TraceDx: Criticality-Weighted Atomic Facts as a Training Signal for Sequential Clinical Diagnosis

Jiayi Li[1] **Sunny Chung**[2] **Don Chamil Codipilly**[3] **George Marek**[3]
Varan Perananthan[3] **Dennis L. Shung**[3] **Bradly C. Stadie**[1]
[1]Northwestern University [2]Yale University [3]Mayo Clinic

Abstract

In clinical decision-making, success depends on taking the right action to obtain the evidence that matters. One targeted clinical workup can identify a diagnosis that a battery of nonspecific tests missed. Large language models have shown considerable promise in clinical diagnosis, but most current approaches assume that all relevant patient information is available upfront, which does not reflect how diagnostic evidence is gathered in practice. Even when models gather evidence iteratively, challenges arise in determining which action to take, as they receive no feedback on which findings are diagnostically decisive. To solve this problem, we introduce TraceDx, a reinforcement learning framework for sequential clinical diagnosis. TraceDx is built on the idea that unstructured clinical records can be decomposed into atomic clinical findings, each scored by its diagnostic value. This signal rewards agents for discovering decisive evidence in addition to reaching the correct diagnosis. Consistent with this design, a case-level analysis shows that recovery of critical evidence is the only statistically significant predictor of diagnostic success. Direct case comparisons further show that TraceDx uncovers more diagnostically important evidence, and uses it to resolve the key diagnostic uncertainty. TraceDx-trained open-weight models outperform multiple frontier baselines on both MIMIC-CDM and a rare gastrointestinal disease dataset.

1 Introduction

Clinical diagnosis is inherently a sequential process in which evidence is acquired through questions asked, tests ordered, and results interpreted [4]. However, most approaches to clinical diagnosis with Large Language Models (LLMs) reduce this process to inference over a fixed input [21, 15, 16, 9, 5, 3]. This approach assumes all pertinent information is available upfront and overlooks how evidence is gathered iteratively in real clinical practice. A few benchmarks let agents acquire information iteratively [17, 12, 6], but they rely on LLMs without task-specific training.

This raises the question of how to train models to operate in clinical settings, wherein diagnosis depends on gathering the right evidence through a sequence of actions. The challenge is that not all clinical evidence is equally informative. Some findings are decisive, such as a CT confirming diverticulitis or a markedly elevated lipase indicating pancreatitis, whereas other findings provide only weak diagnostic signals or none at all. Clinicians succeed in part by prioritizing the evidence that best distinguishes among competing hypotheses, selectively pursuing tests that resolve uncertainty, rather than gathering information indiscriminately [18, 24, 2].

Training an agent to recognize these differences in diagnostic importance requires feedback on what its actions actually reveal, since the value of a diagnostic action lies in the evidence it uncovers. Yet such feedback is hard to obtain. The patient record presents an unstructured stream of notes, test results, and observations, mixing raw findings with clinical interpretations layered on top. At this granularity, the record offers no scorable unit. The findings are there, but not as discrete items

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the agent can be credited for recovering. We address this by decomposing each patient record into atomic clinical facts, drawing on prior work in factuality evaluation [14]. Each atomic fact is a single observation, a lab value, an imaging finding, or a physical exam result.

Atomization gives us scorable units, but no way to weigh them. Prior factuality work treats extracted facts as equally weighted and uses them to evaluate model outputs. In clinical decision making, equal weighting fails because findings differ in how much they narrow the diagnosis. We score each fact on an ordinal scale by its diagnostic criticality. The scale runs from clinically irrelevant evidence to pathognomonic findings that settle the diagnosis. Atomization and criticality scoring give us a way to value each piece of evidence. The remaining problem is how to leverage these per-fact values to assess action quality across a trajectory.

We propose TraceDx, a reinforcement learning framework that turns these per-fact values into a trajectory-level reward, crediting agents for the value of the evidence their actions discover. A trajectory that uncovers more decisive evidence receives higher credit, even for the same diagnosis. The agent operates in an unconstrained clinical simulator, where actions and diagnoses are proposed in free text with no predefined menu. In response, the simulator generates observations grounded in the patient record. The simulator environment mirrors what a clinician faces in practice, leaving the agent to decide which tests are worth pursuing. The goal is to train an agent that learns not just which diagnosis to reach, but which actions to take and why.

