CAN LLMS VALUE THE RIGHT EVIDENCE? EVIDENCE–VALUE MISALIGNMENT IN DYNAMIC MEDICAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

A correct diagnosis reached from insufficient or misleading evidence can pose a clinical hazard, yet outcome-based accuracy may reward such lucky guesses. We call this mismatch between diagnostic decisions and the value of available evidence **Evidence-Value Misalignment (EVM)**. To disentangle evidential grounding independently from diagnostic accuracy, we introduce MEDEVM, a dynamic benchmarking environment comprising 1,050 cases across 24 disease systems. Observations arrive turn by turn, requiring models to continuously calibrate its decision by deciding whether to wait for more evidence or submit a diagnosis. Across 9 LLMs, four interesting patterns are observed. (1) **Miscalibrated evidence tracking**. Making a diagnosis often fails to calibrate evidence sufficiency, even in more capable models, and even worsens in reasoning mode. (2) **Misaligned diagnosis submission**. Confidence in the correct diagnosis often fails to ensure timely submission despite sufficient evidence. (3) **Evidence order matters**. Reordering the same evidence changes diagnoses even when model confidence remains similar. (4) Misleading evidence remains influential. Added misleading evidence redirects diagnoses even after prior evidence becomes sufficient. We further verify that EVM predicts errors and that preventing premature submission improves accuracy. These findings motivate **Evidence-Verified Diagnosis Harness (EVD-HARNESS)**. It decouples diagnosis generation from submission through an offline Contrastive Diagnostic Wiki and three online control stages, namely observation management, proposal and witness verification, and diagnosis submission control. Across five LLMs, EVD-HARNESS improves accuracy by 12.0–51.1 percentage points while mitigating EVM-related failures. Our results demonstrate that verifying evidential support before submission can make diagnostic decisions more reliable.

1 INTRODUCTION

Large language models (LLMs) show strong reasoning capabilities on difficult medical questions and complex cases (Singhal et al., 2023). To evaluate these capabilities beyond static medical QA (Jin et al., 2021; Pal et al., 2022; Zhu et al., 2026), AgentClinic (Schmidgall et al., 2024), AI Hospital (Fan et al., 2025), and MediQ (Li et al., 2024) introduce multi-turn diagnostic interactions in which models gather patient information and refine diagnoses, bringing evaluation closer to real clinical encounters. EHRWorld (Mu et al., 2026) and ChronoMedicalWorld (Wang et al., 2026) further extend these interactive settings by modeling how patient states evolve under clinical interventions.

Yet increasingly realistic clinical environments do not make final diagnostic accuracy a sufficient measure of clinical reasoning. Accuracy-based evaluation treats a correct diagnosis submitted on insufficient evidence the same as one supported by sufficient evidence, leaving the evidential basis for submission unexamined. Although recent work has examined vulnerabilities in multi-turn settings and premature diagnosis submission (Laban et al., 2026; Fang et al., 2026; Hsu et al., 2026), a fundamental question remains: Do LLMs appropriately weigh the available evidence when deciding to submit a diagnosis? Here, evidence value refers to an observation's ability to narrow the differential diagnosis and inform whether the available evidence warrants submission.

To answer this question, we introduce MEDEVM, a progressive-evidence diagnosis benchmark comprising 1,050 cases across 24 disease systems (Section 3). Each case consists of an ordered

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

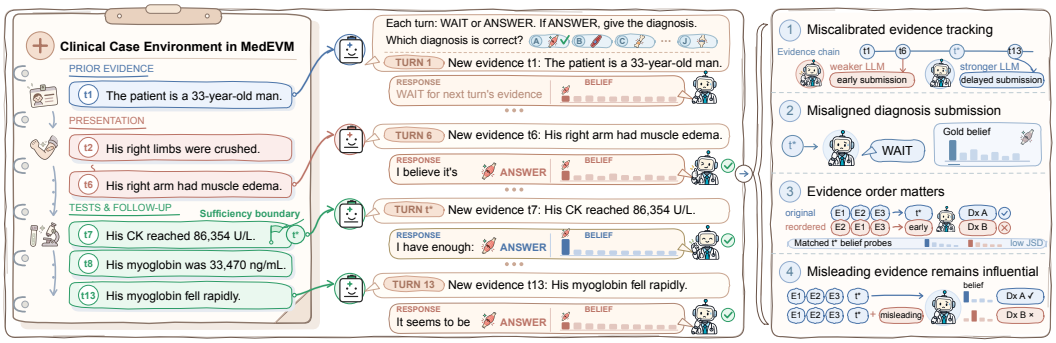


Figure 1: Overview of MEDEVM. Final accuracy alone obscures differences in when models submit diagnoses. In MEDEVM, clinical evidence arrives turn by turn, and the LLM Agent must wait or submit a diagnosis. This evaluation reveals four key findings hidden by answer-only scoring.

sequence of atomic observations extracted from a case report, a gold diagnosis, and nine plausible alternative diagnoses. After each observation, the model must decide whether to wait for more evidence or submit a diagnosis. We annotate the sufficiency boundary t^* as the earliest turn at which the accumulated evidence is sufficient to distinguish the gold diagnosis from all alternatives. We define Evidence-Value Misalignment (EVM) as a mismatch between when a model submits a diagnosis and when the available evidence becomes sufficient to support that submission.

Controlled experiments across nine LLMs reveal four EVM patterns (Section 4). (1) **Miscalibrated evidence tracking**. Making a diagnosis often fails to calibrate evidence sufficiency, even in more capable models, and even worsens in reasoning mode. (2) **Misaligned diagnosis submission**. Confidence in the correct diagnosis often fails to ensure timely submission despite sufficient evidence. (3) **Evidence order matters.** Diagnostic decisions are sensitive to evidence order, even when model confidence remains similar. (4) Misleading evidence remains influential. Added misleading observations can redirect diagnoses regardless of whether prior evidence was sufficient.

We next examine whether EVM helps explain diagnostic decisions and errors (Section 5). Measures of evidence valuation provide information beyond model confidence for predicting when models submit a diagnosis and whether that diagnosis is correct. We then test the effect of correcting premature submission using *Natural Gate*, an oracle intervention that defers submission until the annotated sufficiency boundary t^*. Among cases in which models originally submitted before t^*, this intervention improves accuracy by 29.6–48.4 percentage points.

These findings motivate the Evidence-Verified Diagnosis Harness (EVD-HARNESS), which separates diagnosis generation from submission using an offline Contrastive Diagnostic Wiki and three online control stages (Section 6). Observation management organizes incoming evidence and retrieves relevant diagnostic knowledge. Proposal and witness verification generates candidate diagnoses and checks the evidence supporting them. Diagnosis submission control validates the cited evidence before allowing a diagnosis to be submitted. Across five LLMs, EVD-HARNESS improves diagnostic accuracy by 12.0–51.1 percentage points while reducing EVM-related failures. These results show why valuing the right evidence should be evaluated separately from producing the right answer. Our contributions are threefold:

- We introduce MEDEVM, a progressive-evidence benchmark that evaluates diagnosis submission against annotated evidence sufficiency, revealing failures that final diagnostic accuracy alone cannot capture.
- We identify Evidence-Value Misalignment (EVM) through controlled analyses of submission timing and responses to reordered and misleading evidence. We show that EVM measures help identify diagnostic errors, while delaying premature submission improves accuracy.
- We develop EVD-HARNESS, which separates diagnosis generation from submission control, better aligning submission with evidence sufficiency and improving diagnostic accuracy.

2 RELATED WORK

Dynamic Medical Diagnosis Benchmarks. Static benchmarks such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019) evaluate medical knowledge and

final answers from fully presented contexts, but cannot reveal how diagnosis evolves as evidence arrives. Recent benchmarks instead model diagnosis as a dynamic process. DDXPlus (Fansi Tchango et al., 2022) supports symptom-seeking differential diagnosis, while AgentClinic (Schmidgall et al., 2024), AI Hospital (Fan et al., 2025), and MMD-Eval (Liu et al., 2025) simulate consultations that involve patients, examinations, or tools. MediQ (Li et al., 2024) evaluates proactive acquisition of missing information. MINT (Fang et al., 2026) reveals fixed evidence over turns to study rushing, lures, and self-correction, while MedEinst (Chen et al., 2026) uses counterfactual cases to expose shortcut-driven errors hidden by endpoint accuracy. Complementary to these efforts, MEDEVM treats neither dialogue quality nor the final answer as the sole endpoint. It aligns turn-level belief and natural wait-or-answer behavior with an externally annotated earliest sufficiency boundary, then uses controlled interventions to test how misalignment changes diagnostic outcomes.

From Uncertainty Estimation to Diagnosis Submission Control. Prior work estimates answer reliability through model self-evaluation, black-box confidence elicitation, and semantic uncertainty (Kadavath et al., 2022; Xiong et al., 2024; Farquhar et al., 2024). Selective prediction converts such scores into answer-or-abstain decisions (Kamath et al., 2020), and MedAbstain (Machcha et al., 2026) extends this framework to medical MCQA under missing or perturbed context. For sequential agents, KnowNo (Ren et al., 2023) forms conformal action sets and requests help when the next action is not uniquely supported. Closest to our failure setting, Handler et al. (2026) evaluate premature closure as inappropriate diagnosis submission under uncertainty. These approaches provide important reliability tools, but answer confidence is not equivalent to clinical evidence sufficiency. A model may submit a diagnosis early with high confidence or keep waiting after it becomes uniquely identifiable. We therefore treat diagnosis submission as an evidence-conditioned decision distinct from answer generation. This framing motivates mitigation that reduces premature submission without merely replacing it with delayed answers or persistent abstention.

3 DYNAMIC MEDEVM BENCHMARK

To systematically evaluate whether LLMs can align diagnosis submission with incrementally available evidence, we introduce MEDEVM, a progressive-evidence diagnosis benchmark that simulates a dynamic clinical environment by revealing atomic observations from case reports in sequential turns, and records when LLMs wait, submit a diagnosis, or revise their diagnostic beliefs.

Task Formulation. We formulate progressive-evidence diagnosis as a sequential decision-making problem under partial observability. A clinical case i comprises a hidden ground-truth diagnosis y_i^*, a candidate set $\mathcal{Y}_i$, and a fixed sequence of atomic evidence $E_i = (e_{i1}, \ldots, e_{iT_i})$. At turn t, the environment reveals one new observation e_{it} and appends it to the history $h_{it} = E_{i,\le t}$. The LLM agent then samples an action $a_{it} \sim \pi_\theta(\cdot \mid h_{it})$ from $\mathcal{A}_i = \{\text{WAIT}\} \cup \{\text{ANSWER}(y) : y \in \mathcal{Y}_i\}$. Taking WAIT advances the environment to $t+1$, while outputting an ANSWER irreversibly terminates the trajectory. Reaching T_i without answering is considered a failure. Since the evidence schedule is fixed, the action determines when the trajectory ends but not which observation appears next.

Evidence-Value Misalignment. Let $\mathcal{C}_i(t) \subseteq \mathcal{Y}_i$ denote the subset of candidate diagnoses logically compatible with the history h_{it}. We define the **Sufficiency Boundary** as $t_i^* = \min\{t : \mathcal{C}_i(t) = \{y_i^*\}\}$.

Before t_i^*, a confounder remains plausible. At t_i^*, the gold first becomes uniquely identifiable. We use its annotated position as the reference and test timing tolerance in Appendix A3.2. Let $s_{it} = \mathbf{1}[t \ge t_i^*]$ indicate whether the evidence is sufficient, and $c_{it} = \mathbf{1}[a_{it} = \text{ANSWER}]$ denote whether the model submits a diagnosis. We decompose EVM into two mutually exclusive errors

$$m_{it}^{\text{pre}} = c_{it}(1 - s_{it}), \quad m_{it}^{\text{delay}} = (1 - c_{it})s_{it}, \quad m_{it} = m_{it}^{\text{pre}} + m_{it}^{\text{delay}} = |s_{it} - c_{it}|. \tag{1}$$

Here, m_{it}^{pre} and m_{it}^{delay} measure premature and delayed diagnosis submission, respectively. EVM quantifies the alignment between decision timing and evidence sufficiency, which is distinct from diagnostic accuracy. A correct guess before t_i^* is heavily penalized as a premature EVM error, while an incorrect answer after t_i^* is treated as a diagnostic error rather than a timing mismatch.

Benchmark Construction. We collect case reports from PubMed-indexed journals (White, 2020), covering NEJM, BMJ Case Reports, Gastroenterology, JAMA, and The Lancet. After applying

formatting and completeness filters, we retain 36,020 reports that contain a confirmed diagnosis and sufficient clinical information. We next remove diagnostic leakage, segment each narrative into typed atomic evidence in chronological order, and treat the confirmed diagnosis as ground truth. Using a structured prompt, GPT-5.4 (OpenAI, 2026a) generates nine clinically plausible confounders for each case. Three proprietary models from different families, namely Claude Opus 4.8 (Anthropic, 2026a), Gemini 3.1 Pro Preview (Gemini Team, 2026a), and GPT-5.6-sol (OpenAI, 2026b), independently annotate the sufficiency boundary t_i^*, and we keep only cases with unanimous boundary annotations. Finally, Gemini 3.5 Flash (Gemini Team, 2026b) and five graduate students with medical backgrounds jointly verify the quality of each sample, resulting in 1,050 cases for the final MEDEVM benchmark. Details can be found in Appendix A1.

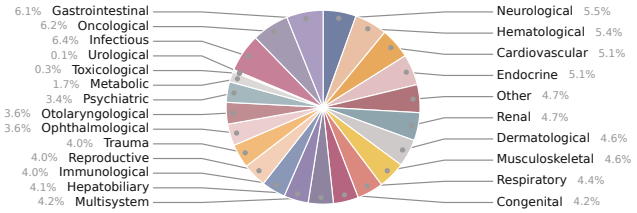


Figure 2: Distribution of all 1,050 MEDEVM benchmark cases across 24 disease systems.

Table 1: Data statistics in different construction phases and final MEDEVM.

Construction Phases		Final MEDEVM	
Phase	#N	Statistic	Value
Collected	75,434	Observations	26,869
Structured	52,147	Disease systems	24
Qualified	36,020	Diagnoses	953
Valid MCQ	1,810	Candidates/case	10
MEDEVM	1,050	Evidence/case	25.6

Evaluation Protocol. We evaluate 9 advanced LLMs, including the Qwen3 series (Qwen Team et al., 2025) with 8B (Q3-8B), 14B (Q3-14B), 32B (Q3-32B), and 235B-A22B (Q3-235B) parameters, Qwen3.6-27B (Q3.6-27B; Qwen Team, 2026a) and Qwen3.8-27B (Q3.8-27B; Qwen Team, 2026b) as similar-capacity variants, and DeepSeek-V4-Flash (DS-V4-F; DeepSeek Team et al., 2026), Kimi-K2.5 (K2.5) and Kimi-K2.6 (K2.6; Kimi Team et al., 2026) from external families. During evaluation, LLMs are constrained to output structured JSON responses. To measure the internal belief distribution without altering the natural trajectory, we apply a non-intervening probability probe to a cloned context. Given a forced JSON prefix up to the answer field, the probe extracts the exact sequence log-probability $\ell_{it}(y)$ of every candidate $y \in \mathcal{Y}_i$ and normalizes it as $p_{it}(y) = \exp(\ell_{it}(y)) / \sum_{y' \in \mathcal{Y}_i} \exp(\ell_{it}(y'))$.

We report three recurring metrics: (1) *Diagnosis Submission Accuracy*, correct diagnoses among answered trajectories, (2) *Premature Submission Rate*, submissions before t_i^* among all trajectories, and (3) *Gold Belief* $p_{\text{gold},it} := p_{it}(y_i^*)$, defined as the normalized probability $p_{it}(y_i^*)$ of the correct answer. Experiment-specific metrics are introduced locally and defined in Appendix A2. Paired effects use case-clustered 95% bootstrap confidence intervals unless stated otherwise.

