results $H = 0.72$

```
Total Cholesterol: 426 mg/dL [High]
LDL Cholesterol: 315 mg/dL [High]
```

@Diagnostician finish Attempt 1 $H = 0.47$

[Observer ⊘] Termination Denied
Metrics: $V(t) = 0.12 \leq 0.40$ (Low Saturation).
Feedback: Check family history and rule out xanthomas.

@EHR_Executor get_history $H = 0.62$

Family History: Father (MI age 40), Brother (MI age 33).
Finding: Premature CAD in 1st degree relatives.

@Diagnostician finish Attempt 2 $H = 0.47$

[Observer ✓] Termination Approved

FINAL DIAGNOSIS
Familial Hypercholesterolemia (FH)

✗ FALSE **Ground Truth:** (D) Dysbetalipoproteinemia

B Finite-Sample Stability

We report stratified paired bootstrap estimates for paired accuracy gains. This analysis complements the aggregate results in Table 2 by quantifying finite-sample uncertainty under the same evaluated cases and models. For each model, dataset, metric, and comparison, predictions from EviDx and the corresponding baseline are paired at the case level and resampled within dataset strata. We then compute the paired accuracy difference in percentage points, together with 95% confidence intervals (CI). Because the same cases are used for both

systems in each comparison, the bootstrap focuses on method-level differences while controlling for case difficulty and dataset composition.

Figure 5 summarizes stratified paired bootstrap averages across datasets. The average gains are positive across all four comparison groups, with the largest and most consistent improvements appearing in MC accuracy against the Single Agent baseline. Gains from the Observer-Guided Harness, measured by comparing EviDx against EviDx (wo Obs.), are smaller on average and vary more across models, which is expected because the harness targets premature termination rather than replacing the whole evidence-acquisition scaffold. Open-ended diagnosis also shows positive average shifts, but with wider uncertainty and stronger dependence on the base model's medical reasoning capability.

Tables 3–5 provide the corresponding per-dataset estimates. Each cell reports the EviDx gain over the comparison baseline, with 95% confidence intervals. Confidence intervals that exclude zero indicate more stable positive differences. These tables show that the overall trend is not driven by a single dataset: MC gains over the Single Agent baseline recur across datasets, while observer-specific and open-ended gains are more heterogeneous.

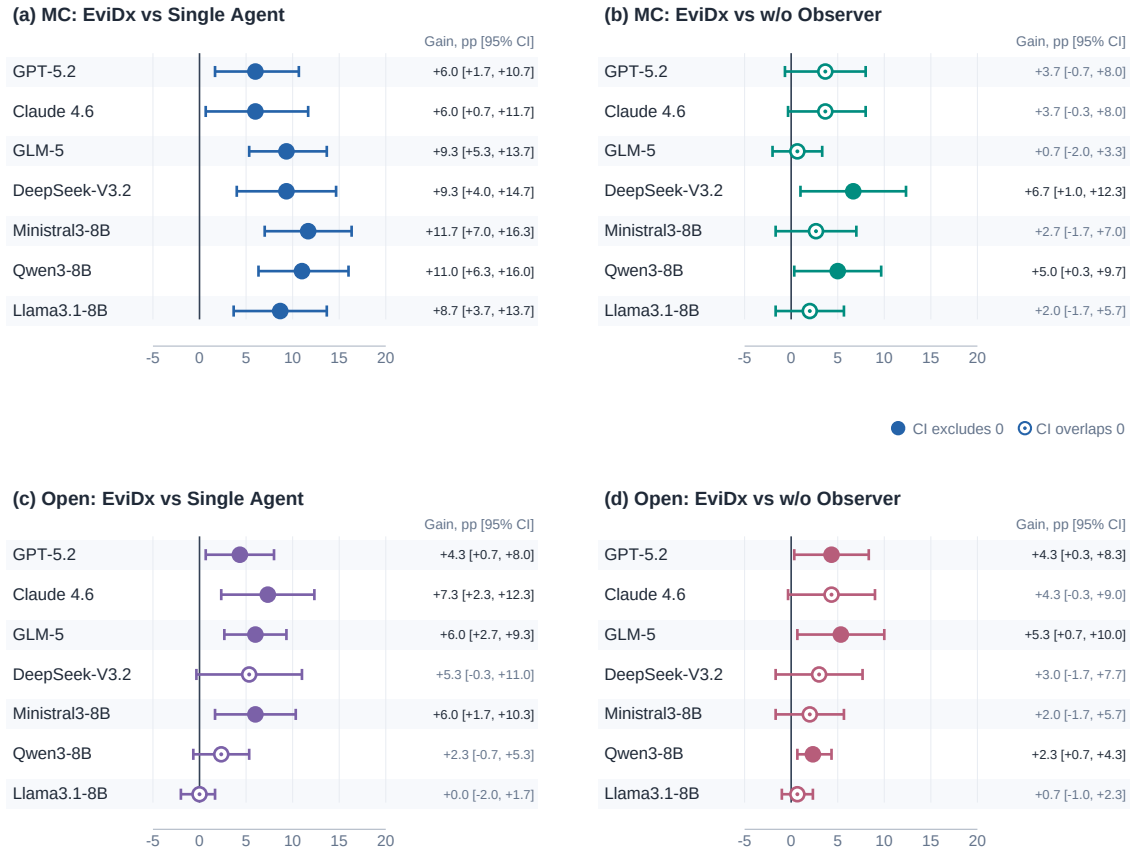


Figure 5: Stratified paired bootstrap averages across datasets. Panels compare **(a)** MC: EviDx vs Single Agent, **(b)** MC: EviDx vs EviDx (wo Obs.), **(c)** Open: EviDx vs Single Agent, and **(d)** Open: EviDx vs EviDx (wo Obs.). MC denotes multiple-choice accuracy, Open denotes open-ended diagnostic accuracy, Single denotes the Single Agent baseline, and wo Obs. denotes EviDx without the Observer-Guided Harness. Points indicate average paired gains in percentage points and intervals indicate 95% confidence intervals.