We evaluate TraceDx on two clinical diagnostic datasets. MIMIC-CDM [6] contains real emergency department cases derived from MIMIC-IV, spanning common abdominal pathologies. The rare disease dataset [10] draws from published medical case reports covering uncommon gastrointestinal and hepatobiliary conditions. The two settings present different challenges. In MIMIC-CDM, the agent must isolate a handful of decisive findings within hundreds of routine measurements per case, while in the rare disease set it must recognize the hallmark of a condition it has barely encountered. Across both settings, TraceDx-trained open-weight models match or exceed most frontier baselines, despite operating at a smaller scale. A case-level analysis identifies recovery of critical evidence as the only significant predictor of diagnostic success, and case studies show TraceDx-trained agents resolving diagnostic uncertainty with more decisive actions.

2 Related Work

Clinical diagnosis with LLMs. Prior work on LLMs for clinical diagnosis has largely operated in a non-interactive setting where the model receives a complete case and produces a diagnosis or differential. This includes multiple-choice medical QA [21, 22] and complex case vignettes drawn from published case reports [5, 3]. Several efforts move toward interaction, where the model must request evidence before committing to a diagnosis. Hager et al. [6] introduce MIMIC-CDM, an interactive diagnostic benchmark built on MIMIC-IV. Subsequent benchmarks evaluate stepwise evidence acquisition on clinical case vignettes [17, 12] and across multiple specialties [19]. All of these efforts evaluate fixed models rather than training them on specific tasks.

A smaller body of work addresses optimization. AMIE [28] trains a conversational agent for history-taking via supervised fine-tuning on self-play dialogues with a simulated patient. Yu et al. [31] train policies for cost-effective sequential test ordering over tabular lab values, and ED-Copilot [25] demonstrates LM-guided diagnostic ordering in emergency settings. A concurrent effort, LA-CDM [1], applies SFT and GRPO to a two-agent system on MIMIC-CDM, with a sparse terminal correctness reward augmented by per-test cost penalties, plus a separate GRPO objective for confidence calibration. LA-CDM operates over a predefined menu of four abdominal diagnoses and twelve test types. TraceDx differs in two ways. First, it trains a single-agent policy with a reward that quantifies the diagnostic value of discovered evidence, rewarding evidence quality rather than just terminal correctness. Second, it operates in an unconstrained simulator where actions and diagnoses are proposed in natural language, supporting open-ended exploration over the full action space. These choices let TraceDx learn what evidence matters, not just what diagnosis to output.

Multi-turn RL for LLM agents. Search-R1 [8] trains LLM agents with a binary terminal reward on the final answer. TraceDx replaces this with a richer terminal reward that credits trajectories not only for correctness but for discovering diagnostically decisive evidence.

Fact decomposition. Fact decomposition is the first step in the decompose-then-verify pipeline for long-form factuality evaluation introduced by FActScore [14] and extended in subsequent work [29, 23]. MedScore [7] adapts this paradigm to free-form medical text by introducing condition-aware claim decomposition for verification against medical corpora. While some work uses decompose-then-verify scores as training signals [27, 13], these efforts decompose model outputs into atomic facts and use the resulting score as a preference signal for offline alignment of a generator LLM. All these methods treat extracted facts as equally informative, which fails in diagnostic reasoning where findings differ in how much they narrow the diagnosis. Like these methods, we apply LLM-based decomposition to extract atomic facts. We depart from this paradigm in three ways. We decompose ground-truth patient records rather than model outputs, weight each fact by its diagnostic importance, and use the resulting signal as an online RL reward to train a diagnostic agent.