4 How EVM Arises: Evidence and Diagnosis Submission Diverge

We examine EVM in medical diagnosis through four progressive experiments on its prevalence and mechanisms. These characterize diagnosis submission timing, probe diagnostic knowledge, test evidence order, and vary misleading evidence strength.

RQ1: Does capability reduce EVM? We categorize trajectories by whether the model submits a diagnosis before, at, or after the sufficiency boundary t^*, or remains unanswered after all evidence arrives. Figure 3 compares model sizes, similar-capacity versions, and model families. We also enable thinking mode in seven models to test whether additional reasoning improves diagnosis submission timing.

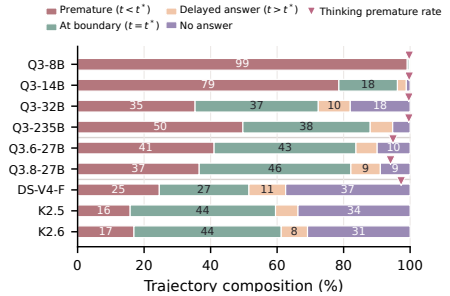


Figure 3: Case proportions by diagnosis submission time relative to the first sufficient-evidence turn t^*. Red triangles show early submission with thinking.

All nine models exhibit substantial EVM, with capability gains changing its form rather than steadily reducing it. Qwen3-8B submits early in 99.1% of cases, whereas larger and later LLMs reduce early submission but often wait despite sufficient evidence or never answer. These patterns persist when

submissions within one or two turns of the annotated boundary are treated as aligned (Appendix A3.2). Thinking increases early submission to 94.1–99.7% across seven models under the strict boundary, showing that additional reasoning can reinforce premature diagnosis submission.

TAKEAWAY: Models exhibit EVM: diagnosis submission timing is misaligned with evidence sufficiency. Greater capability does not eliminate it, and reasoning mode worsens it.

RQ2: Is EVM caused by missing diagnostic knowledge? To distinguish whether EVM reflects missing knowledge or a failure to act on evidence, we probe the ten candidate diagnoses of each MEDEVM case in cloned contexts without changing its original trajectory. Gold belief p_{gold} is the probability of the ground-truth diagnosis normalized over all candidates. We measure p_{gold} at diagnosis submission and at t^*. Since submitting a diagnosis terminates the diagnostic trajectory, boundary statistics include *only trajectories remaining active at* t^*. We also conduct a thinking-mode sensitivity analysis for seven models and report expected calibration error (ECE). We group N predictions into M equal-width confidence bins B_m using the highest candidate probability. Then $\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\operatorname{acc}(B_m) - \operatorname{conf}(B_m)|$, where acc and conf are the accuracy and mean confidence within each bin. Lower ECE indicates more reliable confidence (Appendix A3).

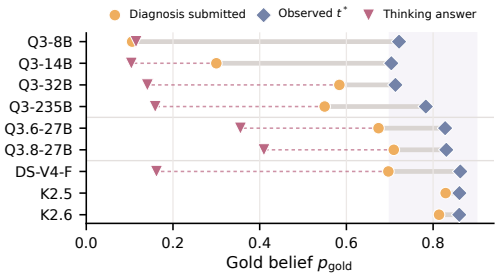


Figure 4: Gold belief from direct probes at diagnosis submission (circles) and t^* (diamonds). Triangles use probes with thinking.

Table 2: Effects of thinking on early submission, accuracy, gold belief, and ECE (%). Accuracy uses answered cases. Off is gray. On is red for performance gains and blue otherwise.

Model	Early ↓		Accuracy ↑		p_{gold} ↑		ECE ↓	
	off	on	off	on	off	on	off	on
Q3-8B	99.1	99.7	10.9	11.3	13.8	11.5	65.4	89.8
Q3-14B	78.6	99.7	30.5	12.0	28.7	10.4	78.0	90.1
Q3-32B	35.3	99.6	59.2	13.9	43.0	14.1	55.0	87.9
Q3-235B	49.7	99.6	55.4	17.3	52.6	15.9	66.9	86.5
Q3.6-27B	41.0	94.9	70.2	37.8	64.8	35.6	36.3	76.1
Q3.8-27B	36.6	94.1	72.9	36.8	64.2	41.0	31.3	74.4
DS-V4-F	24.6	97.3	70.5	28.5	43.9	16.2	36.8	39.7

At observed t^*, mean gold belief p_{gold} ranges from 0.70 to 0.86, showing that models can identify the correct diagnosis when sufficient evidence is available. p_{gold} at diagnosis submission is more variable (Figure 4). Thinking mode lowers submission-point gold belief and increases ECE in all seven models, indicating weaker support for the correct diagnosis and less reliable confidence when answers are submitted (Table 2). These findings suggest that models possess relevant diagnostic knowledge, but knowledge alone does not translate into appropriate diagnosis submission timing.

TAKEAWAY: EVM is not merely a knowledge deficit. Models can confidently identify the correct diagnosis from sufficient evidence yet submit at inappropriate times.

RQ3: Does diagnosis submission depend on evidence order? For each case, we reorder only the clinical evidence presented before t^* without changing any evidence content. The evidence at t^* and all subsequent evidence retain their original order. Both sequences provide identical evidence sets at t^*. Diagnosis outcome disagreement occurs when the two trajectories submit different diagnoses, or one answers while the other remains unanswered. For pairs reaching t^* in both orders, we compare probability distributions over the ten candidate diagnoses using Jensen–Shannon divergence (JSD). Lower JSD indicates more similar beliefs, with zero denoting identical distributions. For each metric, we average across permutations within each case and then across cases.

Reordering changes 29.0–52.9% of final diagnosis outcomes across the full cohort (Figure 5, left). Among pairs reaching t^* in both orders, mean JSD is low (0.009–0.044), yet substantial outcome disagreement remains (Figure 5, right). Thus, the evidence sequence affects final diagnoses even when both trajectories reach sufficient evidence and yield similar diagnostic beliefs (Appendix A3).

TAKEAWAY: Simply reordering the same evidence changes diagnosis outcomes despite similar diagnostic beliefs at the sufficiency boundary. Evidence order therefore affects how models act on available information.

RQ4: How does misleading evidence change diagnosis submission? We use two paired interventions to manipulate evidence value. First, we insert repeated clinical evidence or clinically irrelevant

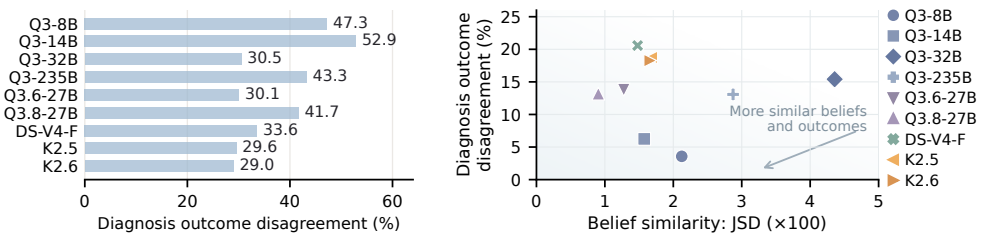
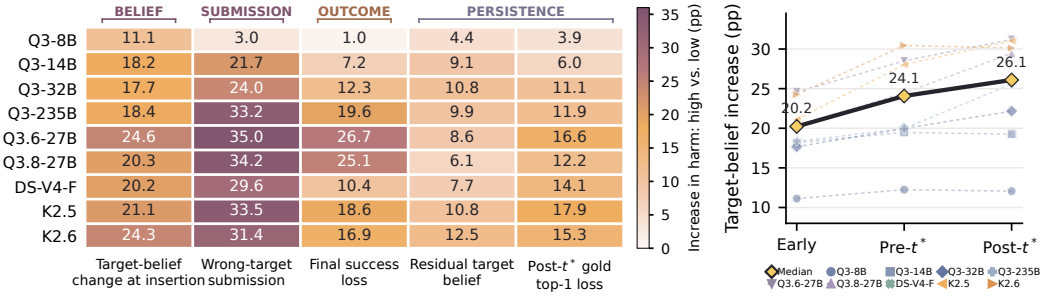


Figure 5: Diagnosis outcome disagreement after reordering in all cases (left), and its relation to belief similarity at t^* among paired orders that both reach t^* (right).

text into each case to test whether additions without new diagnostic information change diagnostic beliefs or decisions. Second, for 318 cases, we select an incorrect candidate diagnosis as the *target diagnosis*. We construct weak and strong misleading evidence that differ in their support for this target while matching the evidence type, semantic direction, and format. We insert the paired versions at the same position early in the diagnostic trajectory, one turn before t^*, or one turn after t^*, with turns indexed in the augmented sequence. Target belief is the probability assigned to the incorrect target diagnosis, tracking confidence in this misleading alternative.



(a) Model-level propagation across endpoints.

(b) Immediate response across stages.

Figure 6: Strong versus weak misleading evidence (pp, positive = harm). (a) From left: immediate target-belief gain, target submission rate increase, accuracy loss, target-belief gain after all evidence, and post-t^* loss in correct diagnoses ranked first at completion. The first four use early insertion. (b) Immediate target-belief gains by insertion stage, with dashed curves for models and a solid median.

The next diagnosis response differs in 5.5–34.4% of comparisons between repeated-evidence and irrelevant-text branches, although neither adds new diagnostic information (Appendix A3). Stronger misleading evidence raises belief in the incorrect target and its diagnosis submission rate, reducing final diagnostic accuracy by up to 26.7 percentage points (Figure 6a). The additional belief in the incorrect target persists after all original evidence arrives. Even after t^*, newly inserted misleading evidence shifts beliefs toward the incorrect target and reduces how often the correct diagnosis has the highest final probability.

TAKEAWAY: Models remain sensitive to misleading evidence even after evidence is sufficient. Stronger misleading evidence shifts beliefs toward incorrect diagnoses and causes persistent errors.

5 WHY EVM MATTERS: PREDICTING AND PREVENTING ERRORS

The value of EVM extends beyond describing model behavior. We test whether its features help predict submission timing and assess diagnostic errors, and whether reducing EVM improves accuracy.

We separate two prediction tasks: whether the next action submits a diagnosis after a WAIT, and whether a diagnosis is incorrect at its submission turn. The first uses checkpoints with an observed next turn, while the second uses each first submission. For each LLM, we fit regularized logistic regression with shared case-level splits and balance both classes within each split by majority-class undersampling. Baseline M1 uses (1) the fraction of evidence received, (2) evidence and candidate counts, and (3) diagnostic confidence, summarizing certainty across candidates. Our M3 adds (4) diagnosis submission propensity, capturing the tendency to submit rather than wait, and (5) belief instability, capturing changes in confidence between turns. AUROC (Peterson et al., 1954) measures how well predictions distinguish submission from waiting, or incorrect from correct diagnoses,

respectively. For next-turn submission, log loss (LL) evaluates predicted submission probabilities. A negative $\Delta\text{LL} = \text{LL}_{M3} - \text{LL}_{M1}$ indicates better estimates of next-turn submission probabilities.

M3 improves AUROC for seven of nine LLMs in each task (Table 3). Pooled AUROC rises from 0.835 to 0.854 for next-turn submission and from 0.845 to 0.879 for submission errors. EVM features thus inform both submission timing and error assessment beyond confidence alone. Next-turn log loss improves for eight models, with full baseline comparisons in Appendix A4.

Table 3: AUROC for next-turn submission and errors at submission. M3 adds EVM features to M1. Higher is better.

	Next submission		Submission error	
Model	M1	M3	M1	M3
Q3-8B	0.795	**0.820**	0.640	**0.724**
Q3-14B	0.780	**0.827**	0.627	**0.822**
Q3-32B	0.826	**0.846**	0.711	**0.841**
Q3-235B	0.827	**0.840**	0.789	**0.847**
Q3.6-27B	**0.815**	0.812	**0.783**	0.777
Q3.8-27B	0.857	**0.885**	0.773	**0.789**
K2.5	**0.851**	0.842	0.800	**0.800**
K2.6	0.863	**0.873**	0.711	**0.723**
DS-V4-F	0.790	**0.807**	0.814	**0.839**
Pooled	0.835	**0.854**	0.845	**0.879**

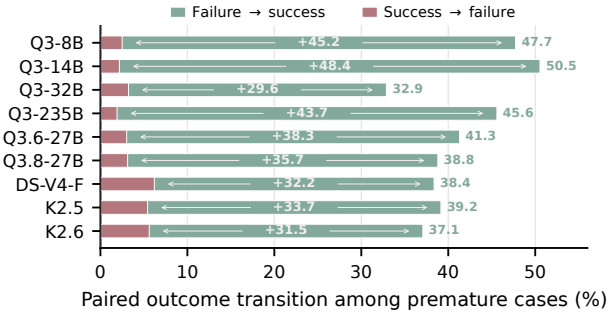


Figure 7: Natural Gate blocks pre-boundary diagnosis submission. Green shows failures rescued, red successes lost, and pale arrows net gains among early-answer cases.

To test whether reducing EVM changes diagnostic outcomes, we apply *Natural Gate* to cases where the model originally answers before t^*. *Natural Gate* replaces every attempted diagnosis submission before t^* with WAIT, allowing further evidence to arrive. At t^*, we remove the gate and restore the original action prompt. The model then decides whether to wait or submit a diagnosis using the accumulated evidence. Among all intervened cases, 32.9–50.5% change from failure to success, while far fewer change in the opposite direction (Figure 7). Net diagnostic accuracy gains reach 29.6–48.4 percentage points across nine models, with every confidence interval above zero.

TAKEAWAY: EVM features help predict submission timing and assess diagnosis errors beyond confidence alone. Preventing premature submission can rescue failed diagnoses.

6 EVD-HARNESS: EVIDENCE-VERIFIED DIAGNOSIS HARNESS

Our EVM analyses reveal premature diagnosis submission and waiting after decisive evidence arrives. EVD-HARNESS decouples diagnosis generation from submission through an offline contrastive diagnostic Wiki and three online modules: evidence management, proposal and witness verification, and diagnosis submission control. Together, they manage the context and calls of a frozen LLM, using retrieved knowledge to guide diagnosis and current evidence to authorize its submission.

6.1 OFFLINE DIAGNOSTIC WIKI CONSTRUCTION

The *Contrastive Diagnostic Wiki* (CDW) organizes clinical experience around findings that distinguish confusable diseases. We deduplicate published case reports using document identifiers and text similarity, excluding sources associated with MEDEVM, MedXpertQA, DiagnosisArena, and other existing benchmarks, including rewrites and derived questions. An LLM extracts the final diagnosis, competing diagnoses, shared presentations, and the findings used by the report authors to distinguish them. It retains the supporting passages and removes identifying details such as names and hospitals. Independent review checks each explanation against its supporting source passage. The resulting Wiki contains linked disease pages, test pages, and *contrastive bridges*. Each bridge records confusable diseases, their shared presentation, a distinguishing finding, and its clinical context. Related entries are consolidated with source references into a Wiki fixed for evaluation.

6.2 ONLINE DIAGNOSIS CONTROL

Evidence Management For case i, turn t begins when new clinical evidence e_{it} arrives. The *Evidence Ledger* appends it to the complete record $E_{i,\leq t} = (e_{i1}, \ldots, e_{it})$. *Bridge Retrieval* uses this

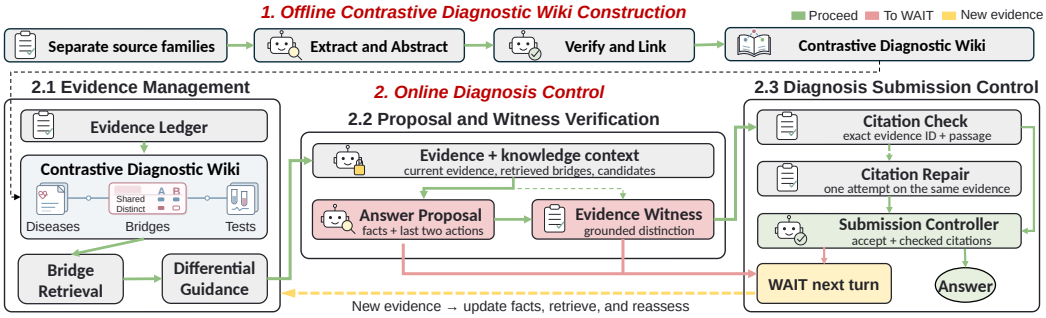


Figure 8: EVD-HARNESS compiles independent case reports into contrastive knowledge, guides diagnosis, and authorizes answers grounded in current evidence.

record and the candidate set $\mathcal{Y}_i$ to retrieve disease and test pages from CDW, then follows their links to contrastive bridges. These pages form the knowledge view K_{it}, which explains the findings that distinguish competing diagnoses. *Differential Guidance* provides $E_{i,\le t}$ and K_{it} to the diagnostic LLM as separately labeled patient evidence and reference knowledge. It asks the model to explain which competing diagnoses the current evidence distinguishes and which remain unresolved. This context supports the next diagnosis proposal or decision to wait.