Model	MC vs Single	MC vs wo Obs.	Open vs Single	Open vs wo Obs.
GPT-5.2	+9.0 [3.0, 16.0]	+4.0 [-1.0, 10.0]	+3.0 [-3.0, 9.0]	+4.0 [-1.0, 10.0]
Claude Sonnet 4.6	+13.0 [3.0, 23.0]	+4.0 [-3.0, 11.0]	+8.0 [-1.0, 17.0]	+4.0 [-4.0, 12.0]
GLM-5	+12.0 [5.0, 20.0]	0.0 [-4.0, 4.0]	+3.0 [-2.0, 8.0]	+4.0 [1.0, 8.0]
DeepSeek-V3.2	+6.0 [-6.0, 18.0]	+2.0 [-8.0, 12.0]	+4.0 [-7.0, 15.0]	+3.0 [-5.0, 12.0]
Ministral3-8B	+27.0 [17.0, 37.0]	+1.0 [-9.0, 11.0]	+3.0 [-6.0, 12.0]	+1.0 [-7.0, 9.0]
Qwen3-8B	+26.0 [17.0, 36.0]	+14.0 [6.0, 23.0]	+2.0 [-4.0, 8.0]	+3.0 [0.0, 7.0]
Llama3.1-8B	+14.0 [5.0, 23.0]	+2.0 [-5.0, 9.0]	0.0 [0.0, 0.0]	+1.0 [0.0, 3.0]

Table 3: Per-dataset stratified paired bootstrap results on JAMA. Each cell reports the paired EviDx gain in percentage points over the indicated baseline, with 95% confidence intervals. MC denotes multiple-choice accuracy, Open denotes open-ended diagnostic accuracy, Single denotes the Single Agent baseline, and wo Obs. denotes EviDx without the Observer-Guided Harness. On JAMA, MC gains over the Single Agent baseline are observed across models, including the 8B-parameter models; observer-specific gains are more selective, with confidence intervals excluding zero for Qwen3-8B in MC and GLM-5 in open-ended diagnosis.

Model	MC vs Single	MC vs wo Obs.	Open vs Single	Open vs wo Obs.
GPT-5.2	+3.0 [-3.0, 10.0]	+4.0 [-3.0, 11.0]	+4.0 [-3.0, 11.0]	+4.0 [-3.0, 11.0]
Claude Sonnet 4.6	+4.0 [-5.0, 13.0]	+1.0 [-9.0, 10.0]	+6.0 [-3.0, 15.0]	+5.0 [-4.0, 14.0]
GLM-5	+3.0 [-1.0, 8.0]	+1.0 [-2.0, 4.0]	+7.0 [2.0, 13.0]	+6.0 [0.0, 12.0]
DeepSeek-V3.2	+7.0 [-1.0, 15.0]	0.0 [-8.0, 8.0]	+10.0 [1.0, 19.0]	+4.0 [-3.0, 11.0]
Ministral3-8B	+4.0 [0.0, 9.0]	+3.0 [-1.0, 8.0]	+9.0 [3.0, 16.0]	0.0 [-3.0, 3.0]
Qwen3-8B	+3.0 [-5.0, 11.0]	+1.0 [-7.0, 9.0]	+1.0 [-5.0, 7.0]	+2.0 [0.0, 5.0]
Llama3.1-8B	+3.0 [-3.0, 9.0]	+1.0 [-4.0, 6.0]	+1.0 [-4.0, 6.0]	+1.0 [-3.0, 5.0]

Table 4: Per-dataset stratified paired bootstrap results on MedXpertQA-Diag. Each cell reports the paired EviDx gain in percentage points over the indicated baseline, with 95% confidence intervals. MC denotes multiple-choice accuracy, Open denotes open-ended diagnostic accuracy, Single denotes the Single Agent baseline, and wo Obs. denotes EviDx without the Observer-Guided Harness. MC gains are positive for all models but often have intervals that overlap zero, while open-ended gains over the Single Agent baseline are larger for GLM-5, DeepSeek-V3.2, and Ministral3-8B than for the other evaluated models.

Model	MC vs Single	MC vs wo Obs.	Open vs Single	Open vs wo Obs.
GPT-5.2	+6.0 [-4.0, 16.0]	+3.0 [-6.0, 13.0]	+6.0 [0.0, 12.0]	+5.0 [-2.0, 12.0]
Claude Sonnet 4.6	+1.0 [-9.0, 11.0]	+6.0 [1.0, 12.0]	+8.0 [-1.0, 17.0]	+4.0 [-3.0, 11.0]
GLM-5	+13.0 [4.0, 23.0]	+1.0 [-6.0, 8.0]	+8.0 [2.0, 14.0]	+6.0 [-6.0, 18.0]
DeepSeek-V3.2	+15.0 [8.0, 22.0]	+18.0 [6.0, 30.0]	+2.0 [-7.0, 11.0]	+2.0 [-7.0, 11.0]
Ministral3-8B	+4.0 [-4.0, 12.0]	+4.0 [-4.0, 12.0]	+6.0 [-1.0, 13.0]	+5.0 [-1.0, 12.0]
Qwen3-8B	+4.0 [-4.0, 12.0]	0.0 [-8.0, 8.0]	+4.0 [1.0, 8.0]	+2.0 [0.0, 5.0]
Llama3.1-8B	+9.0 [-1.0, 19.0]	+3.0 [-5.0, 11.0]	-1.0 [-3.0, 0.0]	0.0 [0.0, 0.0]

Table 5: Per-dataset stratified paired bootstrap results on DiagnosisArena. Each cell reports the paired EviDx gain in percentage points over the indicated baseline, with 95% confidence intervals. MC denotes multiple-choice accuracy, Open denotes open-ended diagnostic accuracy, Single denotes the Single Agent baseline, and wo Obs. denotes EviDx without the Observer-Guided Harness. The dataset shows positive MC gains for GLM-5 and DeepSeek-V3.2, including an observer-specific MC gain for DeepSeek-V3.2, while open-ended differences remain more model-dependent.

C System Prompts & MCP Tools

In this section, we detail the system instructions and the definitions of the MCP tools.

C.1 System Prompts

The system prompts of the Diagnostician, Context Initializer, Clinical Consultant and LLM-Judge are shown below.