3 Preliminaries

Sequential Diagnostic Reasoning. We frame medical diagnosis as a sequential decision-making problem and model it as a finite-horizon POMDP in which the hidden state is the complete patient record M, fixed throughout the episode. An agent π_θ begins with an initial patient presentation o_0 and iteratively gathers further clinical evidence through a series of actions a_t, each yielding a clinical observation

$$o_{t+1} \sim \mu(\cdot \mid M, \tau_t, a_t),$$

where μ is the clinical simulator (Section 4.1), and $\tau_t = (o_0, a_0, o_1, \ldots, a_{t-1}, o_t)$ is the interaction history up to time t. Since M is never directly observed, the agent takes actions based on the history, yielding a history-dependent policy $\pi_\theta(a_t \mid \tau_t)$. Each episode terminates when the agent commits to a final diagnosis $\hat{d}$ or reaches the maximum horizon $T_{\max}$. The resulting trajectory $\tau = (o_0, a_0, o_1, a_1, \ldots, a_{T-1}, o_T, \hat{d})$ records the full diagnostic process. Here $T \leq T_{\max}$ is the number of actions taken. If the horizon is reached without a committed diagnosis, $T = T_{\max}$ and $\hat{d} = \emptyset$.

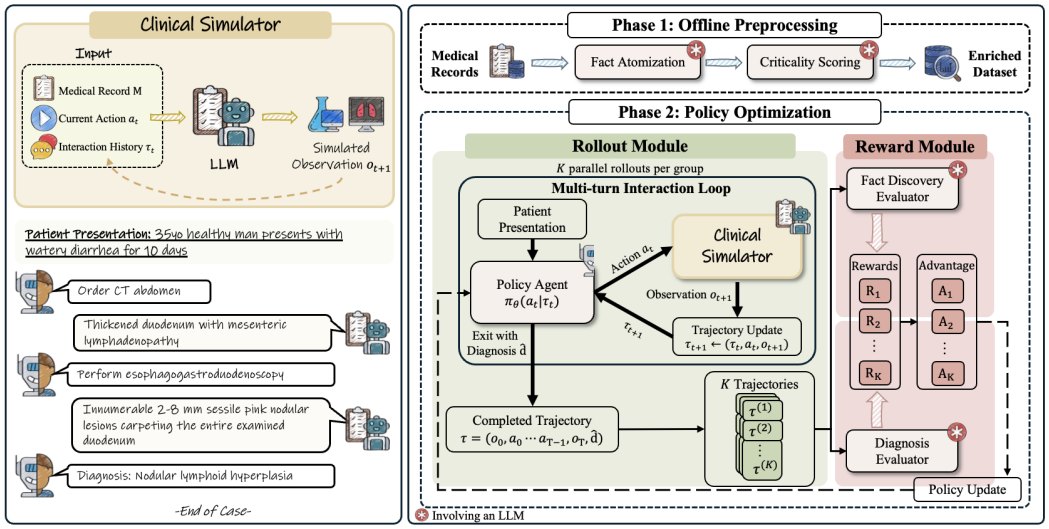

Figure 1: Overview of the TraceDx framework. *Left:* The agent issues a diagnostic action a_t at each step and the simulator returns an observation based on the patient record M and interaction history τ_t. *Right:* Two-phase training pipeline. Phase 1 extracts atomic clinical facts from each record offline and scores them by diagnostic importance. Phase 2 trains the agent with the criticality-weighted reward $R(\tau)$.

4 Method

We propose TraceDx, a framework that turns fact-level criticality into a training signal, rewarding agents for discovering the facts that actually drive diagnosis rather than merely for arriving at the right answer. TraceDx consists of four components: (1) an unconstrained clinical simulator for multi-step diagnostic interaction, (2) an offline fact atomization and criticality scoring pipeline, (3) a criticality-weighted reward, and (4) policy optimization.

4.1 Clinical Simulator Setup

The simulator μ is implemented as an LLM grounded in the patient record M. Both the action a_t and the eventual diagnosis $\hat{d}$ are produced in free-text form, without a predefined action set or diagnosis label space. This design captures the open-ended nature of clinical workups, where clinicians choose tests and diagnoses freely rather than from a fixed menu. The gold diagnosis d^* is used solely for reward computation and is never provided to the agent or simulator. Providing d^* to the simulator would risk generating confirmatory observations that bias the trajectory toward the correct answer, introducing a signal that would not be available in real clinical practice. An example conversation between the agent and the simulator is shown in the left panel of Figure 1. Further implementation details, limitation analysis, and validation results are in Appendix C.