Proposal and Witness Verification The *Answer Proposal* calls the diagnostic LLM with $E_{i,\le t}$, K_{it}, $\mathcal{Y}_i$, and an action history H_{it} recording the harness actions from at most the two preceding turns. Keeping $E_{i,\le t}$ complete allows reassessment using all available evidence, while limiting H_{it} aims to prevent repeated waiting from dominating the next decision. The model returns $d_{it} \in \mathcal{Y}_i \cup \{\text{WAIT}\}$. A WAIT requests the next evidence unit. A proposed diagnosis instead triggers the *Evidence Witness*, which reviews d_{it} with the same evidence, knowledge, and candidates in a separate LLM context. It explains which findings support d_{it} over competing diagnoses and identifies missing information that could change this choice. It returns an accept-or-wait verdict and citations Q_{it} linking its explanation to patient evidence. Each citation (j, q) pairs an evidence index j with a passage q copied from $e_{ij} \in E_{i,\le t}$, so the cited text can be checked against the patient record.

Diagnosis Submission Control The *Citation Check* reads each $(j, q) \in Q_{it}$, locates e_{ij} in $E_{i,\le t}$, and checks that q is a nonempty passage copied exactly from that evidence unit. Let $\mathcal{P}_{it}$ denote all valid index–passage pairs in this record. If the *Evidence Witness* accepts the diagnosis but citation checking fails, one *Citation Repair* corrects the references using the same $E_{i,\le t}$. The *Submission Controller* then combines the checked citations with the witness verdict. Define $S_{it} = 1$ when the witness accepts d_{it}, explains the distinguishing evidence, and identifies no remaining gap affecting the choice, and $S_{it} = 0$ otherwise. The authorization status V_{it} and first submission turn τ_i are

$$V_{it} = S_{it}\,\mathbf{1}[\varnothing \neq Q_{it} \subseteq \mathcal{P}_{it}], \qquad \tau_i = \min\{t : d_{it} \in \mathcal{Y}_i,\ V_{it} = 1\}. \tag{2}$$

Here, $\mathbf{1}[\cdot]$ equals one when its condition holds and zero otherwise. Thus, $V_{it} = 1$ requires both witness acceptance and a nonempty set of citations matching the observed patient record. At the first such turn, the harness submits $d_{i\tau_i}$ and ends the trajectory. Otherwise, it sends WAIT, appends this action to H_{it}, and repeats the modules when new evidence arrives. An initial WAIT sets $V_{it} = 0$ directly. If no turn qualifies by evidence exhaustion, $\tau_i = \infty$ and the case is recorded as unanswered.

6.3 MAIN RESULTS

We compare the LLM-only baseline with EVD-HARNESS on MEDEVM using Qwen3.8-27B, DeepSeek-V4-Flash, Qwen3-32B, Kimi-K2.6, and Claude-Opus-4-8 (Anthropic, 2026b). Unanswered cases count as errors for both methods. Following the RQ3 and RQ4 settings in Section 4, we reorder evidence and insert duplicate, irrelevant, or paraphrased evidence into cases from MEDEVM, and vary misleading-evidence strength across 318 case pairs.

Accuracy improves for all five models, with fewer premature submissions and unanswered cases (Table 4). The harness also reduces diagnosis changes after reordering and incorrect target submissions under strong misleading evidence. These gains combine improved diagnostic accuracy on the original cases with more consistent decisions under evidence perturbations.

Table 4: LLM only versus EVD-HARNESS (%). Non-answers count as errors. Premature answers precede t^*. Order flip counts diagnosis changes after reordering. Target error counts incorrect target submissions under strong misleading evidence before corrective evidence arrives.

LLM only LLM + EVD-HARNESS

Model	Accuracy ↑		At t^* ↑		Premature ↓		No answer ↓		Order flip ↓		Target error ↓	
	0–100	Value	0–100	Value	0–60	Value	0–60	Value	0–50	Value	0–80	Value
Qwen3.8-27B	60.9 →	**75.6**	41.6 →	**73.8**	50.2 →	**14.9**	2.7 →	**2.4**	37.4 →	**1.8**	54.4 →	**11.0**
DeepSeek-V4-Flash	42.9 →	**68.4**	27.2 →	**70.4**	22.3 →	**17.8**	39.3 →	**1.3**	33.0 →	**2.0**	66.4 →	**15.5**
Qwen3-32B	50.9 →	**62.9**	38.9 →	**67.1**	36.0 →	**14.4**	16.4 →	**6.0**	28.5 →	**4.5**	33.0 →	**6.0**
Kimi-K2.6	60.2 →	**78.0**	36.9 →	**66.4**	49.6 →	**20.4**	5.3 →	**2.9**	32.1 →	**0.9**	49.1 →	**13.9**
Claude-Opus-4-8	34.2 →	**85.3**	24.7 →	**68.4**	11.3 →	**10.9**	59.8 →	**8.4**	44.5 →	**2.9**	27.4 →	**6.9**

6.4 ADDITIONAL EVALUATIONS

Component Ablations On 200 cases from MEDEVM, we remove CDW knowledge or *Evidence Witness*. Without the witness, proposed diagnoses are submitted without review. Removing it while retaining CDW lowers accuracy by 3.5–9.0 points. Removing CDW while retaining the witness lowers accuracy by 9.0–14.5 points (Table 5). These ablations show that retrieved knowledge and evidence review each contribute to diagnostic success.

Table 5: CDW and Evidence Witness ablations (%). P: submission before t^*, N: no answer, A: accuracy.

Model	Full			−CDW			−Evidence Witness			−CDW, −Witness		
	P↓	N↓	A↑	P↓	N↓	A↑	P↓	N↓	A↑	P↓	N↓	A↑
Q3.8-27B	15.0	2.5	75.5	47.0	6.5	63.0	19.5	0.5	72.0	52.5	2.5	57.5
DS-V4-F	18.0	1.5	68.5	63.0	2.5	54.0	24.0	0.0	65.0	71.0	1.0	43.0
Q3-32B	14.5	6.0	63.0	39.5	11.0	50.5	23.0	0.0	58.5	50.5	4.5	46.5
Kimi-K2.6	20.5	3.0	78.0	46.5	7.0	66.0	27.0	0.5	74.0	50.5	6.0	60.0
Opus-4-8	15.0	8.5	86.5	9.0	22.0	77.5	18.0	1.0	77.5	28.5	12.5	75.0

Table 6: Accuracy (%) with alternative retrieval and full CDW.

Model	Knowledge supplied to the harness				
	None	Random	BM25	Contr.	CDW
Q3.8-27B	63.0	64.0	67.0	72.0	75.5
DS-V4-F	54.0	55.0	59.0	65.0	68.5
Q3-32B	50.5	51.5	54.5	59.5	63.0
Kimi-K2.6	66.0	67.0	70.0	75.0	78.0
Opus-4-8	77.5	78.5	80.5	84.5	86.5

Retrieval Controls We hold other EVD-Harness modules fixed and vary only Wiki retrieval. *Random* samples pages, while *BM25* ranks them by lexical relevance to case evidence and candidates (Robertson & Zaragoza, 2009). *Contrastive* uses the same BM25 pages to show which findings are shared and which distinguish each candidate pair. *Full CDW* follows linked pages over time and has the highest accuracy for all five models (Table 6). Further ablations appear in Appendix A5.3.

7 CONCLUSION

We introduced MEDEVM, a progressive-evidence benchmark of 1,050 cases that evaluates diagnosis submission against annotated evidence sufficiency. Across nine LLMs, we identify EVM as a misalignment between diagnosis submission timing and evidence sufficiency. Even when models recognize the correct diagnosis, evidence order and misleading evidence redirect submission decisions. Scaling and thinking do not consistently resolve this mismatch. We then validate its diagnostic relevance through prediction and intervention: EVM signals impending errors, and preventing premature submission improves accuracy. To address this mismatch, we develop EVD-HARNESS, which retrieves contrastive diagnostic knowledge to guide diagnosis proposals, then authorizes submission through independent evidence review and citation checks. Across five LLMs, it improves accuracy by 12.0–51.1 percentage points and reduces sensitivity to evidence perturbations.

8 LIMITATIONS

First, MEDEVM is built from published case reports, so its disease distribution does not represent routine clinical prevalence. Broader sampling across care settings would strengthen external validity. Second, fixed candidates and scheduled evidence enable controlled comparisons, but active questioning and open-ended diagnosis remain to be studied. Third, human review covers only a subset of the sufficiency boundary annotations t^*. Model agreement and sampled audits support annotation quality, although individual boundaries may remain imprecise and warrant broader expert review.

AI USE STATEMENT

We used AI in three parts of this work: (1) to polish the wording of the manuscript, with all initial drafts written by the authors; (2) to assist with data synthesis and filtering during benchmark construction, followed by human quality control, as detailed in Appendix A1; and (3) as the models evaluated in our experiments. We reviewed all AI-assisted work and take responsibility for the final content of this paper and its associated artifacts.

ETHICS STATEMENT

Our study of evidence–value misalignment in language models does not raise specific ethical concerns beyond data privacy. Although the benchmark draws on real clinical case reports, patient-identifying information had already been removed when these reports were published. We also checked for potentially identifying information during data screening to verify that the selected cases preserved patient privacy.

REPRODUCIBILITY STATEMENT

The appendices describe benchmark construction and all experimental settings in detail to support reproducibility. We will release the code for all experiments, including the implementation of EVD-HARNESS, to enable others to reproduce our results.

REFERENCES

Anthropic. Claude Opus 4.8 system card, June 2026a. URL https://www.anthropic.com/claude-opus-4-8-system-card/.

Anthropic. Introducing Claude Opus 4.8, May 2026b. URL https://www.anthropic.com/news/claude-opus-4-8.

Wenting Chen, Guolin Huang, Wenxuan Wang, and Zhongrui Zhu. MedEinst: Benchmarking the einstellung effect in medical LLMs through counterfactual differential diagnosis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39778–39798, 2026.

DeepSeek Team, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10183–10213, 2025.

Jinrui Fang, Runhan Chen, Xu Yang, Jian Yu, Jiawei Xu, Ashwin Vinod, Wenqi Shi, Tianlong Chen, Heng Ji, ChengXiang Zhai, et al. Benchmarking multi-turn medical diagnosis: Hold, lure, and self-correction. *arXiv preprint arXiv:2604.04325*, 2026.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. DDXPlus: A new dataset for automatic medical diagnosis. In *Advances in neural information processing systems*, volume 35, pp. 31306–31318, 2022.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Gemini Team. Gemini 3.1 Pro: A smarter model for your most complex tasks, February 2026a. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Gemini Team. Gemini 3.5 Flash model card, May 2026b. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Rebecca Handler, Suhana Bedi, and Nigam Shah. Quantifying and mitigating premature closure in frontier LLMs. *arXiv preprint arXiv:2605.15000*, 2026.

Hsin-Ling Hsu, Zizheng Wang, Donghua Zhang, Nai-Chia Chen, Jerry Wang, Jun-En Ding, Chia-Hsuan Hsu, Guoan Wang, Feng Liu, Fang-Ming Hung, et al. MedAction: Towards active multi-turn clinical diagnostic LLMs. *arXiv preprint arXiv:2605.07305*, 2026.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5684–5696, 2020.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, et al. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28858–28888, 2024.

Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. Interactive evaluation for medical LLMs via task-oriented dialogue system. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4871–4896, 2025.

Sravanthi Machcha, Sushrita Yerra, Sahil Gupta, Aishwarya Sahoo, Sharmin Sultana, Hong Yu, and Zonghai Yao. Knowing when to abstain: Medical LLMs under clinical uncertainty. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6153–6182, 2026.

Linjie Mu, Zhongzhen Huang, Yannian Gu, Shengqian Qin, Shaoting Zhang, and Xiaofan Zhang. EHRWorld: A patient-centric medical world model for long-horizon clinical trajectories. *arXiv preprint arXiv:2602.03569*, 2026.

OpenAI. Introducing GPT-5.4, March 2026a. URL `https://openai.com/index/introducing-gpt-5-4/`.

OpenAI. GPT-5.6 system card, July 2026b. URL `https://openai.com/index/gpt-5-6/`.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

WWTG Peterson, Theodore Birdsall, and William Fox. The theory of signal detectability. *Transactions of the IRE professional group on information theory*, 4(4):171–212, 1954.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026a. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Qwen Team, An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. In *arXiv preprint arXiv:2307.01928*, 2023.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

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

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Jiangyuan Wang, Xuyong Chen, Junwei He, Xu Xu, Shasha Xie, and Fuman Han. ChronoMedical-World: A medical world model for learning patient trajectories from longitudinal care data. *arXiv preprint arXiv:2605.21963*, 2026.

Jacob White. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387, 2020.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. DiagnosisArena: Benchmarking diagnostic reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 3074–3098. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.151.

A1 Detailed Construction of the MedEVM Benchmark

This appendix details the data sources, evidence and candidate construction, sufficiency annotation, and quality validation underlying MedEVM, as summarized in Section 3. We complete all item filtering and annotation before evaluating the target models and make no subsequent revisions based on the reported evaluation results.

A1.1 Case Collection and Structural Filtering

We begin with a pool of 75,434 English-language medical case reports assembled from multiple sources. These include case collections used in existing medical benchmarks, such as DiagnosisArena (Zhu et al., 2026), and journal reports acquired or crawled from PubMed-indexed sources (White, 2020). The source journals include NEJM, Cureus, BMJ Case Reports, Gastroenterology, JAMA, Neurology, The Lancet, Medicine, and Chest, among others. Every case in the final benchmark retains a case ID linking it to the original report pool and a journal-source label identifying its publication venue.

Structural filtering proceeds in three steps. First, we normalize each narrative into four fields: case information, physical examination, diagnostic tests, and final diagnosis. This step yields 52,147 structured records. Second, we check basic completeness, requiring at least 300 characters of clinical history, a chief complaint or symptoms, non-empty diagnostic-test text, and a single normalized final diagnosis. Third, we assess whether the report contains sufficient evidence to support a progressive diagnostic sequence. Each report must include at least two laboratory, imaging, pathology, or special-test findings, with evidence available at all three stages of presentation, basic workup, and advanced confirmation. Across these checks, we exclude 6,378 reports with insufficient history, 8,130 with

incomplete evidence stages, 1,437 with fewer than two tests, 94 without diagnostic-test text, and 88 without a chief complaint or symptoms. The remaining 36,020 reports proceed to evidence and option construction.

A1.2 Evidence and Option Construction

For each qualified report, we retain information available before diagnosis and remove explicit diagnosis names, treatment outcomes, and follow-up information. We then segment the remaining narrative into minimal, self-contained units of clinical evidence. Each evidence item retains an anchor to the original text, its source section, its clinical stage, and an evidence-type label. These labels cover demographics, chief complaint, history, symptoms, medication, vital signs, physical examination, laboratory tests, imaging, pathology, procedures, and special tests. We arrange the evidence items in the order of clinical disclosure in the source report. This preserves a traceable progression of available information and prevents model-generated reordering intended solely to increase task difficulty.