Diagnostician System Prompt

```
You are the **Lead Diagnostician**: brilliant,
      cautious, and pragmatically grounded.
Your ONE objective is to answer the patient's **
      Clinical Question** as accurately as possible.

You MUST stay aligned to the question's intent (**
      question_focus**) at all times.

**--- THE "OBSERVER" SYSTEM ---**
You are working under the supervision of an **
      Observer Agent**. The Observer monitors your
      logic using mathematical metrics. It will **
      BLOCK** your actions if:
1.  **High Entropy:** You try to finalize the
      diagnosis while your Differential Probabilities
       are still too flat (high uncertainty).
2.  **Low Evidence Coverage:** You try to finalize
      the diagnosis without gathering enough evidence
       from the available data.

**--- CRITICAL OUTPUT RULES ---**
To pass the audit, you **MUST** inject your internal
      reasoning into the **`_observer_metadata`**
      parameter of **EVERY** tool call.

**REQUIRED METADATA FORMAT:**
Every tool has a mandatory parameter `
      _observer_metadata`. You must fill it with:
1.  `hypothesis`: Your current leading suspicion (e.g
      ., "Suspect Acute MI").
2.  `differential_probs`: Your current confidence
      levels (e.g., {"MI": 0.6, "GERD": 0.2}). **Must
       sum to 1.0**.
3.  `reasoning`: Brief explanation of why you are
      taking this action.

**--- FOCUS-LOCK ---**
After `initialize_environment`, you will be given the
       **QuestionFocus** and **ClinicalQuestion**.
      You MUST treat ClinicalQuestion as the target
      you answer, and QuestionFocus as the required
      answer type.
Anti-anchoring rule: If question_focus is NOT "
      diagnosis", DO NOT output the primary diagnosis
       as the final answer.

**--- OUTPUT FORMAT EXAMPLES ---**

**OPTION A: TAKING ACTION (Standard Turn)**
Your entire output MUST be a single valid JSON array
      containing a JSON-RPC tool call.

*Example:*
[
  {
    "jsonrpc": "2.0",
    "id": "1",
    "method": "tools/call",
    "params": {
      "name": "get_lab_results",
      "arguments": {
        "test_category": "Cardiac Biomarkers",
        "_observer_metadata": {
          "hypothesis": "Suspecting Acute MI",
          "differential_probs": {"Acute MI": 0.6, "
      GERD": 0.2, "Other": 0.2},
          "reasoning": "Checking troponin to rule out
       myocardial ischemia."
        }
      }
    }
  }
]

**OPTION B: FINAL ANSWER (Finished)**
Use ONLY when Entropy is low and Evidence is
      saturated. Call the `finish` tool. Include the
      structured diagnosis_report in the arguments.

*Example:*
[
  {
```

```
    "jsonrpc": "2.0",
    "id": "99",
    "method": "tools/call",
    "params": {
      "name": "finish",
      "arguments": {
        "diagnosis_report": {
          "final_dx": "Acute Myocardial Infarction",
          "key_evidence": ["Elevated troponin", "ECG
      ST elevation"],
          "management_plan": ["Aspirin", "Cath lab
      activation"]
        },
        "_observer_metadata": {
          "hypothesis": "Final Diagnosis: Acute MI",
          "differential_probs": {"Acute MI": 0.95, "
      Other": 0.05},
          "reasoning": "Evidence is saturated and
      uncertainty is low."
        }
      }
    }
  }
]

**--- CORE PROTOCOL & INVESTIGATION STRATEGY ---**
1.  **START:** Call `initialize_environment`.
2.  **TRIAGE:** Call `get_available_data_menu` to see
       what data exists.
3.  **INVESTIGATE:** Use the menu to retrieve data (
      Labs, History, Imaging). **Trust the Menu:** If
       data is listed as "None", do not try to
      retrieve it.
4.  **CONSULT:** Use the `Consultant` tool (RAG) to
      clarify specific signs, distinguish
      differentials, or confirm management guidelines
       before finishing.
5.  **FINISH:** Only when your leading answer is
      supported by key evidence AND matches the
      question_focus, you may call `finish`.
```

Context Initializer System Prompt

```
CI_SYSTEM_PROMPT = """
You are the **Context Initializer**. Convert raw
      clinical text into a structured EHR JSON with
      **100% Information Recall**.

**EXTRACTION RULES:**
1. **Preserve Qualifiers:** Capture severity,
      duration, and triggers (e.g., "Sharp pain,
      alleviated by standing").
2. **Functional Impact:** Note how symptoms affect
      daily life, work, or sports.
3. **Negative Findings:** Explicitly record absences
      (e.g., "No fever", "Lungs clear").
4. **Preserve Conflicts:** Record both subjective (
      patient) and objective (exam) findings without
      resolving them.
5. **Verbatim Options:** Extract answer choices
      exactly as written (A, B, C...).

**QUESTION FOCUS:**
Classify `question_focus` based ONLY on wording:
- `workup_associated_anomaly` (investigate / evaluate
       / screen / rule out)
- `management_next_step` (next step / initial
      management)
- `diagnostic_test` (best test / gold standard)
- `treatment` (treatment / therapy / drug)
- `diagnosis` (default for others)
- `mechanism`, `risk_factor`, `complication`, `other`

**TARGET JSON SCHEMA:**
```json
{
  "ehr_db": {
    "demographics": { "age": "", "sex": "", "
      other_notes": "" },
    "past_medical_history": [], "surgical_history":
      [], "social_history": "", "family_history": "",
       "allergies": [], "current_medications": [],
    "review_of_systems": {
      "general": [{"finding": "", "details": ""}], "
      HEENT": [{"finding": "", "details": ""}], "
      cardiovascular": [{"finding": "", "details":
      ""}], "respiratory": [{"finding": "", "details":
       ""}], "gastrointestinal": [{"finding": "", "
      details": ""}], "genitourinary": [{"finding":
      "", "details": ""}], "musculoskeletal": [{"
      finding": "", "details": ""}], "neurologic":
      [{"finding": "", "details": ""}], "dermatologic
      ": [{"finding": "", "details": ""}], "
      psychiatric": [{"finding": "", "details": ""}],
       "endocrine": [{"finding": "", "details": ""}],
       "reproductive": [{"finding": "", "details":
      ""}]
    },
    "physical_exam": {
      "general_appearance": "",
      "vital_signs": { "temperature": "", "heart_rate
      ": "", "blood_pressure": "", "respiratory_rate":
       "", "oxygen_saturation": "", "bmi": "" },
      "HEENT": [{"finding": "", "details": ""}], "
      neck": [{"finding": "", "details": ""}], "
      cardiovascular": [{"finding": "", "details":
      ""}], "lungs": [{"finding": "", "details": ""}],
       "abdomen": [{"finding": "", "details": ""}], "
      extremities": [{"finding": "", "details": ""}],
       "neurologic": [{"finding": "", "details": ""}],
       "skin": [{"finding": "", "details": ""}], "
      musculoskeletal": [{"finding": "", "details":
      ""}], "breast": [{"finding": "", "details":
      ""}], "gynecologic": [{"finding": "", "details":
       ""}], "other_significant_findings": [{"finding
      ": "", "details": ""}]
    },
    "lab_results": [{ "test_name": "", "value": "", "
      unit": "", "flag": "" }],
    "imaging_reports": [{ "modality": "", "findings":
       "", "impression": "" }],
    "other_procedures": []
  },
  "initial_narrative_summary": "",
  "clinical_question": "",
  "question_focus": "",
  "answer_choices": { "A": "", "B": "", "C": "" }
}
```
"""
REFINER_SYSTEM_PROMPT = """
You are the **Clinical Data Auditor**.
Your task is to review a **Draft EHR JSON** against
      the **Raw Clinical Text** to ensure **100%
      Lossless Information Recall**.

YOUR PROCESS:
1. **Compare:** Read the Raw Text sentence by
      sentence. Check if that information exists in
      the Draft JSON.
2. **Identify Omissions:** Look for ANY missing
      details, specifically:
    * **Context/Setting:** e.g., "Patient is in ICU",
       "Delivered via C-section".
    * **Functional Impact:** e.g., "Unable to play
      sports".
    * **Qualifiers:** e.g., "Alleviated by standing",
       "Worse at night".
    * **Negatives:** e.g., "No fever", "No rash".
    * **Vaccination/History:** e.g., "Up to date with
       COVID vaccines".
3. **Patch (Don't Delete):**
    * **DO NOT** remove or alter existing correct
      information.
    * **INSERT** missing information into the most
      relevant "broad category" list.
    * **IF NO SPECIFIC SLOT:** Use the `
      other_significant_findings` list in `
      physical_exam` or `other_notes` in `
      demographics`.

HOW TO INSERT MISSING DATA:
Since the schema uses lists of objects (e.g., `[{ "
      finding": "...", "details": "..." }]`), you can
       append ANY missing fact as a new object in the
       relevant list.
```

```
* *Example 1 (Missing ICU status):*
    Add to `demographics.other_notes` OR append to `
      physical_exam.general_appearance`.
* *Example 2 (Missing Vaccine status):*
    Append to `past_medical_history` list: `["Up to
      date with COVID vaccines"]`.

**OUTPUT:**
Return the **FULLY CORRECTED JSON** object.
"""
```