4.2 Fact Atomization and Criticality Scoring

The patient record is unstructured and heterogeneous, mixing raw findings with clinical interpretations. To obtain a fine-grained training signal, we decompose patient records into atomic clinical facts and assign each fact a criticality score reflecting its diagnostic importance. This produces, for every case, a set of scorable units annotated with their relative diagnostic value (Figure 2). These scores are used in Section 4.3 to construct a trajectory-level reward.

Before training, we preprocess each patient record M in the dataset to extract a set of atomic clinical facts. Formally, we define a fact extraction function ϕ such that $F = \phi(M) = \{f_1, \ldots, f_n\}$, where each f_i is an atomic clinical finding, a single observation such as a lab value, imaging finding, or physical examination result, stated without interpretation or inference.

Each fact is then scored by its diagnostic importance. Given the fully extracted fact set F, gold diagnosis d^*, and initial patient presentation o_0, an LLM-based scorer assigns each fact a criticality score $c(f) \in \{0, 1, 2, 3\}$ conditioned on d^* and o_0. Higher scores indicate greater clinical importance: $c = 0$ for clinically irrelevant evidence, $c = 1$ for supportive but non-specific evidence, $c = 2$ for strongly supportive evidence, and $c = 3$ for pathognomonic or hallmark findings. This scoring is performed offline prior to training, adding no computational overhead during rollout. Full details are in Appendix D.

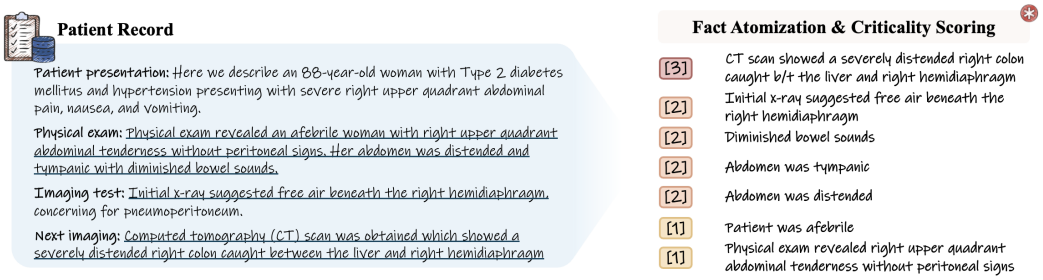

Figure 2: Fact atomization and criticality scoring. A patient record is decomposed into atomic findings, each scored by diagnostic importance $c(f) \in \{0, 1, 2, 3\}$. Underlined text marks content extracted as atomic facts. Initial patient presentation and clinical interpretations are excluded.

4.3 Criticality-Weighted Reward

A naive baseline is to optimize the policy using a binary outcome-based reward that depends only on the final diagnostic accuracy: $R_{\text{naive}}(\tau) = \text{acc}(\hat{d}, d^*)$, where $\text{acc}(\hat{d}, d^*) \in \{0, 1\}$ is evaluated by an LLM judge assessing clinical equivalence rather than exact string match. Diagnostic reasoning involves both gathering evidence and reasoning over it, but terminal diagnostic accuracy supervises only the latter, leaving the agent without a signal to learn the former.

To address this limitation, we design a mechanism to measure how much diagnostically relevant information is uncovered along a trajectory, which serves as a proxy for the quality of the actions taken. Given the fact set F and a completed trajectory τ, we let $D(\tau) \subseteq F$ denote the subset of facts revealed during the trajectory, computed by an LLM-based evaluator that parses the trajectory and outputs indices of observed facts (see Appendix E.1).

We quantify the amount of diagnostically important evidence recovered using the criticality recall:

$$CR(\tau) = \frac{\sum_{f \in D(\tau)} c(f)}{\sum_{f \in F} c(f)}.$$

We restrict training to cases with $\sum_{f \in F} c(f) > 0$ so that $CR(\tau)$ is well-defined.