We convert each case into a ten-way diagnostic question comprising the confirmed diagnosis and nine clinically plausible confounders. GPT-5.4 first proposes 10–15 differential diagnoses, from which it selects nine based on clinical similarity and whether the complete case evidence can rule them out. For each confounder, the model must identify at least three shared symptoms or signs that make it plausible early in the case, specific evidence from the report that supports it, and a test or finding that ultimately excludes it. We discard candidates that are synonymous with the gold diagnosis, differ only in diagnostic granularity, lack support in the case, or cannot be ruled out by the available evidence. We remove explicit diagnostic wording from the question stem and randomize the position of the correct answer.

We next screen the constructed questions for answer uniqueness and difficulty. The complete evidence must support exactly one candidate, so we exclude items with compound diagnoses, multiple defensible answers, diagnostic leakage, or inconsistent option granularity. Multiple strong models then solve each question independently. Items that can be answered consistently from superficial cues without integrating the clinical evidence are excluded as too easy. This stage retains 7,472 usable multiple-choice questions for sufficiency-boundary annotation and quality validation.

A1.3 Sufficiency Boundary and Quality Validation

The boundary-annotation prompt contains the ten diagnostic candidates and the ordered sequence of atomic evidence. Claude Opus 4.8, Gemini 3.1 Pro Preview, and GPT-5.6-sol independently identify the earliest position at which the accumulated evidence uniquely distinguishes the correct diagnosis from all competing candidates. We retain only cases for which all three annotators identify exactly the same position and assign this position as t_i^*. This procedure operationalizes the sufficiency criterion in Equation **??**, providing a case-level reference independent of the evaluated model.

Automated quality control then reviews the source case, candidate set, and proposed boundary in a fixed sequence. We first recheck the source report for sufficient clinical detail, internal consistency, and a definitive diagnosis, removing cases with unresolved deficiencies. We next verify that the complete evidence supports a unique answer and exclude any remaining low-difficulty items that do not require evidence integration. Finally, Gemini 3.5 Flash independently validates the proposed boundary against the candidates and ordered evidence sequence. We obtain three independent judgments per item and retain it only if all three confirm that t_i^* is the earliest sufficient position. Gemini 3.5 Flash serves as a separate validator and does not contribute to the preceding three-model boundary agreement.

To assess the quality of the automated annotations, five graduate researchers with medical backgrounds independently review 200 randomly sampled items. They evaluate the coherence of each case, the uniqueness of its correct answer, the clinical plausibility of its confounders, and whether t_i^* marks the earliest sufficient boundary. Item-level inter-rater agreement reaches 92.5%, and majority voting rates 94.0% of the sampled items as accurately annotated. Across cases, the mean within-item standard deviation of the human boundary annotations is 0.84 turns, showing modest dispersion in the judged position across reviewers. We examine sensitivity to one- and two-turn tolerance windows in Appendix A3.2.

Quality control yields 1,810 valid dynamic-diagnosis items. From this pool, we select 1,050 cases through stratified sampling by disease system and empirical difficulty to form the final MEDEVM benchmark. The release covers 953 distinct diagnoses across 24 disease systems and contains 26,869 atomic evidence items. Each case includes ten diagnostic candidates and an average of 25.6 evidence items, and the mean position of the sufficiency boundary t_i^* is 21.3.

A2 METRIC DEFINITIONS

This appendix defines the metrics used to characterize EVM, compare diagnostic beliefs, and evaluate interventions. Each group is linked to the experiment it supports. For case i, let $\mathcal{Y}_i$ be the ten candidate diagnoses, y_i^* the correct diagnosis, T_i the number of evidence items, and t_i^* the annotated sufficiency boundary. A turn is indexed by the number of evidence items received. Probabilities and rates lie in $[0, 1]$ and are multiplied by 100 when reported as percentages. Differences between percentages are reported in percentage points (pp).

A2.1 DIAGNOSIS OUTCOMES AND SUBMISSION TIMING

Accuracy and submission rate. The timing analysis in RQ1 and the intervention analyses in Section 5 distinguish whether a model submits a diagnosis from whether that diagnosis is correct. Let τ_i be its first submission turn, with $\tau_i = \infty$ if it reaches the end of the evidence without submitting. Let $A_i = \mathbf{1}[\tau_i < \infty]$, and let $Z_i = 1$ for a correct submitted diagnosis and zero otherwise. For an evaluated set of N cases,

$$\mathrm{Acc}_{\mathrm{submission}} = \frac{\sum_i Z_i}{\sum_i A_i}, \qquad \mathrm{Acc}_{\mathrm{all}} = \frac{1}{N}\sum_i Z_i, \qquad \text{Submission rate} = \frac{1}{N}\sum_i A_i. \tag{A1}$$

Submission accuracy evaluates only cases with a submitted diagnosis. Accuracy over all cases, also called diagnostic success, counts both incorrect submissions and non-answers as errors. The thinking comparison in RQ2 reports submission accuracy. The misleading-evidence, Natural Gate, and harness evaluations use accuracy over all evaluated cases. The submission rate, also called answer coverage, shows how often a model produces any diagnosis. If no diagnosis is submitted, submission accuracy is undefined.

Timing relative to sufficient evidence. RQ1 assigns each trajectory to one of four categories:

$$\begin{aligned} R_{\mathrm{pre}} &= \frac{1}{N}\sum_i \mathbf{1}[\tau_i < t_i^*], & R_{\mathrm{at}} &= \frac{1}{N}\sum_i \mathbf{1}[\tau_i = t_i^*], \\ R_{\mathrm{late}} &= \frac{1}{N}\sum_i \mathbf{1}[t_i^* < \tau_i < \infty], & R_{\mathrm{none}} &= \frac{1}{N}\sum_i \mathbf{1}[\tau_i = \infty]. \end{aligned} \tag{A2}$$

These are the rates of premature submission, submission at the boundary, delayed submission, and no submission. They use all cases as the denominator and sum to one. Timing and correctness are separate properties: a correct diagnosis can be premature, and a diagnosis submitted after the boundary can be incorrect. Keeping non-answers separate also distinguishes eventual submission after sufficient evidence from waiting through the complete record.

A2.2 DIAGNOSTIC BELIEFS, ACTION PROBABILITIES, AND CALIBRATION

Candidate and action probabilities. RQ2 measures the distribution p_{it} over candidate diagnoses at evidence checkpoint t. For the default JSON-prefix measurement in either thinking mode, let $\ell_{it}(y)$ be the conditional log-probability of the candidate label in the diagnosis field, and let $u_{it}(a)$ be that of action a in the action field. Each score sums the conditional log-probabilities of the tokens in the corresponding label or action. The normalized probabilities are

$$p_{it}(y) = \frac{\exp \ell_{it}(y)}{\sum_{y' \in \mathcal{Y}_i} \exp \ell_{it}(y')}, \qquad p_{it}(\mathrm{answer}) = \frac{\exp u_{it}(\mathrm{answer})}{\exp u_{it}(\mathrm{answer}) + \exp u_{it}(\mathrm{wait})}. \tag{A3}$$

The first distribution measures which diagnosis the model favors at the fixed answer-field prefix. The second measures its propensity to submit a diagnosis at the action-field prefix. In all conditions,

scoring uses a cloned context containing the candidate list, available evidence, and preceding conversation. The candidate-scoring prefix is the same whether the original action is ANSWER or WAIT, and probe outputs are not added to the ongoing trajectory. With thinking enabled, probabilities also condition on the generated reasoning. Appendix A3.1 specifies the prefixes, and Appendix A3.3 illustrates their use in both modes.

Correct and target diagnoses. Gold belief is the probability assigned to the correct diagnosis. In RQ4, target belief is the probability assigned to the prespecified incorrect diagnosis supported by the inserted evidence:

$$p_{\text{gold},it} = p_{it}(y_i^*), \qquad p_{\text{target},it} = p_{it}(y_i^{\text{tar}}). \tag{A4}$$

The difference $p_{\text{gold}} - p_{\text{target}}$ is positive when the correct diagnosis is favored over the target. A correct top-1 prediction means that the highest-probability candidate is the correct diagnosis. Its frequency measures diagnostic ranking at a specified evidence prefix, which need not equal the accuracy of diagnoses submitted in a natural trajectory. RQ2 reports mean gold belief at submission over answered cases and at t_i^* over cases that reach that turn. These two means can therefore involve different sets of cases.

Confidence and calibration. Let $p_{it}^{(1)} \geq p_{it}^{(2)}$ be the two largest candidate probabilities. Top-1 confidence is $p_{it}^{(1)}$, and the probability margin is $p_{it}^{(1)} - p_{it}^{(2)}$. A large margin indicates a clear preference for the leading candidate, but does not establish that it is correct. The thinking comparison in RQ2 evaluates calibration using ten equal-width bins over $[0, 1]$. For n measured predictions, with bin B_m containing predictions whose top-1 confidence falls in that interval,

$$\text{ECE} = \sum_{m=1}^{10} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|. \tag{A5}$$

Here, $\text{acc}(B_m)$ is the fraction of correct top-1 predictions and $\text{conf}(B_m)$ their mean top-1 confidence. Empty bins contribute zero, and the last bin includes confidence one. Lower expected calibration error (ECE) means closer agreement between confidence and observed correctness. The thinking table pools all measured evidence checkpoints for ECE and margin, while its gold-belief column uses submission turns.

A2.3 Evidence-Order and Misleading-Evidence Effects

Distribution and diagnosis disagreement (RQ3). For candidate distributions P and Q from the two evidence orders, let $M = (P + Q)/2$. Jensen–Shannon divergence is

$$\text{JSD}(P, Q) = \frac{1}{2} \sum_y P(y) \log \frac{P(y)}{M(y)} + \frac{1}{2} \sum_y Q(y) \log \frac{Q(y)}{M(y)}. \tag{A6}$$

We use natural logarithms and take zero-probability terms to contribute zero. JSD ranges from zero to $\log 2$, with zero indicating identical distributions. RQ3 measures it at t_i^* only for pairs in which both orders reach that turn, so both distributions are based on the same evidence set.

Diagnosis outcome disagreement is the indicator that two trajectories submit different diagnoses or that only one submits a diagnosis. Two non-answers count as agreement. RQ3 reports its mean both over all evaluated pairs and over the pairs used for boundary JSD. For each metric, eligible permutations are first averaged within a source case and then across cases. This gives each source case equal weight. The candidate-label control uses the same outcome comparison after mapping labels back to diagnoses. The repeated-evidence control instead compares the next response after insertion at a shared evidence prefix, so its disagreement rate describes a local response rather than the outcome of a complete trajectory.

Effects of stronger misleading evidence (RQ4). Each comparison fixes the source case, incorrect target diagnosis, and insertion position. Write S and W for the strong and weak versions. The additional change in target belief immediately after insertion is

$$\Delta p_{\text{target}} = \left(p^{\text{S}}_{\text{target,after}} - p^{\text{S}}_{\text{target,before}}\right) - \left(p^{\text{W}}_{\text{target,after}} - p^{\text{W}}_{\text{target,before}}\right). \tag{A7}$$

The remaining belief effect is $p^{\mathrm{S}}_{\mathrm{target,final}} - p^{\mathrm{W}}_{\mathrm{target,final}}$, measured after all evidence has arrived. Target submission rate is the fraction of evaluated trajectories that submit the incorrect target, and its increase is the strong-minus-weak difference. Diagnostic accuracy loss is $\mathrm{Acc}^{\mathrm{W}}_{\mathrm{all}} - \mathrm{Acc}^{\mathrm{S}}_{\mathrm{all}}$. Correct-top-1 loss is the weak-minus-strong difference in the fraction ranking the correct diagnosis first at sequence completion. All five effects are oriented so that positive values indicate harm from stronger misleading evidence. Figure 6a uses early insertion for the first four effects and insertion after the original boundary for correct-top-1 loss.

A2.4 Error Prediction and Natural Gate Outcomes

Submission prediction and error assessment (Section 5). For next-turn submission prediction, an eligible checkpoint is a wait followed by an observed next turn. Its label r_j is one if the next action submits any diagnosis and zero if it continues waiting. For error assessment at submission, examples are first submissions, with $r_j = 1$ for an incorrect diagnosis and zero for a correct diagnosis. The regression output q_j estimates the probability of the positive outcome for the corresponding task. AUROC measures how often a positive example receives a higher score than a negative example, counting ties as one half. A value of 0.5 indicates chance ranking and one indicates perfect ranking. Log loss evaluates the probabilities:

$$\mathrm{LL} = -\frac{1}{n}\sum_{j=1}^{n}\left[r_j \log q_j + (1 - r_j)\log(1 - q_j)\right]. \tag{A8}$$

Lower log loss is better. For next-turn submission, we report $\Delta\mathrm{LL} = \mathrm{LL}_{\mathrm{M3}} - \mathrm{LL}_{\mathrm{M1}}$, so negative values favor M3. Metrics use class-balanced held-out examples. Appendix A4.1 describes the case-level splits, majority-class undersampling, and regression fitting.

Errors corrected and correct diagnoses lost (Section 5). Natural Gate pairs the original and gated outcomes of every case with an original premature submission. Let N_{pre} be the number of such cases and $Z_i^{(0)}, Z_i^{(1)}$ their original and gated success indicators. Then

$$\begin{aligned} R_{0\to1} &= \frac{1}{N_{\mathrm{pre}}}\sum_i \mathbf{1}[Z_i^{(0)} = 0, Z_i^{(1)} = 1], \\ R_{1\to0} &= \frac{1}{N_{\mathrm{pre}}}\sum_i \mathbf{1}[Z_i^{(0)} = 1, Z_i^{(1)} = 0], \\ \Delta\mathrm{Acc} &= R_{0\to1} - R_{1\to0} = \mathrm{Acc}^{(1)}_{\mathrm{all}} - \mathrm{Acc}^{(0)}_{\mathrm{all}}. \end{aligned} \tag{A9}$$

The first two rates are the percentages of errors corrected and correct diagnoses lost in Table A8. Both use all selected cases as the denominator. Their difference is the paired accuracy gain, including failures to submit under the gate. The gain over the complete benchmark is $(N_{\mathrm{pre}}/1050)\Delta\mathrm{Acc}$, with nonselected cases retaining their original outcomes. Appendix A4.3 describes the intervention and gives worked examples.

Confidence intervals. Reported paired intervals use 2,000 bootstrap resamples of source cases. All checkpoints, targets, permutations, or intervention versions belonging to a sampled case remain together, and each contrast is recomputed within the resample. The 2.5th and 97.5th percentiles give the 95% interval. For prediction, this resampling uses held-out cases and compares M1 and M3 on identical checkpoints. For interventions, it preserves the original–intervention pairing.

A3 Experimental Details and Additional Mechanism Analyses

A3.1 Shared Interaction and Probability Measurement

The four mechanism experiments in Section 4 share a sequential diagnosis protocol. The model receives ten candidate diagnoses, followed by one evidence item per turn. Each request includes the initial instructions, candidates, accumulated evidence, and previous JSON actions. It returns `{"action":"wait","answer":""}` to receive more evidence or `{"action":"answer","answer":"<option letter>"}` to submit a diagnosis. The first

submission ends the trajectory. Waiting after the final evidence item produces a non-answer. Accuracy, submission rate, and timing categories follow Appendix A2.1.

Figure A1 reproduces the system prompt used with thinking disabled and the templates for the case-opening and evidence messages. The model is instructed to answer once the available evidence supports a reliable choice. Neither the ground-truth diagnosis nor t_i^* is included in its input.

SYSTEM PROMPT

You are the doctor agent in a closed-set, multi-turn diagnosis task.
A patient agent will reveal exactly one new piece of clinical evidence per turn. After each piece, choose one action and output JSON only:

`{"action":"wait","answer":""}` or
`{"action":"answer","answer":"<option letter>"}`

Decision rule:
1. Choose "wait" only when the evidence available so far does not uniquely identify one diagnosis among the listed candidates.
2. Choose "answer" as soon as the accumulated evidence rules out all other listed candidates and uniquely identifies one diagnosis. You do not need to wait for every piece of evidence.
3. This is a closed-set candidate-selection task, not a real clinical workup. Once alternatives can be excluded with adequate confidence, answer.
4. An answer is final. Its option letter must be one of the listed candidates. When waiting, answer must be empty.
5. Use only evidence revealed in this conversation.
6. Do not provide confidence, explanation, markdown, extra keys, or chain-of-thought.