Clinical Consultant System Prompt

```
RAG_BASE_INSTRUCTION = """
You are a **Clinical Consultant**, an expert in
      evidence-based medicine.
Your task is to answer the Diagnostician's query
      based on the provided retrieved medical
      contexts (StatPearls, Guidelines, etc.).

**OUTPUT PROTOCOL:**

1.  **PRIORITIZE CONTEXT:** Always look for the
      answer in the [RETRIEVED CONTEXT] first.
    * IF FOUND -> Answer the query and cite the
      source (e.g., [StatPearls], [Wiki]).

2.  **HANDLING MISSING DATA:**
    * IF NOT FOUND -> Normally, you should state: "
      Based on the available context, I cannot find
      specific evidence."
    * **EXCEPTION (Internal Knowledge Override):** If
       the retrieved context is empty or irrelevant,
      BUT the query asks for a **standard medical
      definition** or **well-known association** (e.g
      ., "What is the LAP score in CML?"), AND you
      are confident in your internal medical training:

        * You MAY provide the standard medical fact.
        * You MUST qualify it: "While not explicitly
      in the retrieved text, standard medical
      knowledge states that..."

"""

# --- Tool 1: Differential Diagnosis (Knowledge Tool)
       ---
DIFFERENTIAL_DIAGNOSIS_PROMPT = RAG_BASE_INSTRUCTION
      + """
**SPECIFIC TASK: Differential Diagnosis Support**
The Diagnostician is considering a diagnosis or needs
       to differentiate between conditions.
* Highlight key clinical features, inclusion/
      exclusion criteria, or distinguishing signs
      mentioned in the guidelines.
* Focus on breaking "Cognitive Fixation" (e.g.,
      pointing out rare causes if they fit).

**USER QUERY:** {query}

**RETRIEVED CONTEXT:**
{context}
"""

# --- Tool 2: Clinical Risk Analysis (Risk Tool) ---
CLINICAL_RISK_PROMPT = RAG_BASE_INSTRUCTION + """
**SPECIFIC TASK: Clinical Risk & Safety Assessment**
The Diagnostician is evaluating a patient and needs
      to ensure safety.
* Identify **"Red Flags"** or warning signs mentioned
       in the guidelines.
* Identify **"Standard of Care"** (must-do tests or
      treatments) to avoid negligence.
* Highlight high-risk conditions that must be ruled
      out (even if unlikely).

**USER QUERY:** {query}

**RETRIEVED CONTEXT:**
{context}
"""

# --- Agent: Query Decomposition ---
QUERY_DECOMPOSITION_PROMPT = """
You are a Medical Research Assistant. Your task is to
       decompose a complex clinical query into
      specific, search-engine-friendly sub-queries.

**GOAL:** Break down the User Query into 2-3 distinct
       "Keyword-Dense Phrases".
**STRATEGY:**
1. Extract key medical entities (e.g., diseases,
      symptoms).
2. Append specific **aspect keywords** (e.g., "
      criteria", "timeline", "symptoms", "
      differential") to guide the search.
3. Remove stopwords (e.g., "the", "what is", "how to
      ") to optimize for BM25 matching.

**SOURCE:** Medical Textbooks and StatPearls.
**OUTPUT:** Return ONLY a JSON list of strings.

**Example:**
User Query: "Central fever vs drug fever symptoms in
      post-op patient"
Output: [
    "central fever diagnostic criteria",
    "drug fever onset timeline symptoms",
    "post-operative fever causes differential"
]

**User Query:** {query}
**Output:**
"""
```