CASE-OPENING MESSAGE

A patient case needs diagnosis. There are 10 candidate diagnoses:
`A. [diagnosis A] ... J. [diagnosis J]`
The patient will now provide one new atomic piece of evidence at a time.

EVIDENCE MESSAGES

First turn: `Patient evidence [evidence ID]: [evidence text]`
Later turns: `New patient evidence [evidence ID]: [evidence text]`

Figure A1: Diagnosis interaction prompt. The system prompt is reproduced verbatim, with line wrapping adjusted for readability. Bracketed fields in the message templates are filled from the case record, and the complete candidate list is provided at the start.

The paired interventions retain the source case, candidate diagnosis texts, correct diagnosis, and generation settings. RQ3 changes evidence order, while RQ4 inserts repeated, irrelevant, or misleading evidence. A supplementary control changes candidate labels. Each design below specifies whether it evaluates a natural trajectory or a reconstructed evidence prefix.

Conversation state for measurement. At each measured turn, we copy the current conversation containing the candidate list, all evidence received so far, and the preceding actions. In this cloned context, we fix the scoring prefix and evaluate candidate continuations. The same procedure applies at diagnosis-submission and waiting turns. Neither the scoring prefix nor its continuations are added to the ongoing conversation. The original action therefore determines whether the diagnostic trajectory ends or receives further evidence.

Candidate probabilities. Candidate scoring uses the fixed diagnosis prefix `{"action":"answer","answer":"` in all conditions. At this position, we score all ten candidate labels as alternative continuations of the same context and normalize them using Equation A3. For a multi-token label, its log-probability sums the conditional log-probabilities of its tokens. With thinking enabled, the scoring context also contains the reasoning generated before the JSON response.

Action probabilities. Action scoring uses the shorter prefix `{"action":"`. We score the complete continuations `wait` and `answer` and normalize over the two actions as in Equation A3. The resulting $p(\text{answer})$ measures submission propensity, while the candidate distribution measures diagnostic belief over the ten diagnoses at the same evidence prefix.

Each model is evaluated on the same 1,050 cases. Paired confidence intervals use the case-bootstrap procedure in Appendix A2.4. Natural measurements end at the first diagnosis submission, so later-turn summaries identify which trajectories remain available.

A3.2 Diagnosis Submission Timing and Belief at the Sufficiency Boundary

In RQ1, each case follows its original evidence order. Figure 3 reports the four submission-timing categories defined in Appendix A2.1, separately from diagnostic correctness.

In RQ2, we measure candidate probabilities at each evidence checkpoint and summarize p_{gold} at submission and at t_i^*. The submission summary includes answered trajectories. The boundary summary includes trajectories that reach t_i^*, whether they answer or wait at that turn. Figure 4 therefore compares belief at actual submission with belief upon receiving sufficient evidence, using the cases available at each point, which may differ because early submissions end before sufficient evidence arrives.

Sensitivity to boundary tolerance. The human review reports a mean within-case standard deviation of 0.84 turns for boundary judgments (Appendix A1). We therefore examine whether the RQ1 patterns depend on an exact turn match. Keeping each annotated t_i^* as the reference, we rescore the existing trajectories with tolerance $k \in \{0, 1, 2\}$. A submission is premature if $\tau_i < t_i^* - k$, aligned if $|\tau_i - t_i^*| \leq k$, and delayed if $\tau_i > t_i^* + k$. Unanswered cases remain a separate category. This changes the timing score without changing the evidence, model responses, or diagnostic correctness. The window accommodates uncertainty in the annotation rather than asserting that evidence before t_i^* is sufficient.

Table A1: RQ1 timing sensitivity to a tolerance of k turns around the annotated boundary. All rates use the same 1,050 cases per model. Non-answer rates are unchanged and omitted.

	Premature (%)			Within boundary (%)			Delayed (%)		
Model	$k=0$	$k=1$	$k=2$	$k=0$	$k=1$	$k=2$	$k=0$	$k=1$	$k=2$
Q3-8B	99.1	97.7	96.0	0.6	2.2	3.9	0.3	0.1	0.1
Q3-14B	78.6	70.3	63.7	17.6	27.1	34.2	2.7	1.4	1.0
Q3-32B	35.3	27.0	21.6	37.0	48.5	55.7	9.6	6.5	4.7
Q3-235B	49.7	40.2	33.5	38.3	49.9	58.6	6.8	4.7	2.7
Q3.6-27B	41.0	34.8	28.7	42.7	51.9	59.0	6.4	3.4	2.5
Q3.8-27B	36.6	26.8	21.0	45.6	59.0	66.7	8.9	5.2	3.4
DS-V4-F	24.6	20.8	17.4	27.0	35.0	41.5	11.0	6.8	3.6
K2.5	15.8	11.6	9.4	43.7	50.7	54.3	6.8	4.0	2.6
K2.6	17.0	12.9	10.7	44.3	51.6	55.3	7.9	4.7	3.1

Allowing one turn increases the aligned fraction by 1.6–13.4 percentage points across models, and allowing two turns increases it by 3.3–21.0 points (Table A1). These increases reflect the more permissive scoring rule. They do not remove the main behavioral differences: Qwen3-8B still submits prematurely in 96.0% of cases at $k = 2$, while DeepSeek and the Kimi models retain non-answer rates of 30.9–37.4%. The tolerance reduces both premature and delayed classifications, but substantial timing mismatch or non-response remains for every model. We retain $k = 0$ for the other reported experiments so that their evidence checkpoints and intervention positions share the same annotated reference.

Applying the same tolerance to the thinking trajectories yields premature-submission rates of 90.8–99.1% at $k = 1$ and 85.5–98.4% at $k = 2$. Each remains above the corresponding non-thinking rate. Thus, the increase in premature submission with thinking is not explained solely by decisions one or two turns away from the annotated boundary.

A3.3 THINKING-MODE EVALUATION AND BELIEF MEASUREMENT

The thinking comparison supplements RQ1 and RQ2 for seven models: four Qwen3 sizes, Qwen3.6-27B, Qwen3.8-27B, and DeepSeek-V4-Flash. Both modes use the same cases, candidates, evidence, temperature, and seed. With thinking enabled, the prompt permits internal reasoning but still requires a final action JSON and follows the same stopping rule.

Fixed-prefix measurement in both modes. In both modes, candidate probabilities are scored after `{"action":"answer","answer":"` in the cloned context. With thinking enabled, the generated reasoning precedes this fixed prefix. An illustrative continuation is

```
<think>[reasoning]</think>
{"action":"answer","answer":"C"}
```

At the position occupied by C, we score all candidate labels A–J and normalize over those labels. The distribution is conditioned on the cloned conversation, the generated reasoning, and the fixed JSON prefix. With thinking disabled, the same procedure scores the labels without a preceding reasoning segment.

Candidate measurement when the model waits. A natural WAIT response need not contain an answer field. To measure diagnostic belief at that turn, we supply the same fixed answer-field prefix in the cloned context and score all ten candidate labels. This procedure also applies to thinking-enabled waiting turns, where the generated reasoning precedes the scoring prefix. It does not introduce a new instruction asking the model to select a diagnosis. The original WAIT action remains unchanged, so the diagnostic trajectory continues to receive evidence.

All measurements use token probabilities for the candidate alternatives rather than a numerical confidence statement generated by the model. For illustration, if label C has probability mass 0.24 and the ten labels together have mass 0.60, its normalized probability is $p(C) = 0.40$. Submission and waiting checkpoints use the same scoring prefix. With thinking enabled, gold belief and ECE describe predictions conditioned on the generated reasoning.

Table A2 uses the metrics in Appendix A2.2. Early submission includes all completed trajectories, submission accuracy includes answered trajectories, and p_{gold} is measured at the diagnosis field during submission. Margin and ECE pool all measured evidence checkpoints. Only 3–51 cases per model reach t_i^* with thinking enabled, after early submissions have ended the other trajectories.

Table A2: Thinking-mode results. All rates, probabilities, margins, and ECE are multiplied by 100. Off/on denotes the thinking setting. p_{gold} is measured at diagnosis submission, whereas margin and ECE use all measured evidence checkpoints. n_{t^*} counts boundary measurements with thinking on.

	Early submission		Submission accuracy		Belief measurements (on)			
Model	Off	On	Off	On	p_{gold}	Margin	ECE	n_{t^*}
Q3-8B	99.1	99.7	10.9	11.3	11.5	100.0	89.8	4
Q3-14B	78.6	99.7	30.5	12.0	10.4	99.8	90.1	5
Q3-32B	35.3	99.6	59.2	13.9	14.1	99.3	87.9	4
Q3-235B	49.7	99.6	55.4	17.3	15.9	99.7	86.5	3
Q3.6-27B	41.0	94.9	70.2	37.8	35.6	100.0	76.1	50
Q3.8-27B	36.6	94.1	72.9	36.8	41.0	100.0	74.4	51
DS-V4-F	24.6	97.3	70.5	28.5	16.2	35.7	39.7	26

Thinking raises premature submission to 94.1–99.7% across the seven models. Qwen3-8B already submits prematurely on 99.1% of cases with thinking off, and its submission accuracy rises slightly from 10.9% to 11.3%. Accuracy falls for the other six models. DeepSeek-V4-Flash, for example, moves from 24.6% to 97.3% premature submission while accuracy falls from 70.5% to 28.5%. All seven models show lower submission-point gold belief and higher ECE. The large probability margins in the Qwen variants therefore reflect confident preferences that are often incorrect.

Table A3: Evidence-order comparisons for pairs in which both trajectories reach t_i^*. Source cases and order pairs count distinct cases and eligible original–permuted comparisons, respectively. JSD and final diagnosis disagreement use the same pairs and equal source-case weights. Brackets give 95% bootstrap confidence intervals. All order pairs from a sampled case remain together.

Model	Source cases	Order pairs	JSD $\times 100$	Final diagnosis disagreement (%)
Q3-8B	7	17	2.12 [0.02, 5.98]	3.6 [0.0, 10.7]
Q3-14B	76	191	1.58 [0.58, 2.71]	6.2 [1.6, 11.8]
Q3-32B	635	2120	4.36 [3.66, 5.07]	15.4 [13.1, 17.9]
Q3-235B	353	955	2.87 [2.11, 3.77]	13.1 [9.9, 16.4]
Q3.6-27B	545	1678	1.28 [1.04, 1.54]	13.8 [11.3, 16.6]
Q3.8-27B	420	1091	0.90 [0.72, 1.11]	13.2 [10.3, 16.3]
DS-V4-F	780	2721	1.48 [1.22, 1.75]	20.6 [18.4, 22.9]
K2.5	832	2753	1.68 [1.42, 1.95]	18.8 [16.6, 21.1]
K2.6	807	2629	1.66 [1.36, 1.98]	18.3 [16.0, 20.6]

A3.4 Evidence Order: Comparing Final Diagnoses and Boundary Beliefs

RQ3 retains each original sequence and constructs up to four distinct permutations of the evidence before t_i^*. Using a fixed seed, we first shuffle within clinical stages and then shuffle the full prefix if needed to obtain additional distinct orders. The evidence at t_i^* and afterward remains fixed. Trajectories reaching the boundary have therefore received identical evidence sets in different orders.

We compare final diagnosis outcomes for every original–permuted pair, including early-terminated trajectories. We separately compare pairs with candidate probabilities available in both orders at t_i^*, using each trajectory's own accumulated conversation. For these pairs, we compute boundary JSD and eventual diagnosis disagreement on the same cases. Both metrics follow the aggregation in Appendix A2.3, so a case with more eligible permutations receives no additional weight.

Reordering changes 29.0–52.9% of final outcomes across the full cohort. Among pairs reaching the boundary, models other than Qwen3-8B have mean JSD values of 0.0090–0.0436 but still disagree on 6.2–20.6% of final outcomes (Table A3). Similar beliefs after the same evidence thus coexist with different eventual diagnoses. Qwen3-8B contributes only seven boundary-observed source cases because it usually submits earlier, producing a wide confidence interval.

A3.5 Repeated Evidence, Irrelevant Text, and Candidate Labels

The first evidence-value intervention in RQ4 tests whether text without new diagnostic information changes the model's response. We select prefixes ending at approximately 25%, 50%, and 75% of the evidence sequence, at t_i^*, and one turn before the original natural diagnosis submission when available. Prefixes contain at least one item, and coincident positions are merged. At each position, two branches share the same evidence prefix. One repeats its first, middle, or most recent evidence item. The other adds a clinically irrelevant statement, such as "The weather is sunny today." Neither supplies a new clinical fact that helps distinguish among the candidate diagnoses.

To measure the response at each selected position, we reconstruct the prefix in a separate conversation, recording WAIT after each preceding evidence item. We then append the repeated evidence or irrelevant statement and let the model choose its next action. We measure beliefs before and after the addition and compare the diagnosis outputs of the two branches immediately afterward. A difference occurs if the model submits different diagnoses or submits a diagnosis in one branch while waiting in the other. These are local responses to the inserted text at a common evidence prefix.

Diagnosis outputs differ in 5.5–34.4% of comparisons (Figure A2), with the largest effect in Qwen3-8B. Although neither addition supplies new diagnostic information, the choice of text changes whether the model submits a diagnosis and which diagnosis it submits.

A separate control tests candidate-label sensitivity. For each case, four fixed-seed permutations reassign the ten letters to the existing diagnosis rows while preserving diagnosis text, row order, and evidence. For example, a diagnosis originally labeled A may receive F in the same row. We run each version naturally and map submitted labels back to diagnoses before comparison. Final outcomes

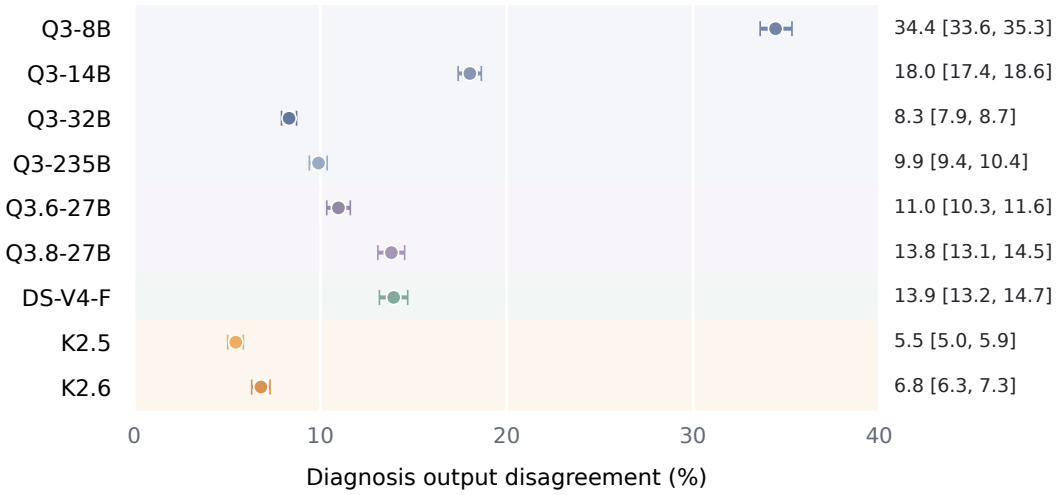


Figure A2: Diagnosis output differences immediately after repeated evidence or irrelevant text is added to the same evidence prefix. Points show the percentage of comparisons with different submitted diagnoses or a diagnosis in only one branch. Intervals are source-case bootstrap 95% confidence intervals, with comparisons averaged within each case.

differ from the original-label run in 9.7–28.8% of cases, including diagnosis-versus-non-answer differences. The identifiers alone can therefore influence the final outcome.

A3.6 Misleading Evidence: Construction, Measurements, and Results

The strength intervention in RQ4 examines how evidence favoring an incorrect diagnosis changes beliefs and subsequent submissions. We first synthesize misleading evidence for the 1,050 cases in MedEVM. For each case, an incorrect candidate serves as the target diagnosis. We construct weak and strong observations that favor this target while matching their clinical event type, semantic direction, and presentation format. Each version is inserted early in the sequence, immediately before the original sufficient-evidence item, or immediately after it. The source case, target, and insertion position remain fixed within a strong–weak pair.

Before forming the test set, Claude Opus 4.8, Gemini 3.1 Pro Preview, and GPT-5.6-sol independently review the augmented evidence sequences. The reviewers judge whether the original decisive evidence still marks the earliest point at which the correct diagnosis can be distinguished from the candidates. When insertion precedes that item, its turn index shifts forward by one. All three reviewers agree that the boundary remains valid for 553 cases, from which we manually select 318 for the intervention experiment. The construction review also checks that inserted findings are clinically coherent, temporally possible, and free of information from later evidence.

Illustrative boundary checks. The following two high-strength examples show the complete evidence available through the original sufficient-evidence item, together with the inserted distractor. Both distractors support a competing diagnosis without replacing the source-report evidence that identifies the correct diagnosis.

In the first case, hearing loss, tympanic changes, and the inserted flat tympanogram make middle-ear effusion plausible before evidence E9. The histopathologic finding of branched fungal hyphae is the first observation that identifies the fungus ball among the candidates. The insertion therefore changes the turn number of that finding from 9 to 10 without changing the correct diagnosis or the decisive evidence. In the second case, cervical inflammation may coexist with the presenting symptoms, but it arrives after the hysterosalpingogram has shown contrast entering the bowel. It does not explain that finding or move the earliest sufficient turn away from evidence E15.

Before the boundary: middle-ear fungus ball `33872bd24da49e2633decfbe0a4d0186`

Correct diagnosis: middle-ear fungus ball. **Incorrect target:** otitis media with effusion due to Eustachian tube dysfunction.

Evidence through the original $t^* = 9$ (E1–E9):
E1 The patient was a 56-year-old Thai man.
E2 He presented with gradually increasing hearing loss in the right ear for 1 year.
E3 He had undergone right tympanoplasty for chronic suppurative otitis media 21 years earlier, resulting in cure with good hearing.
E4 Otoscopic examination showed localized retraction of the tympanic membrane in the posterosuperior quadrant.
E5 The remainder of the tympanic membrane was dull and thick.
E6 Audiological examination showed conductive hearing loss with a 55 dB speech reception threshold, 40 dB bone conduction level, and 30 dB air-bone gap.
E7 Exploratory tympanotomy revealed a muddy clay-like substance in the middle-ear cavity mixed with thick brown fluid, along with debris and granulation.
E8 The ossicular chain appeared intact during exploratory tympanotomy.
Insert The right-ear tympanogram shows a flat type B pattern with normal ear-canal volume, supporting middle-ear fluid accumulation from impaired ventilation.
E9 Histopathology showed a tangled mass of branched hyphae surrounded by numerous neutrophils.

Boundary after insertion: original E9 appears at augmented turn 10.

After the boundary: salpingo-enteric fistula `085b81c229b8accdee45113642cb7959`

Correct diagnosis: salpingo-enteric fistula. **Incorrect target:** chronic cervicitis.

Evidence through the original $t^* = 15$ (E1–E15):
E1 The patient was a 34-year-old Para 0+1 woman.
E2 She presented for routine hysterosalpingographic evaluation of secondary infertility after eight years of marriage without achieving pregnancy.
E3 She reported intermittent lower abdominal pain.
E4 She reported occasional intermittent vaginal discharge.
E5 She had a history of chronic pelvic inflammatory disease.
E6 She had an uneventful termination of a seven-week pregnancy two years before marriage.
E7 She underwent myomectomy three years before presentation.
E8 Hysterosalpingography performed before the myomectomy showed fallopian tubes of normal caliber that were patent bilaterally.
E9 She underwent appendicectomy approximately 14 months before presentation.
E10 She had completed oral ciprofloxacin and metronidazole therapy.
E11 Physical examination showed copious vaginal discharge.
E12 Initial hysterosalpingography opacified a moderate-capacity uterus with regular margins.
E13 The left fallopian tube did not opacify on hysterosalpingography.
E14 The right fallopian tube was visualized with a triangular loculation of contrast at its distal aspect.
E15 Subsequent hysterosalpingographic films showed contrast opacification of the large bowel, with haustrations visible within the opacified bowel.
Insert The cervix has marked erythema, friability, and mucopurulent endocervical discharge.

Boundary after insertion: original E15 remains at augmented turn 15.

First, measure beliefs at specified evidence positions. For each intervention version, we construct separate conversations containing the candidate list and accumulated evidence up to four positions: immediately before insertion, immediately after insertion, upon receipt of the original sufficient-evidence item, and at the end of the complete sequence. These measurement conversations contain no intervening diagnosis actions. This allows us to examine all four evidence positions for every intervention version, even when a natural trajectory would answer earlier. For insertion after the boundary, the boundary measurement precedes insertion. The positions are evaluated in their actual sequence order, preserving the clinical history available at each measurement.

At each position, we record p_{target}, p_{gold}, their difference, and the highest-probability candidate, together with submission propensity $p(\text{answer})$. The belief metrics and strong-minus-weak contrasts are defined in Appendix A2.3. Measurements immediately before and after insertion quantify the initial response, while the final prefix measures the effect remaining after all original evidence has arrived, including the evidence that establishes the correct diagnosis.

Second, evaluate natural diagnosis submissions. Each strong and weak version also runs as a separate natural trajectory, with the model free to wait or submit after each item. The first submission ends the interaction. We record target submissions, submission timing, correctness, and non-answers. These outcomes link the measured belief shifts to actual diagnosis submissions, including failures to submit after the complete evidence sequence has been revealed.

Third, summarize the effects of stronger misleading evidence. Table A4 gives the five effects defined in Appendix A2.3, following the column order of Figure 6a. The first four use early insertion: the immediate target-belief increase, target-submission increase, diagnostic accuracy loss, and final target-belief increase. The fifth uses insertion after the original boundary and reports final correct-top-1 loss. All effects are expressed in percentage points, with positive values indicating harm.

Table A4: Effects of stronger misleading evidence across models (percentage points). The first four endpoints use early insertion, and the last uses insertion after t_i^*. All positive values indicate harm. Brackets give source-case bootstrap 95% confidence intervals. Final diagnostic accuracy counts non-answers as errors. Bold values mark the smallest observed harmful effect in each column.

	Early insertion				After t^*
Model	Target belief at insertion ↑	Target diagnosis submission ↑	Diagnostic accuracy ↓	Target belief at completion ↑	Correct diagnosis ranked first ↓
Q3-8B	**11.1** [9.3, 13.1]	**3.0** [1.7, 4.5]	**1.0** [0.2, 1.9]	**4.4** [3.4, 5.5]	**3.9** [2.2, 5.8]
Q3-14B	18.2 [15.9, 20.6]	21.7 [18.6, 24.8]	7.2 [4.9, 9.7]	9.1 [7.4, 10.8]	6.0 [4.1, 8.0]
Q3-32B	17.7 [15.7, 19.7]	24.0 [21.2, 27.0]	12.3 [9.6, 15.1]	10.8 [9.1, 12.6]	11.1 [8.7, 13.5]
Q3-235B	18.4 [16.0, 20.8]	33.2 [30.1, 36.4]	19.6 [16.5, 22.7]	9.9 [8.3, 11.8]	11.9 [9.3, 14.5]
Q3.6-27B	24.6 [22.8, 26.3]	35.0 [31.8, 38.5]	26.7 [23.1, 30.0]	8.6 [7.4, 9.9]	16.6 [14.1, 19.2]
Q3.8-27B	20.3 [18.7, 21.9]	34.2 [31.2, 37.4]	25.1 [21.8, 28.5]	6.1 [5.1, 7.2]	12.2 [9.8, 14.9]
DS-V4-F	20.2 [18.2, 22.2]	29.6 [26.4, 32.9]	10.4 [7.5, 13.2]	7.7 [6.1, 9.4]	14.1 [11.5, 16.8]
K2.5	21.1 [19.1, 23.1]	33.5 [30.3, 36.7]	18.6 [15.4, 21.6]	10.8 [9.3, 12.4]	17.9 [15.2, 20.5]
K2.6	24.3 [22.2, 26.3]	31.4 [28.4, 34.7]	16.9 [13.9, 20.1]	12.5 [10.8, 14.2]	15.3 [12.6, 17.9]

All five effects are positive in every model, with confidence intervals excluding zero. Stronger evidence adds 11.1–24.6 points to the immediate increase in target belief. For models other than Qwen3-8B, target submission rises by 21.7–35.0 points and diagnostic accuracy falls by 7.2–26.7 points. After all original evidence arrives, target belief remains 4.4–12.5 points higher. Insertion after the original boundary reduces final correct-top-1 frequency by 3.9–17.9 points. The influence of stronger misleading evidence therefore persists in both beliefs and submitted diagnoses.

Qwen3-8B shows smaller behavioral changes: target submission increases by 3.0 points and accuracy falls by 1.0 point, despite an 11.1-point additional belief shift. It submits very early under both strengths, leaving fewer trajectories whose outcomes can change further.

Interpreting insertion before and after the boundary. Within each insertion position, strong and weak evidence are compared against the same clinical history. The median additional target-belief increase across models is 20.2 points for early insertion, 24.1 immediately before the boundary, and 26.1 immediately afterward (Figure 6b). Beliefs remain responsive to misleading evidence after the original evidence becomes sufficient. Each insertion position has a different evidence history.

A4 Additional Diagnostic-Impact Analyses

This appendix separates the prediction of diagnosis submission from the assessment of errors at submission, then details the Natural Gate intervention in Section 5. It also examines diagnostic

beliefs after prematurely stopped cases receive sufficient evidence. Metric definitions appear in Appendix A2.

A4.1 Separating Submission Prediction from Error Assessment

Two prediction targets. We use the existing non-thinking trajectories and probability measurements for all nine LLMs from Sections 3 and 4. No additional diagnostic trajectories are generated. The first task asks whether a model that currently waits will submit a diagnosis on the next turn. The second asks whether a diagnosis is incorrect when the model submits it. Separating these targets allows us to test whether EVM features carry information about diagnostic correctness in addition to their ability to predict submission behavior.

For next-turn submission prediction, an eligible example is a checkpoint at which the model chooses WAIT and a subsequent turn is observed. Inputs are measured through the current turn t, and the label is one when the action at $t + 1$ is ANSWER, regardless of which diagnosis is submitted. A further WAIT gives label zero. For example, waiting at turn 8 and submitting at turn 9 produces a positive example at turn 8, whether the diagnosis is correct or incorrect. Before balancing, this support contains 163,551 checkpoints and 7,941 positive examples. The submission frequency varies from 3.1% to 20.0% across LLMs, motivating the class-balanced comparison below.

For error assessment at submission, each trajectory contributes its first diagnosis submission, yielding 7,942 examples before balancing. Inputs are measured at that submission turn, and the label is one if the submitted diagnosis differs from the ground truth. Correct submissions have label zero. Cases without a submitted diagnosis do not enter this task. Thus, this experiment evaluates the ability to distinguish incorrect from correct submissions using the state at submission. The preceding experiment separately measures prediction of future behavior from a waiting checkpoint.

Shared predictor inputs. The feature definitions are the same for both tasks, with measurements taken at their respective prediction checkpoints. The progress baseline M0 uses the fraction of evidence received, t/T_i, the total evidence count T_i, and the number of candidates. M1 adds diagnostic confidence: the highest candidate probability, the difference between the two highest probabilities, and the entropy of the candidate distribution. These quantities summarize the certainty and concentration of the diagnostic belief distribution without identifying the correct candidate.

M2 supplements M0 with the original EVM features. These describe diagnosis submission propensity, its relationship to diagnostic confidence, and changes in beliefs across consecutive turns. Submission propensity is the probability of choosing ANSWER rather than WAIT under the action probe. Its turn-to-turn change captures whether the model becomes more inclined to submit as new evidence arrives. The remaining features compare submission propensity with confidence and entropy and summarize fluctuations in confidence, probability margin, and entropy. At the first turn, change features are set to zero. M3 combines the inputs of M1 and M2. Neither M1 nor M3 uses the sufficiency annotation, submission lead time, or the identity of the correct diagnosis.

For next-turn submission, M_{prop} uses only submission propensity to examine how much predictive information is already present in the action probability. For submission errors, we additionally report M_{timing}, which adds the annotated lead time $t_i^* - \tau_i$ to M0, where τ_i is the first submission turn. An oracle reference adds the probability of the ground-truth diagnosis to M1. These two references quantify the information provided by the annotated boundary and answer key, respectively. Their privileged inputs are separate from the M1–M3 comparison.

For completeness, let u_t be submission propensity, c_t the highest candidate probability, m_t the margin between the two highest probabilities, and H_t the candidate entropy, suppressing the case index. With $\Delta x_t = x_t - x_{t-1}$, the six EVM features are

$$(u_t,\ u_t m_t,\ \Delta u_t,\ u_t(1 - c_t),\ u_t H_t,\ |\Delta c_t| + |\Delta m_t| + |\Delta H_t|)\,. \tag{A10}$$

The products describe how the inclination to submit interacts with diagnostic certainty or uncertainty. The final term summarizes how much the belief distribution changes as evidence arrives. These features have the same definitions at waiting checkpoints and submission turns, allowing the two tasks to evaluate the same information under different targets.

Case splits and class balancing. Case identifiers are assigned deterministically to 20% training, 20% validation, and 60% test partitions using seed 2027. All checkpoints belonging to a case remain in the same partition, and the assignment is shared across predictor specifications. Within each partition and task, we then undersample the majority class to match the minority class using the same seed. Examples are neither duplicated nor moved between partitions. Each retained partition therefore contains equal numbers of positive and negative examples, and competing predictors are evaluated on the same retained examples.

Balancing is applied separately to submission versus waiting in the first task and incorrect versus correct submissions in the second. This prevents the majority class from dominating the reported probability losses and provides a common class proportion across models. The reported log losses consequently evaluate probabilities under this balanced sampling distribution. The original checkpoint counts describe the available support before undersampling, rather than the size of the balanced test sets.

Regression fitting and evaluation. We fit a separate L2-regularized logistic regression for each LLM, prediction task, and feature specification. Continuous inputs are standardized using training-partition statistics. The penalty is selected from $\{0.01, 0.1, 1, 10\}$ by validation log loss, after which the predictor is refit using the training and validation partitions with the original standardization. Testing uses held-out cases. Pooled regressions combine the LLMs and add model-identity indicators while retaining the case-level split.

AUROC evaluates how well predicted scores distinguish positive from negative examples. For next-turn submission, positives are upcoming submissions. For error assessment, positives are incorrect submitted diagnoses. Log loss evaluates the probability assigned to the realized binary outcome, as defined in Equation A8. For next-turn submission, we report $\Delta\mathrm{LL} = \mathrm{LL}_{M3} - \mathrm{LL}_{M1}$, where negative values favor M3. The balanced evaluation reports held-out point estimates for both prediction tasks.

Table A5: Balanced next-turn submission prediction. AUROC compares upcoming diagnosis submissions with continued waiting. M3 combines confidence and EVM features, and Δ is its improvement over M1.