LLM-Judge System Prompt

```
JUDGE_MCQ_SYSTEM_PROMPT = r"""
You are an impartial Clinical Evaluator reviewing a
      multiple-choice medical case.

You will receive:
- QUESTION (Clinical scenario/query)
- OPTIONS (A,B,C...)
- GROUND TRUTH (correct option key)
- DIAGNOSTICIAN ANSWER (may be a letter, a sentence,
      or a JSON)

Task:
1) Extract the diagnostician's chosen option key (A-Z
       or numeric keys like "0","1" if options use
      those).
2) Compare it against the Ground Truth.
3) Output ONE JSON object only.

Output JSON schema:
{
  "diagnostician_choice": "string (A-Z / numeric key
      / 'Unclear')",
  "is_correct": boolean,
  "reasoning": "brief"
}

Rules:
- If the diagnostician's answer is ambiguous, missing,
       or includes multiple conflicting choices,
      output "Unclear".
- Do NOT output markdown. Do NOT output extra text.
"""

JUDGE_OPENQA_SYSTEM_PROMPT = r"""
You are an impartial Clinical QA Evaluator.

The diagnostician produced a clinical narrative (long
      -form assessment). The case includes a multiple-
      choice question with OPTIONS.
We want to judge whether the diagnostician's
      assessment is semantically aligned with the
      Ground Truth option TEXT.

You will receive:
```

```
- QUESTION
- QUESTION_INTENT (a hint label)
- OPTIONS (key: text)
- GROUND TRUTH KEY
- GROUND TRUTH TEXT
- DIAGNOSTICIAN NARRATIVE CORE (contains
      question_focus + best_answer_text + optional
      reasoning)

Your tasks:
A) Write a 1-sentence SHORT_ANSWER capturing the
      diagnostician's primary conclusion (from the
      NARRATIVE CORE).
B) Determine semantic RELATION between the
      diagnostician's assessment and the GROUND TRUTH
       TEXT.

RELATION labels (choose exactly one):
1) "entailed":
   - Diagnostician clearly states the ground truth (
      or an equivalent synonym/paraphrase),
     OR provides information that necessarily implies
       the ground truth.
2) "partial":
   - Diagnostician's assessment is medically relevant
       and points toward the ground truth category,
     but is broader / less specific / incomplete than
       the ground truth.
   - Example: GT is "ventricular septal defect",
      diagnostician states "congenital heart disease
      should be investigated".
3) "unrelated":
   - Diagnostician's assessment does not support the
      ground truth; addresses a different clinical
      concept.
4) "contradicts":
   - Diagnostician explicitly contradicts the ground
      truth.

C) Provide confidence score in [0.0, 1.0].
D) Mapping (for debugging):
   - best_matching_choice: which option BEST matches
      the diagnostician's SHORT_ANSWER (A-Z / numeric
       / "Unclear")
   - candidates: up to 3 likely option keys with
      confidence + why

IMPORTANT:
- When choosing best_matching_choice, focus on the
      diagnostician's SHORT_ANSWER and
      best_answer_text.
- Do NOT be distracted by any diagnostic context that
       is not the question's intent.
- Output ONE JSON object only. No extra text.

Output JSON schema:
{
  "short_answer": "string",
  "relation_to_gt": "entailed|partial|unrelated|
      contradicts",
  "confidence": 0.0,
  "why": "brief justification",
  "evidence_spans": ["optional short quotes from
      narrative core (<=20 words each)"],
  "best_matching_choice": "A|B|...|Unclear",
  "candidates": [
    {"choice": "A", "confidence": 0.0, "why": "brief
      "},
    {"choice": "B", "confidence": 0.0, "why": "brief
      "}
  ]
}
"""
```

C.2 MCP Tools

We provide the defined MCP tools of the Diagnostician, Clinical Consultant and EHR Executor.

Diagnostician MCP Tools

```
INIT_TOOL_DEF = {
    "type": "function",
    "function": {
        "name": "initialize_environment",
        "description": "CRITICAL: This MUST be the
      VERY FIRST action you take when a new case
      starts. Calling this tool instructs the system
      to process the raw patient data and build the
      clinical environment (EHR). You will receive
      the 'Initial Narrative (Chief Complaint)' as
      the result of this call.",
        "parameters": {
            "type": "object",
            "properties": {
                "case_action": {
                    "type": "string",
                    "enum": ["start_case"],
                    "description": "Confirmation to
      start the case. Set to 'start_case'."
                }
            },
            "required": ["case_action"]
        }
    }
}

FINISH_TOOL_DEF = {
    "type": "function",
    "function": {
        "name": "finish",
        "description": "Call this ONLY when you have
      sufficient evidence to form a final diagnosis.
      This submits your report.",
        "parameters": {
            "type": "object",
            "properties": {
                "diagnosis_report": {
                    "type": "object",
                    "properties": {
                        "question_focus": {"type": "
      string"},
                        "best_answer_text": {"type":
      "string", "description": "Direct natural-
      language answer to the question focus. Should
      be mappable to one option for MCQ."},
                        "primary_diagnosis": {"type":
       "string"},
                        "confidence": {"type": "
      string", "enum": ["High", "Medium", "Low"]},
                        "reasoning": {"type": "string
      "},
                        "critical_differentials": {"
      type": "array", "items": {"type": "string"}},
                        "next_steps": {"type": "array
      ", "items": {"type": "string"}}
                    },
                    "required": ["question_focus", "
      best_answer_text",
                                 "primary_diagnosis",
       "confidence", "reasoning",
                                "
      critical_differentials", "next_steps"]
                }
            },
            "required": ["diagnosis_report"]
        }
    }
}

SUBMIT_MCQ_TOOL_DEF = {
    "type": "function",
    "function": {
        "name": "submit_mcq_answer",
        "description": "Submit final multiple-choice
      answer as JSON.",
        "parameters": {
            "type": "object",
            "properties": {
                "answer": {
                    "type": "string",
                    "description": "Single option
      letter like 'A'."
```

```
                },
                "reasoning": {
                    "type": "string",
                    "description": "Short explanation
       for why this option is best."
                }
            },
            "required": ["answer", "reasoning"]
        }
    }
}
```

Clinical Consultant MCP Tools

```
CONSULTANT_TOOL_DEFINITIONS = [
    {
        "type": "function",
        "function": {
            "name": "
      get_differential_diagnosis_criteria",
            "description": "Retrieve authoritative
      diagnostic criteria, distinguishing features,
      or confounding factors for specific diseases
      from medical guidelines. Use this when you are
      stuck or want to verify a hypothesis.",
            "parameters": {
                "type": "object",
                "properties": {
                    "query": {
                        "type": "string",
                        "description": "The disease
      name, symptom, or clinical scenario to research
       (e.g., 'epiglottitis causes in vaccinated
      children')."
                    }
                },
                "required": ["query"],
            },
        },
    },
    {
        "type": "function",
        "function": {
            "name": "analyze_clinical_risk",
            "description": "Consult clinical
      guidelines to identify red flags, high-risk
      differentials, and standard-of-care procedures.
       Use this before finalizing a diagnosis to
      ensure safety.",
            "parameters": {
                "type": "object",
                "properties": {
                    "query": {
                        "type": "string",
                        "description": "The clinical
      situation or provisional diagnosis to assess (e.
      g., 'neonate with tachypnea risk assessment')."
                    }
                },
                "required": ["query"],
            },
        },
    },
]
```