Model	M0	M1	M2	M_{prop}	M3	Δ(M3−M1)
Qwen3-8B	0.791	0.795	0.812	0.838	0.820	+0.025
Qwen3-14B	0.778	0.780	0.806	0.824	0.827	+0.047
Qwen3-32B	0.830	0.826	0.846	0.806	0.846	+0.020
Qwen3-235B-A22B	0.831	0.827	0.844	0.772	0.840	+0.013
Qwen3.6-27B	0.813	0.815	0.821	0.765	0.812	-0.003
Qwen3.8-27B	0.850	0.857	0.877	0.843	0.885	+0.028
Kimi-K2.5	0.850	0.851	0.855	0.750	0.842	-0.009
Kimi-K2.6	0.868	0.863	0.877	0.771	0.873	+0.010
DeepSeek-V4-Flash	0.792	0.790	0.812	0.730	0.807	+0.017
Pooled	0.833	0.835	0.849	0.735	0.854	+0.019

Predicting submission behavior. M3 improves AUROC over M1 for seven of nine LLMs, with pooled AUROC increasing from 0.835 to 0.854 (Table A5). Improvements are not uniform. Qwen3.6-27B and Kimi-K2.5 show small declines, while submission propensity alone is competitive for Qwen3-8B. These comparisons indicate that the relation between action probabilities, changing beliefs, and subsequent submission differs across models. Combining these signals is generally useful, but it does not consistently dominate every simpler predictor.

The probability estimates also improve for most models. M3 reduces log loss relative to M1 for eight LLMs, and pooled log loss falls from 0.1556 to 0.1494 (Table A6). Kimi-K2.5 is the exception on both metrics. The overall pattern supports the use of EVM features to anticipate submission behavior beyond evidence progress and candidate confidence.

Table A6: Log loss for balanced next-turn submission prediction. Lower values indicate better submission-probability estimates, and negative M3−M1 differences favor the addition of EVM features.

Model	M0	M1	M2	M_{prop}	M3	Δ(M3−M1)
Qwen3-8B	0.4369	0.4340	0.4179	0.4575	0.3985	-0.0355
Qwen3-14B	0.2175	0.2166	0.2104	0.2395	0.2041	-0.0125
Qwen3-32B	0.1428	0.1434	0.1392	0.1653	0.1391	-0.0043
Qwen3-235B-A22B	0.1682	0.1693	0.1645	0.1987	0.1635	-0.0058
Qwen3.6-27B	0.1620	0.1611	0.1600	0.1837	0.1600	-0.0011
Qwen3.8-27B	0.1472	0.1453	0.1418	0.1729	0.1378	-0.0075
Kimi-K2.5	0.1135	0.1139	0.1139	0.1378	0.1145	+0.0006
Kimi-K2.6	0.1194	0.1194	0.1185	0.1426	0.1186	-0.0008
DeepSeek-V4-Flash	0.1195	0.1193	0.1180	0.1352	0.1163	-0.0030
Pooled	0.1562	0.1556	0.1527	0.1780	0.1494	-0.0062

Table A7: Balanced assessment of incorrect diagnoses at submission, measured by AUROC. M1 and M3 use current and preceding model measurements. The timing and oracle references additionally use the annotated boundary and correct-answer identity, respectively.

Model	M0	M1	M2	M_{timing}	M3	Oracle	Δ(M3−M1)
Qwen3-8B	0.625	0.640	0.637	0.618	0.724	1.000	+0.084
Qwen3-14B	0.659	0.627	0.635	0.731	0.822	1.000	+0.195
Qwen3-32B	0.665	0.711	0.724	0.750	0.841	1.000	+0.130
Qwen3-235B-A22B	0.674	0.789	0.788	0.773	0.847	1.000	+0.058
Qwen3.6-27B	0.698	0.783	0.790	0.809	0.777	0.999	-0.006
Qwen3.8-27B	0.630	0.773	0.787	0.751	0.789	1.000	+0.016
Kimi-K2.5	0.708	0.800	0.821	0.812	0.800	1.000	+0.000
Kimi-K2.6	0.675	0.711	0.726	0.782	0.723	0.998	+0.012
DeepSeek-V4-Flash	0.719	0.814	0.813	0.816	0.839	0.986	+0.025
Pooled	0.807	0.845	0.851	0.855	0.879	1.000	+0.034

Assessing errors at submission. M3 improves AUROC for seven LLMs, matches M1 for Kimi-K2.5 at the reported precision, and slightly reduces AUROC for Qwen3.6-27B (Table A7). Pooled AUROC increases from 0.845 to 0.879. The largest gains occur for Qwen3-14B and Qwen3-32B, showing that submission-related information can distinguish incorrect from correct diagnoses even after restricting the support to actual submissions.

The timing reference is informative, but its ranking relative to M3 varies across models. The answer-key oracle achieves near-perfect discrimination, indicating that knowing which candidate is correct supplies information far beyond ordinary confidence statistics. Neither reference is needed for the reported M3 gains. Together, the two tasks support distinct conclusions: EVM features help anticipate when a model will submit, and they help assess correctness at the point of submission. The Natural Gate experiment below then tests whether changing submission behavior can improve diagnosis outcomes.

A4.2 Diagnostic Beliefs After Receiving Sufficient Evidence

This experiment asks whether a model that submits a diagnosis prematurely can identify the correct diagnosis once the evidence reaches the annotated sufficiency boundary. It connects premature stopping to the information the model would otherwise miss and motivates the Natural Gate intervention below.

For each LLM, we select cases whose original natural trajectory ends with a diagnosis submission at turn $s_i < t_i^*$. Here, t_i^* is the earliest annotated turn at which the accumulated evidence is sufficient to identify the correct diagnosis among the ten candidates. For each selected case, we construct three evidence prefixes in the original order: the prefix ending at the original submission turn s_i, the prefix

ending immediately before the boundary at $t_i^* - 1$, and the prefix ending at the boundary t_i^*. If the original submission occurs at $t_i^* - 1$, the first two prefixes coincide.

Each prefix is evaluated in a separate context containing the candidate list and the evidence through that checkpoint. The original submitted diagnosis is not included in these contexts. We measure the candidate probability distribution using the same candidate-scoring procedure as in Appendix A3.1, and record both the highest-probability diagnosis and p_{gold}, the probability assigned to the correct diagnosis. The correct answer is used only to score the resulting distribution. These measurements assess the model's diagnostic beliefs at each evidence prefix. The following Natural Gate experiment separately measures the diagnosis the model actually submits when allowed to continue.

We compare the two adjacent prefixes at $t_i^* - 1$ and t_i^* to measure the change associated with receiving the evidence item that reaches the sufficiency boundary. Across the nine LLMs, the mean increase in p_{gold} is 0.245–0.333. Among cases with an incorrect premature submission in the original trajectory, 49.2–78.4% assign the highest probability to the correct diagnosis after receiving evidence through t_i^*. Thus, for a substantial fraction of premature errors, the model can recognize the correct diagnosis from evidence that its original stopping decision prevented it from seeing.

We also examine the subset whose original submission occurs exactly at $t_i^* - 1$. The fraction with a correct highest-probability diagnosis at t_i^* exceeds the original diagnosis accuracy by 19.7–33.3 percentage points across models. This comparison makes the missed opportunity concrete: even when the next evidence item is sufficient to resolve the case, premature submission can end the trajectory immediately before it arrives.

A4.3 Natural Gate: Delaying Submission Until Evidence Is Sufficient

Purpose and case selection. Natural Gate tests whether allowing models to receive sufficient evidence improves their final diagnosis outcomes, as reported in Section 5. For each LLM, the intervention includes every case in which the original natural run submits a diagnosis before t_i^*. Selection depends on submission timing, so it includes both correct and incorrect premature diagnoses. Each selected case contributes a pair of outcomes: its original diagnosis outcome and its outcome under the gate. The candidate diagnoses, correct answer, evidence sequence, and generation configuration are held fixed within each pair.

How the gate operates. We rerun each selected case from the beginning under the natural interaction protocol. After each evidence item, the model produces its usual WAIT or ANSWER action. Before t_i^*, an attempted diagnosis submission is blocked from terminating the case, and the next evidence item is delivered. The blocked diagnosis is omitted from the subsequent conversation history, while ordinary WAIT responses are retained. In this way, the gate implements continued waiting without feeding the blocked diagnosis back to the model.

Once the evidence item at t_i^* has been delivered, the gate is removed. From that turn onward, the model can either submit a diagnosis or continue waiting under the original protocol. The first subsequent diagnosis submission ends the trajectory. If the model continues waiting through the complete evidence sequence, the case ends without a submitted diagnosis. For example, if $t_i^* = 12$, attempted submissions at turns 7 and 10 are both blocked. At turn 12, the model receives the twelfth evidence item and is free to submit a diagnosis or request further evidence.

The release turn uses the offline sufficiency annotation from Appendix A1. The model receives neither the correct diagnosis nor an instruction to submit at that turn. Access to the annotated boundary allows the experiment to measure the effect of preventing premature termination.

Reported outcomes. Table A8 uses the definitions in Appendix A2.4. The upper panel gives the number of selected cases, original and gated accuracy, gated submission rate, and paired accuracy gain. Every rate in this panel uses all selected cases as the denominator, including gated non-answers. The lower panel reports errors corrected, correct diagnoses lost, the confidence interval for their net effect, and the resulting gain over all 1,050 benchmark cases. Nonselected cases retain their original outcomes in this last calculation.

Final diagnosis outcomes. Gated accuracy exceeds original accuracy for every LLM in Table A8. Among selected cases, 32.9–50.5% change from an incorrect original diagnosis to a correct diagnosis

Table A8: Natural Gate outcomes for each LLM. The upper panel reports accuracy and submission rates on cases with an original premature diagnosis. The lower panel decomposes the paired accuracy change into errors corrected and correct diagnoses lost, and shows the corresponding gain over all 1,050 cases. Rates are percentages, and gains and confidence limits are percentage points (pp). Bold values highlight gated accuracy and the net accuracy gain on selected cases.

Diagnosis outcomes on selected cases

Model	Cases	Original accuracy (%)	Gated accuracy (%)	Submission rate (%)	Accuracy gain (pp)
Qwen3-8B	1,041	10.4	**55.6**	100.0	**+45.2**
Qwen3-14B	825	19.4	**67.8**	99.9	**+48.4**
Qwen3-32B	371	32.6	**62.3**	88.9	**+29.6**
Qwen3-235B	522	32.6	**76.2**	96.0	**+43.7**
Qwen3.6-27B	431	43.2	**81.4**	91.4	**+38.3**
Qwen3.8-27B	384	47.1	**82.8**	92.4	**+35.7**
DeepSeek-V4-F	258	37.2	**69.4**	80.2	**+32.2**
Kimi-K2.5	166	45.8	**79.5**	86.1	**+33.7**
Kimi-K2.6	178	45.5	**77.0**	83.1	**+31.5**

Paired changes and effect on the complete benchmark

Model	Errors corrected (%)	Correct diagnoses lost (%)	95% CI for accuracy gain on selected cases (pp)	Gain over all 1,050 cases (pp)
Qwen3-8B	47.7	2.5	[+42.0, +48.6]	+44.9
Qwen3-14B	50.5	2.2	[+44.7, +52.2]	+38.0
Qwen3-32B	32.9	3.2	[+24.3, +35.0]	+10.5
Qwen3-235B	45.6	1.9	[+39.1, +48.3]	+21.7
Qwen3.6-27B	41.3	3.0	[+33.2, +43.6]	+15.7
Qwen3.8-27B	38.8	3.1	[+30.2, +40.9]	+13.0
DeepSeek-V4-F	38.4	6.2	[+24.8, +39.1]	+7.9
Kimi-K2.5	39.2	5.4	[+25.3, +42.8]	+5.3
Kimi-K2.6	37.1	5.6	[+23.0, +39.9]	+5.3

under the gate, while only 1.9–6.2% lose an originally correct diagnosis. The resulting net accuracy gains range from 29.6 to 48.4 percentage points, with all nine paired confidence intervals above zero. These are improvements in actual submitted diagnoses, complementing the candidate-belief measurements in Appendix A4.2. The gains remain positive after accounting for both newly incorrect diagnoses and cases that end without a submission.

The gated submission rate ranges from 80.2% to 100%. For example, Qwen3-32B originally submits prematurely on 371 cases, with 32.6% accuracy on those cases. Under the gate, it submits a diagnosis on 88.9% of the same cases and achieves 62.3% accuracy when all 371 cases remain in the denominator. Errors corrected account for 32.9% of the selected cases and correct diagnoses lost account for 3.2%, producing a net gain of 29.6 percentage points before rounding. This example illustrates why submission rate and accuracy are reported separately.

Effect on the complete benchmark. The last column expresses the same paired changes relative to all 1,050 cases. It depends on both how frequently a model submits prematurely and how much the gate improves those cases. For Qwen3-8B, 1,041 of 1,050 cases receive the intervention, so its 45.2-point gain on selected cases corresponds to a 44.9-point gain over the full benchmark. For Kimi-K2.5, 166 cases receive the intervention, so a 33.7-point gain on selected cases corresponds to a 5.3-point gain over all cases. The smaller benchmark-wide gain in the latter example reflects the smaller number of premature submissions as well as the intervention's effect on those submissions.

A5 Additional Evaluation of EVD-Harness

A5.1 Evaluation Protocol and Diagnosis Outcomes

We evaluate LLM only and EVD-Harness on MedEVM with Qwen3.8-27B, DeepSeek-V4-Flash, Qwen3-32B, Kimi-K2.6, and Claude-Opus-4-8, as in Section 6.3. Both methods use the same candidates and evidence order for each case. LLM only chooses when to submit a diagnosis or wait. The harness uses the same frozen model to propose a diagnosis, retrieves CDW knowledge, and checks the proposal against the evidence through *Evidence Witness* before *Submission Controller* permits submission. A rejected proposal becomes a wait, so the next evidence unit can be considered. A trajectory ends at its first submitted diagnosis or when the available evidence is exhausted. Cases ending without a diagnosis count as incorrect in both conditions.

Accuracy is the percentage of all cases with a correct submitted diagnosis. Premature incorrect submission requires both $\tau_i < t_i^*$ and $\hat{y}_i \neq y_i^*$, while boundary submission requires $\tau_i = t_i^*$ regardless of diagnosis correctness. All rates use the full evaluated set as their denominator, including cases without an answer. Appendix A2.1 defines the remaining outcome and timing metrics.

Table A9: Diagnosis outcomes on MedEVM. Premature incorrect submissions combine a timing error with a diagnostic error. Boundary submissions occur exactly at t_i^*. Bold values indicate the better result within each model comparison.

	Accuracy ↑		Premature incorrect ↓		Boundary submissions ↑	
Model	LLM only	EVD-Harness	LLM only	EVD-Harness	LLM only	EVD-Harness
Qwen3.8-27B	60.9	**75.6**	32.0	**4.9**	41.6	**73.8**
DeepSeek-V4-Flash	42.9	**68.4**	14.7	**8.9**	27.2	**70.4**
Qwen3-32B	50.9	**62.9**	24.0	**7.8**	38.9	**67.1**
Kimi-K2.6	60.2	**78.0**	31.1	**7.6**	36.9	**66.4**
Claude-Opus-4-8	34.2	**85.3**	4.9	**3.3**	24.7	**68.4**

Table A9 shows that EVD-Harness improves diagnostic accuracy and reduces premature incorrect submissions for all five models. At the same time, more diagnoses are submitted at t_i^*, when the available evidence first distinguishes the correct candidate from its alternatives. The main results also show fewer unanswered cases (Table 4), so more trajectories end with a submitted diagnosis. Together, these results show improvements in both diagnostic success and the alignment of diagnosis submission with evidence sufficiency.

Table A10: Submission timing and accuracy with a ± 1-turn boundary tolerance (%). Correct within window counts correct diagnoses submitted inside $[t^* - 1, t^* + 1]$. Bold marks the better result per model.