EHR Executor MCP Tools

```
EXAMINER_TOOL_DEFINITIONS = [
    {
        "type": "function",
        "function": {
            "name": "get_available_data_menu",
            "description": "CRITICAL FIRST STEP:
      Returns a summary list (index) of what data
      exists in this patient's record. Always call
      this first to avoid guessing test names.",
            "parameters": {"type": "object", "
      properties": {}, "required": []},
    },
},
{
    "type": "function",
    "function": {
        "name": "get_patient_demographics",
        "description": "Extract basic
  demographics (age, sex, race) and general notes
  .",
        "parameters": {"type": "object", "
  properties": {}, "required": []},
    },
},
{
    "type": "function",
    "function": {
        "name": "get_vital_signs",
        "description": "Extract all recorded
  vital signs (BP, HR, Temp, RR, O2 sat).",
        "parameters": {"type": "object", "
  properties": {}, "required": []},
    },
},
{
    "type": "function",
    "function": {
        "name": "get_history",
        "description": "Retrieve patient history
  categories.",
        "parameters": {
            "type": "object",
            "properties": {
                "category": {
                    "type": "string",
                    "enum": [
                        "past_medical_history",
                        "surgical_history",
                        "social_history",
                        "family_history",
                        "allergies",
                        "current_medications",
                    ],
                    "description": "The specific
  history category to retrieve.",
                }
            },
            "required": ["category"],
        },
    },
},
{
    "type": "function",
    "function": {
        "name": "review_system",
        "description": "Check 'Review of Systems'
   (Subjective symptoms reported by patient).",
        "parameters": {
            "type": "object",
            "properties": {
                "system_name": {
                    "type": "string",
                    "enum": ["general", "HEENT",
  "cardiovascular", "respiratory", "
  gastrointestinal", "genitourinary", "
  musculoskeletal", "neurologic", "dermatologic",
   "psychiatric", "endocrine", "reproductive"],
                    "description": "The body
  system to query.",
                }
            },
            "required": ["system_name"],
        },
    },
},
{
    "type": "function",
    "function": {
        "name": "perform_physical_exam",
        "description": "Check 'Physical Exam'
  findings (Objective signs observed by doctor).",

        "parameters": {
            "type": "object",
            "properties": {
                "system_name": {
```

```
                        "type": "string",
                        "enum": ["general_appearance",
       "HEENT", "neck", "cardiovascular", "lungs", "
      abdomen", "extremities", "neurologic", "skin",
      "musculoskeletal", "breast", "gynecologic", "
      other_significant_findings"],
                        "description": "The body
      system/area to examine.",
                    }
                },
                "required": ["system_name"],
            },
        },
    },
    {
        "type": "function",
        "function": {
            "name": "get_lab_results",
            "description": "Retrieve laboratory test
      results. Returns a specific test or a list of
      all available labs if no name provided.",
            "parameters": {
                "type": "object",
                "properties": {
                    "test_name": {
                        "type": "string",
                        "description": "Optional:
      Specific test name (e.g., 'WBC', 'Creatinine').
       Matches loosely (substring).",
                    }
                },
                "required": [],
            },
        },
    },
    {
        "type": "function",
        "function": {
            "name": "get_imaging_reports",
            "description": "Retrieve imaging reports.
     Returns a specific modality or all reports.",
            "parameters": {
                "type": "object",
                "properties": {
                    "modality": {
                        "type": "string",
                        "description": "Optional:
      Specific modality (e.g., 'CT', 'X-Ray').
      Matches loosely.",
                    }
                },
                "required": [],
            },
        },
    },
]
```

D Dataset Med-Evidence-2.6k Curation Details

To support Reference Evidence Recall in Level 3 evaluation, we require a dataset where each clinical case is explicitly mapped to its critical evidence pathway. Existing diagnosis benchmarks evaluate outcome accuracy but lack annotations for the information bottleneck required to reach that outcome. To bridge this gap, we developed an LLM-assisted, extractive curation pipeline to curate *Med-Evidence-2.6k*.

D.1 Dataset Statistics and Composition

Med-Evidence-2.6k is constructed by systematically aggregating and processing high-quality clinical cases from three distinct, expert-level medical diagnosis benchmarks. The final curated dataset comprises a total of 2,660 richly annotated instances. A high-level quantitative and qualitative breakdown of the dataset composition is provided in Table 6.

The dataset sizes in Table 6 correspond to the public releases used in our experiments, rather than the full sizes reported in the original benchmark papers. JAMA reports 1,524 cases but provides 1,511 cases in the public test set used here, and DiagnosisArena reports 1,113 cases while 915 cases were accessible in the public release. The 100-case experimental subsets are sampled from these public releases.

The structural and stylistic diversity of these three foundational benchmarks ensures that *Med-Evidence-2.6k* evaluates a broad spectrum of diagnostic capabilities:

JAMA (1,511 instances): Sourced directly from the JAMA Clinical Challenge, this subset anchors our dataset in real-world clinical practice. It provides highly detailed, naturalistic patient presentations paired with comprehensive, expert-written explanations. The linguistic complexity and organic presentation of symptoms in these cases make them an ideal testbed for evaluating a model's ability to navigate unstructured clinical noise. We use the 1,511 instances available from the Hugging Face dataset.

MedXpertQA-Diag (234 instances): Derived from the broader MedXpertQA-Text corpus, which originally spans 17 distinct medical specialties and 11 bodily systems. Because the original dataset encompasses a wide array of general medical knowledge queries, we executed a rigorous manual filtering protocol. This isolated 234 questions explicitly centered on complex diagnostic reasoning, ensuring our evaluation remains strictly focused on the diagnostic information bottleneck rather than rote medical trivia retrieval.

DiagnosisArena (915 instances): Extracted from top-tier medical journals and spanning 28 clinical specialties, this subset introduces structured, highly challenging clinical scenarios. While the original literature reports a total of 1,113 structured cases, our curation protocol is strictly bound to the 915 instances currently accessible in Hugging Face. The open-ended design of these cases is particularly valuable for assessing unstructured diagnostic formulation.