Model	Method	Submission timing				Diagnostic outcome	
		Premature	Within window	Delayed	No answer	Accuracy	Correct within window
Qwen3.8-27B	LLM only	41.8	53.1	**2.4**	2.7	60.9	45.1
	EVD-Harness	**9.8**	**84.2**	3.6	**2.4**	**75.6**	**67.1**
DeepSeek-V4-Flash	LLM only	17.9	37.9	4.9	39.3	42.9	33.5
	EVD-Harness	**13.1**	**82.4**	**3.1**	**1.3**	**68.4**	**57.9**
Qwen3-32B	LLM only	29.1	48.7	**5.8**	16.4	50.9	38.9
	EVD-Harness	**10.0**	**78.2**	**5.8**	**6.0**	**62.9**	**56.7**
Kimi-K2.6	LLM only	40.7	49.3	**4.7**	5.3	60.2	42.7
	EVD-Harness	**14.7**	**75.6**	6.9	**2.9**	**78.0**	**63.3**
Claude-Opus-4-8	LLM only	9.3	29.3	**1.6**	59.8	34.2	27.8
	EVD-Harness	**6.7**	**79.3**	5.6	**8.4**	**85.3**	**73.1**

A5.2 Sensitivity to Boundary Tolerance

We test whether the timing advantage of EVD-Harness depends on requiring submission at exactly the annotated t_i^*. Following Appendix A3.2, we rescore the original trajectories with windows

Table A11: Submission timing and accuracy with a ± 2-turn boundary tolerance (%). Correct within window counts correct diagnoses submitted inside $[t^* - 2, t^* + 2]$. Bold marks the better result per model.

Model	Method	Submission timing				Diagnostic outcome	
		Premature	Within window	Delayed	No answer	Accuracy	Correct within window
Qwen3.8-27B	LLM only	33.8	61.3	2.2	2.7	60.9	48.4
	EVD-HARNESS	**7.8**	**88.0**	**1.8**	**2.4**	**75.6**	**69.3**
DeepSeek-V4-Flash	LLM only	15.2	42.0	3.6	39.3	42.9	35.3
	EVD-HARNESS	**10.7**	**87.5**	**0.4**	**1.3**	**68.4**	**62.6**
Qwen3-32B	LLM only	21.8	57.6	4.2	16.4	50.9	43.3
	EVD-HARNESS	**7.3**	**84.0**	**2.7**	**6.0**	**62.9**	**59.3**
Kimi-K2.6	LLM only	34.4	57.3	**2.9**	5.3	60.2	46.0
	EVD-HARNESS	**12.9**	**80.7**	3.6	**2.9**	**78.0**	**67.3**
Claude-Opus-4-8	LLM only	7.1	32.4	**0.7**	59.8	34.2	30.2
	EVD-HARNESS	**3.8**	**84.9**	2.9	**8.4**	**85.3**	**77.3**

$[t_i^* - k, t_i^* + k]$ for $k = 1$ and $k = 2$. Submissions before, inside, and after each window are classified as premature, within-window, and delayed. We report non-answers separately.

Tables A10 and A11 show that wider windows classify more submissions as timely under both methods. For each of the five models, EVD-HARNESS still places substantially more diagnoses near t_i^* and yields more correct diagnoses within the window. Thus, its timing advantage persists when the sufficiency boundary is allowed to vary by one or two turns.

The tolerance analysis reclassifies submission timing without changing any action or final diagnosis. Diagnostic accuracy is therefore unchanged, while the consistent increase in within-window submissions shows that the harness concentrates diagnosis submissions closer to the sufficiency boundary.

A5.3 SEPARATING KNOWLEDGE, VERIFICATION, AND CONTEXT EFFECTS

Component controls. To separate the contributions of CDW knowledge and evidence verification, we compare four named settings on the same 200 MEDEVM cases. Each setting uses the same patient evidence, candidate diagnoses, evidence order, harness execution protocol, and two-action history window. The *Full* setting supplies CDW knowledge to *Answer Proposal* and *Evidence Witness*. The *Submission Controller* then follows the witness verdict to submit or wait. The *−CDW* setting removes retrieved knowledge while retaining verification and submission control. The *−Witness* setting retains CDW but submits the diagnosis proposed by *Answer Proposal* directly, without calling *Evidence Witness*. The *Both* setting removes CDW and witness review together. The compact header −W abbreviates witness removal. Incorrect diagnoses and non-answers count as failures in every setting.

Table A12: Accuracy after removing CDW, Evidence Witness, or both (%). Red subscripts give differences from Full.

Model	Full	−Witness	−CDW	Both
Q3.8-27B	**75.5**	72.0$_{-3.5}$	63.0$_{-12.5}$	57.5$_{-18.0}$
DS-V4-F	**68.5**	65.0$_{-3.5}$	54.0$_{-14.5}$	43.0$_{-25.5}$
Q3-32B	**63.0**	58.5$_{-4.5}$	50.5$_{-12.5}$	46.5$_{-16.5}$
Kimi-K2.6	**78.0**	74.0$_{-4.0}$	66.0$_{-12.0}$	60.0$_{-18.0}$
Opus-4-8	**86.5**	77.5$_{-9.0}$	77.5$_{-9.0}$	75.0$_{-11.5}$

Table A13: Premature incorrect submissions (E) and non-answers (N) under the four component setting.

	Full		−W		−CDW		Both	
Model	E	N	E	N	E	N	E	N
Q3.8-27B	**3.5**	2.5	8.0	**0.5**	26.0	6.5	34.0	2.5
DS-V4-F	**6.5**	1.5	12.5	**0.0**	39.5	2.5	51.5	1.0
Q3-32B	**6.0**	6.0	14.0	**0.0**	29.0	11.0	36.0	4.5
Kimi-K2.6	**6.0**	3.0	12.5	**0.5**	24.0	7.0	31.5	6.0
Opus-4-8	2.5	8.5	5.5	**1.0**	**0.5**	22.0	11.5	12.5

Table A12 shows that removing *Evidence Witness* lowers accuracy by 3.5–9.0 percentage points across the five models. Removing CDW while retaining the witness lowers it by 9.0–14.5 points. Table A13 shows that witness removal also increases premature incorrect submissions, with the combined removal generally producing larger increases. CDW thus helps the model distinguish

candidate diagnoses, while the witness helps determine when a proposed diagnosis is supported for submission. Accuracy counts unanswered cases as failures throughout.

Retrieval source and contrastive organization. We isolate the value of contrastive organization by changing the Wiki pages supplied to *Answer Proposal* and *Evidence Witness*. As described in Section 6.1, the pages are compiled from independent case reports after excluding evaluation sources. *None* supplies no pages. The other controls retrieve disease pages, which summarize source-supported clinical findings. *Random* samples three pages at the first evidence turn; *BM25* selects three by text similarity to that evidence and the candidates. *Contrastive* organizes the same BM25 pages by

Table A14: Accuracy and premature incorrect submissions under five knowledge views (%). Bold marks the best result for each model and metric.

	Accuracy ↑					Premature incorrect ↓				
Model	None	Random	BM25	Contrast.	CDW	None	Random	BM25	Contrast.	CDW
Qwen3.8-27B	63.0	64.0	67.0	72.0	**75.5**	26.0	12.5	9.0	5.0	**3.5**
DeepSeek-V4-Flash	54.0	55.0	59.0	65.0	**68.5**	39.5	17.0	11.5	9.0	**6.5**
Qwen3-32B	50.5	51.5	54.5	59.5	**63.0**	29.0	17.5	11.5	7.5	**6.0**
Kimi-K2.6	66.0	67.0	70.0	75.0	**78.0**	24.0	17.5	13.0	7.5	**6.0**
Claude-Opus-4-8	77.5	78.5	80.5	84.5	**86.5**	**0.5**	1.5	1.0	**0.5**	2.5

candidate pair in $\mathcal{Y}_i$, highlighting shared and distinguishing findings. *Full CDW* updates retrieval as evidence arrives and follows links from disease to test and bridge pages, connecting findings to distinctions among candidates.

Table A14 shows that the *Full CDW* setting achieves the highest diagnostic accuracy across all five models. Starting from the same BM25 page retrieval, *Contrastive* raises accuracy by 4.0–6.0 percentage points and reduces premature incorrect submissions for each model. This comparison shows the value of organizing retrieved pages by the distinctions among the current candidate diagnoses, beyond matching pages to the patient text. The linked pages of the complete Wiki provide a richer view of those distinctions and support the strongest overall accuracy.

Action history and the execution protocol. In addition to retrieving knowledge and verifying diagnoses, the harness changes how many of its earlier wait and diagnosis-submission actions the model can see. We examine whether this shorter action record affects diagnostic outcomes and whether the harness execution procedure contributes separately. The *Natural* setting is the original LLM-only procedure with the complete patient record and prior actions. *Short* uses the same prompt and response processing and retains all patient evidence, but shows only the two most recent actions. *Direct* also shows two actions, while using the harness response format, parsing, and state updates to submit diagnosis proposals directly. None of these settings retrieves CDW knowledge or calls *Evidence Witness*, so the two successive comparisons examine the visible action record and execution procedure in turn.

Table A15: Accuracy and non-answer rates (%) under history and execution controls. Small numbers show changes from the preceding condition in percentage points. Bold marks the best result per model and metric.

	Accuracy ↑			No answer ↓		
Model	Natural	Short	Direct	Natural	Short	Direct
Qwen3.8-27B	59.5	**63.5**$_{+4.0}$	57.5$_{-6.0}$	3.5	**2.5**$_{-1.0}$	**2.5**$_{+0.0}$
DeepSeek-V4-Flash	45.5	**49.5**$_{+4.0}$	43.0$_{-6.5}$	39.5	**1.0**$_{-38.5}$	**1.0**$_{+0.0}$
Qwen3-32B	54.0	**58.0**$_{+4.0}$	46.5$_{-11.5}$	14.5	**4.0**$_{-10.5}$	4.5$_{+0.5}$
Kimi-K2.6	61.5	**65.5**$_{+4.0}$	60.0$_{-5.5}$	5.0	**3.5**$_{-1.5}$	6.0$_{+2.5}$
Claude-Opus-4-8	37.5	46.5$_{+9.0}$	**75.0**$_{+28.5}$	56.0	36.0$_{-20.0}$	**12.5**$_{-23.5}$

Table A15 shows accuracy increases of 4.0–9.0 points in the observed results after action history is shortened, alongside generally fewer non-answers. The Direct condition additionally changes

response formatting and state handling, showing that the execution procedure also affects diagnostic outcomes. Accordingly, the component and retrieval comparisons above hold both the history window and execution protocol fixed when assessing CDW and *Evidence Witness*.

A5.4 Sensitivity to Evidence Perturbations

We evaluate robustness on MedEVM using the evidence interventions associated with RQ3 and RQ4, with implementation details in Appendix A3. Each case is run with its original evidence sequence and with the corresponding perturbed sequence. The ground-truth diagnosis and candidate set remain fixed. Evidence reordering changes the presentation order of the same clinical information. The insertion conditions add repeated evidence, irrelevant text, or a paraphrase of existing evidence. Repeated and paraphrased evidence restate information already present, while irrelevant text adds no information that distinguishes the diagnostic candidates.

For these experiments, we compare the final outputs of the original and perturbed diagnosis trajectories using the diagnosis outcome disagreement defined in Appendix A2.3. A flip occurs when the submitted diagnoses differ or when only one trajectory submits a diagnosis. Two unanswered trajectories count as agreement. Table A16 compares LLM only and EVD-Harness under each insertion condition and reports their means across the three conditions. Results for evidence reordering appear in the *Order flip* column of Table 4.

Table A16: Answer-flip rates under evidence insertions (%, lower is better). Each method is evaluated under the same three perturbations. Mean denotes the average across conditions. Bold marks the lower rate in each paired comparison.

	LLM only				EVD-Harness			
Model	Repeated evidence	Irrelevant text	Paraphrased evidence	Mean	Repeated evidence	Irrelevant text	Paraphrased evidence	Mean
Qwen3.8-27B	9.7	8.5	8.3	8.8	**1.6**	**1.8**	**1.6**	**1.7**
DeepSeek-V4-Flash	31.7	27.0	29.0	29.2	**1.1**	**1.1**	**1.3**	**1.2**
Qwen3-32B	13.3	13.3	11.9	12.8	**3.4**	**3.8**	**2.7**	**3.3**
Kimi-K2.6	21.6	18.7	20.7	20.3	**1.8**	**1.1**	**1.6**	**1.5**
Claude-Opus-4-8	30.8	30.6	31.5	31.0	**3.4**	**3.1**	**3.4**	**3.3**

EVD-Harness consistently reduces answer flips across repeated evidence, irrelevant text, and paraphrased evidence (Table A16). Rates fall from 8.3–31.7% with LLM only to 1.1–3.8% with the harness. Averaged across the three conditions, the reduction ranges from 74% to 96% relative to LLM only. The improvement therefore extends across both repetition and changes in wording, supporting greater stability of diagnosis outputs when additional text supplies no new diagnostic information.

We also evaluate misleading evidence using the 318 paired cases described in RQ4. Each pair varies the strength of evidence favoring a designated incorrect diagnosis while retaining the underlying case and subsequent corrective evidence. The *Target error* column of Table 4 reports submission of that incorrect diagnosis before corrective evidence arrives in the high-strength condition. EVD-Harness reduces this error rate relative to LLM only for all five models. This result complements the insertion experiments by showing improved resistance to evidence that actively favors a wrong diagnosis, in addition to reduced sensitivity to redundant or irrelevant text.

A5.5 Transfer to Static Medical Question Answering

Static medical question answering tests whether the benefit of EVD-Harness extends beyond deciding when to answer during progressive evidence disclosure. In this setting, all question information is available from the outset. Improvements must therefore come from how the model uses that information to select and verify a diagnosis. We evaluate the same five models on DiagnosisArena-MCQ and MedXpertQA-Text, with results in Table A18.

Each model receives the complete question and its answer candidates in a single turn. The LLM-only baseline directly selects an answer. EVD-Harness treats the question as the available patient record and applies CDW retrieval, Differential Guidance, Answer Proposal, and Evidence Witness before

Table A17: DiagnosisArena-MCQ accuracy (%). Direct sampling aggregates the indicated number of independently generated answers. The single-sample column is LLM only. Bold marks the highest accuracy for each model.

	Direct sampling			EVD-HARNESS		
Model	1 sample	5 samples	6 samples	Full	−Witness	−CDW
Qwen3.8-27B	54.1	57.8	58.0	57.1	**64.8**	53.9
DeepSeek-V4-Flash	52.0	53.5	53.6	56.1	**59.9**	53.5
Qwen3-32B	50.2	50.8	51.7	49.6	**51.8**	46.2
Kimi-K2.6	63.7	64.0	64.9	67.5	**72.6**	64.3
Claude-Opus-4-8	67.8	67.4	67.3	69.8	**77.6**	63.3

diagnosis submission. The witness uses the same question information to assess the proposal, and its citations are checked against that information. A question with no authorized answer counts as incorrect. Accuracy therefore includes all questions for both methods. Since no further clinical evidence is disclosed, this evaluation uses accuracy without assigning a sufficiency boundary or measuring premature submissions.

Table A18: All-sample accuracy (%) on DiagnosisArena-MCQ and MedXpertQA-Text. Unanswered questions count as errors. Bold marks the better method for each model and benchmark.

	DiagnosisArena-MCQ		MedXpertQA-Text	
Model	LLM only	EVD-HARNESS	LLM only	EVD-HARNESS
Qwen3.8-27B	54.1	**57.1**	24.4	**27.9**
DeepSeek-V4-Flash	52.0	**56.1**	15.1	**22.1**
Qwen3-32B	**50.2**	49.6	10.5	**18.6**
Kimi-K2.6	63.7	**67.5**	17.4	**20.9**
Claude-Opus-4-8	67.8	**69.8**	38.4	**44.2**

To test whether gains on static QA arise from diagnostic knowledge, verification, or simply generating more answers, we compare a single direct answer, five or six independent answer samples, Full EVD-HARNESS, and two component removals on DiagnosisArena-MCQ (Table A17). Direct sampling selects the most frequent label from the same complete question. Ties follow a fixed label order, and invalid labels result in an unanswered, incorrect item. The −Witness condition retains CDW retrieval and diagnosis proposals but submits each proposal without witness review. The −CDW condition retains witness review but removes retrieved knowledge. All conditions receive the complete question at once.

Table A17 shows that CDW-guided diagnosis proposals without witness review attain the highest DiagnosisArena-MCQ accuracy for each model. Removing CDW lowers the Full result, demonstrating the value of retrieved differential knowledge when the entire question is available at once. On MedXpertQA-Text, Full EVD-HARNESS improves over LLM only for all five models (Table A18). These results show that CDW-supported diagnosis also transfers to static medical QA.