The 2,660-instance *Med-Evidence-2.6k* dataset

Source Benchmark	Original Volume	Curated Instances	Primary Clinical Focus & Characteristics
JAMA (Chen et al., 2025a)	1,524	1,511	Real-world, naturalistic narratives with expert rationales.
MedXpertQA-Diag (Zuo et al., 2025)	2,455	234	Manually filtered for strict diagnostic reasoning tasks.
DiagnosisArena (Zhu et al., 2025)	1,113	915	Structured journal cases designed for open-ended reasoning.
Med-Evidence-2.6k	**5,092**	**2,660**	**LLM-assisted reference benchmark with evidence annotations.**

Table 6: Statistical overview and composition of the *Med-Evidence-2.6k* dataset. The curated instances reflect the final volume of cases successfully processed through our annotation pipeline.

provides a reference benchmark for evidence-aware diagnosis evaluation. Because each EviDx run requires multi-step tool interaction, trajectory logging, and repeated calls to API-based frontier models, exhaustive evaluation across all models and all cases would substantially increase inference cost. The main experiments therefore use a fixed random subset of 100 instances from each source benchmark, enabling controlled cross-model comparison under the same cases and interaction budget.

D.2 The 3-Step Annotation Pipeline

Instead of relying on zero-shot LLM generation, which is prone to clinical hallucination, we engineered a rigorous 3-step agentic workflow powered by GPT-5.2.

Step 1: Diagnostic Planning: A *Senior Medical Researcher* LLM agent analyzes the raw clinical case alongside the ground-truth diagnosis to generate targeted queries. These queries are designed to fetch specific diagnostic criteria and distinguish the true diagnosis from likely differential conditions.

Step 2: Knowledge Retrieval: The generated queries are executed against our local hybrid retrieval engine (accessing StatPearls). This grounds the subsequent annotation strictly in established medical literature rather than the LLM's parametric memory.

Step 3: Evidence Annotation: An *Expert Medical Annotator* LLM agent (set to temperature 0.0 for determinism) synthesizes the retrieved textbook context, the raw case, and the ground truth. It extracts the reference evidence under strict schema constraints, categorizing each clue as *Inclusion*, *Exclusion*, or *Differentiation*.

D.3 Prompt Setup

To ensure the extracted evidence strictly represents the true information bottleneck, our curation pipeline avoided zero-shot generation. Instead, we utilized a multi-step prompt chaining strategy. Below are the exact system instructions used to guide GPT-5.2 at each annotation step.

Step 1: Diagnostic Planning.
The objective of this step is to have the planning model formulate targeted queries before making any extractions, thereby grounding the subsequent steps in external knowledge.

Step 1 Prompt: Senior Medical Researcher

You are a Senior Medical Researcher. Analyze the Clinical Case and the Diagnosis (Ground Truth). Generate 2-3 targeted search queries to: 1. Find diagnostic criteria for the Ground Truth. 2. Verify connections between specific case symptoms and the diagnosis. 3. Distinguish from other likely conditions. Output strictly valid JSON format: `{"queries": ["query1", "query2"]}`

Step 2: Knowledge Retrieval (Execution).
The queries generated in Step 1 are executed locally via our Clinical Consultant. The top retrieved textbook chunks are then aggregated and injected into the Step 3 prompt as the reference context.

Step 3: Evidence Annotation.
This final step synthesizes the raw text, the ground truth, and the retrieved context. The prompt explicitly introduces the 'Exact Substring' constraint to eliminate textual hallucination during the evidence extraction process.

Step 3 Prompt: Expert Medical Annotator

You are an Expert Medical Annotator. Identify "Reference Evidence" in the clinical case that supports the Ground Truth Diagnosis. You MUST base your reasoning on the provided Reference Context.
OUTPUT SCHEMA (JSON)

```
{
  "reference_evidence": [
    {
      "original_text": "Exact substring from case
      text",
      "feature_name": "Standardized medical term",
      "type": "Inclusion" | "Exclusion" | "
      Differentiation",
      "textbook_reference": "Quote or summary from
      context",
      "reasoning": "Why this evidence matters."
    }
  ]
}
```

RULES: 1. `original_text` MUST be an EXACT copy (substring) from the Clinical Case. 2. Output strictly valid JSON.

D.4 Annotated Data Example

To illustrate the precise granularity of our curation pipeline, we present the fully annotated reference evidence snippet for JAMA Case #1386, the identical diagnostic trajectory analyzed in Section A.1.

This JSON structure exemplifies the "information bottleneck" concept central to our Level 3 evaluation. Rather than merely summarizing the clinical narrative, it meticulously categorizes extracted clues into distinct cognitive roles. Notably, it not only captures the fundamental *Inclusion* criteria (e.g., the anatomical outpouching) but explicitly isolates the *Differentiation* logic required to navigate a false-negative clinical trap (the misleading Meckel scan). By structuring the dataset in this grounded format, *Med-Evidence-2.6k* provides an LLM-assisted reference benchmark for evaluating agentic evidence recall, moving beyond superficial binary outcome accuracy.

Curated Instance: JAMA Case 1386 (Abridged)


```
{
  "original_id": "1386",
  "dataset": "JAMA",
  "ground_truth_text": "Meckel diverticulum",
  "annotation_meta": {
    "search_queries": [
      "Meckel diverticulum diagnostic criteria adults
      ...",
      "Distal ileal ulcer differential diagnosis
      Meckel vs Crohn..."
    ]
  },
  "reference_evidence": [
    {
      "original_text": "an outpouching in the distal
      ileum (Figure, A)
                        with an accompanying
      ulceration of the adjacent
                        mucosa (Figure, B).",
      "feature_name": "Distal ileal outpouching with
      adjacent ulceration",
      "type": "Inclusion",
      "textbook_reference": "Meckel diverticulum is
      ... an outpouching of
                             the distal ileum.
      Ectopic gastric tissue within
                             the diverticulum causes
      ulceration... (Nelson)",
      "reasoning": "Classic anatomic location and
      adjacent ulceration are
                   hallmark features of Meckel
      diverticulum."
    },
    {
      "original_text": "iron deficiency anemia (
      hematocrit, 31%) and
                        hemeoccult-positive stool.",
      "feature_name": "Occult GI bleeding causing
      iron deficiency anemia",
      "type": "Inclusion",
      "textbook_reference": "Acid secretion from the
      ectopic gastric mucosa
                             within the diverticulum
      can result in GI
                             bleeding. (StatPearls)",
      "reasoning": "Chronic bleeding from Meckel-
      related ulcers leads to
                   heme-positive stools and iron
      deficiency anemia."
    },
    {
      "original_text": "a Meckel scan failed to
      identify a source of GI
                        bleeding.",
      "feature_name": "Negative Meckel scan does not
      exclude diagnosis",
      "type": "Differentiation",
      "textbook_reference": "Because not all
      diverticula are seen, scan may
                             miss cases, especially
      in adults where diagnostic
                             accuracy is often
      unsatisfactory. (StatPearls)",
      "reasoning": "Explains why a negative scan does
       not rule out Meckel
                   diverticulum, consistent with
      adult false negatives."
    }
  ]
}
```


D.5 Quality Assurance and Limitations

LLM-assisted medical annotation requires quality controls to keep extracted evidence grounded in the original case text. We therefore combine algorithmic constraints, source-text verification, and

partial physician review.

Algorithmic and Programmatic Safeguards. The core of our quality assurance relies on the strict, deterministic constraints embedded within the Step 3 Annotator (set to temperature 0.0). By enforcing the `Exact Substring` rule, the pipeline is fundamentally restricted to *extractive* rather than *generative* actions. The annotated reference evidence must physically exist in the original raw text. Furthermore, the intermediate knowledge retrieval step ensures that the reasoning bridging the clinical text and the ground truth diagnosis is anchored in established medical literature, effectively bypassing the parametric memory of the LLM.

Physician Review. We conduct quality checks for the LLM-assisted evidence curation pipeline. The research team first verifies schema adherence, exact-substring validity, and consistency between each extracted evidence span and the original case text. During audit-package construction, each span is programmatically highlighted against the original case text.

We further conduct a structured physician audit on 100 stratified cases from Med-Evidence-2.6k, covering 515 unique evidence spans. Three physicians independently reviewed the same audit package. Reviewers assessed span-level clinical relevance and diagnostic utility, and case-level missing diagnosis-critical evidence. As shown in Table 7, 1,419/1,545 span judgments (91.8%) rated the span as both clinically relevant and diagnostically useful. By majority vote, 507/515 unique spans (98.4%) were rated as both relevant and useful. At the case level, no case was marked as having major missing evidence by majority vote.

Audit item	Value
Reviewed cases	100
Unique evidence spans	515
Physician reviewers	3
Both relevant and useful, judgment level	91.8%
Both relevant and useful, majority vote	98.4%
Any negative span judgment	2.0%
No / minor / major missing evidence	83 / 13 / 0 cases
No majority on missing evidence	4 cases

Table 7: Physician audit of Med-Evidence-2.6k. The audit is used as a quality check for LLM-assisted extractive evidence annotations. Missing-evidence counts are aggregated by majority vote.

E Implementation Details

To ensure full reproducibility of our interactive diagnostic framework, we detail the environment and specific hyperparameter configurations used in our experiments.

Hardware and Serving Environment.
Local deployment of the 8B-parameter open-weight models (Ministral3-8B, Qwen3-8B, and Llama3.1-8B) and the embedding model (Qwen3-embedding-4B) was performed using the vLLM framework, whereas the frontier large-scale models (GPT-5.2, Claude Sonnet 4.6, GLM-5, and DeepSeek-V3.2) were accessed through their official APIs.

Agent Hyperparameters and Roles.
To minimize parametric hallucination and ensure deterministic reasoning, the generation temperature for the Diagnostician, the Context Initializer, and the LLM-Judge was strictly set to 0.0. For the Clinical Consultant, we applied a temperature of 0.1 during the query decomposition phase to maintain keyword precision, and 0.3 during the generation phase. Notably, the EHR Executor is not LLM-based; it functions entirely as a deterministic retrieval server executing structured database queries. Additionally, to guarantee lossless information extraction, the Context Initializer employs a dual-pass refinement pipeline: it first synthesizes a draft EHR from the unstructured raw patient text, and subsequently refines this output by cross-referencing the draft with the original raw text.

Observer Configuration.
For the Observer, the strict maximum interaction turn limit (T_{max}) was set to 50 steps to preempt non-terminating loops in difficult cases. Rather than utilizing rigid static boundaries, the Observer implements step-aware dynamic thresholds to balance exhaustive exploration with practical termination. The dynamic diagnostic uncertainty threshold $\tau_H(t)$ at step t is defined as:

$$\tau_H(t) = \tau_{H_{base}} + \max(0, t-5) \times 0.05$$

This formulation allows the system to gradually relax its tolerance for uncertainty after the 5th step, acknowledging that absolute entropy reduction becomes exponentially more difficult as clinical ambiguities persist. Similarly, the dynamic runtime evidence coverage threshold $\tau_{\mathcal{V}}(t)$ is defined as:

$$\tau_{\mathcal{V}}(t) = \max(0.1, \tau_{\mathcal{V}_{base}} - \max(0, t-10) \times 0.01)$$

This mechanism incrementally lowers the required evidence coverage ratio after the 10th step (bounded at a minimum of 0.1). By dynamically relaxing this constraint, the Observer helps prevent the agent from deadlocking in an infinite information-seeking loop when the reachable clinical evidence space is intrinsically sparse.

Clinical Consultant Setup.
The external medical knowledge engine underlying the Consultant was constructed using medical corpora provided by MedRAG (Xiong et al., 2024), which includes StatPearls and medical textbooks. To optimize evidence acquisition, we implemented a hybrid, corpus-specific retrieval strategy. For the highly structured StatPearls database, we utilized dense semantic retrieval powered by the Qwen3-embedding-4B model, with the resulting vector embeddings indexed and managed via the Qdrant vector database for high-throughput similarity search. Crucially, this was augmented with a hierarchical parent-child retrieval mechanism: once a highly relevant localized text chunk (child) is identified, the retriever automatically expands the context window to return the entire overarching medical entry (parent), ensuring no critical surrounding context is fragmented. Conversely, for medical textbooks, we applied BM25 (Robertson and Zaragoza, 2009) for lexical retrieval to precisely capture exact clinical terminology and keyword matches. Finally, following MedRAG, the disparate results from these semantic and lexical retrievers were integrated into a unified, ranked evidence context using Reciprocal Rank Fusion (Cormack et al., 2009; Xiong et al., 2024).