We define the criticality-weighted reward as

$$R(\tau) = (\alpha \cdot CR(\tau) + \beta)\, \text{acc}(\hat{d}, d^*) - \lambda \frac{T}{T_{\max}}, \tag{1}$$

where $\alpha, \lambda \geq 0$ and $\beta > 0$. Trajectories that uncover more diagnostically decisive facts receive higher reward for the same correct diagnosis, while the positive offset β ensures that correct diagnoses remain rewarded even when criticality recall is low. The step penalty $-\lambda T / T_{\max}$ discourages unnecessary actions and encourages efficient reasoning.

The following proposition establishes two guarantees under approximate calibration: the CR component orders trajectories by true evidence quality whenever scoring noise is small, and the full reward inherits this ordering for correct trajectories whenever their CR gap exceeds the step-penalty difference.

Proposition 1 (Reward order preservation under approximate calibration)**.** *Assume $\sum_{f \in F} c(f) > 0$. Let each fact $f \in F$ have latent relevance $r_f \geq 0$, and suppose the scorer satisfies*

$$c(f) = \kappa r_f + \varepsilon_f, \quad \kappa > 0, \ |\varepsilon_f| \leq \sigma,$$

where $\sigma \geq 0$ is the worst-case per-fact scoring error. For two trajectories $\tau^{(1)}, \tau^{(2)}$ from the same patient case (sharing fact set F) with discovered sets $D_i = D(\tau^{(i)})$, define $\nu(D) = \sum_{f \in D} r_f$. If

$$\nu(D_1) - \nu(D_2) > \frac{\sigma}{\kappa}(|D_1| + |D_2|),$$

then $CR(\tau^{(1)}) > CR(\tau^{(2)})$. Furthermore, if $\alpha > 0$, both trajectories are correct, and $CR(\tau^{(1)}) - CR(\tau^{(2)}) > \frac{\lambda}{\alpha} \cdot \frac{T_1 - T_2}{T_{\max}}$, where T_i is the number of actions in $\tau^{(i)}$, then $R(\tau^{(1)}) > R(\tau^{(2)})$.

In short, the criticality recall ranks trajectories by true evidence quality whenever scoring errors are small relative to the true relevance gap. The full reward inherits this ranking for correct trajectories whenever their CR gap exceeds the step-penalty difference. A formal proof is provided in Appendix J, and validation of the criticality scorer is detailed in Appendix D.

4.4 Policy Optimization

We optimize the policy π_θ using an importance-weighted REINFORCE objective with GRPO-style group-relative advantages [20], omitting PPO-style clipping and the KL penalty. For each patient case, we sample a group of K trajectories $\{\tau^{(1)}, \ldots, \tau^{(K)}\}$, each receiving a scalar terminal reward $R_k = R(\tau^{(k)})$. We form the group-relative advantage $A_k = R_k - \text{mean}(R_1, \ldots, R_K)$.

Each clinical action a_t is realized as a sequence of agent-generated tokens, so we apply the policy gradient at the token level. Our training objective is:

$$\mathcal{L}(\theta) = -\sum_{k=1}^{K} \sum_{l} m_{k,l}\, \rho_{k,l}\, A_k,$$

where $y_{k,l}$ is the l-th token of trajectory $\tau^{(k)}$, $y_{k,<l}$ denotes the preceding tokens in $\tau^{(k)}$, $\rho_{k,l} = \pi_\theta(y_{k,l} \mid y_{k,<l}) / \pi_{\text{old}}(y_{k,l} \mid y_{k,<l})$ is the per-token importance ratio, and $m_{k,l} \in \{0, 1\}$ is 1 for agent-generated tokens and 0 otherwise. We mask non-agent tokens by zeroing their per-token advantage, analogous to retrieved-token masking in Search-R1 [8]. π_{old} denotes the policy used to generate the rollouts, whose log-probabilities are recorded at sampling time and held fixed during the gradient update. Groups with identical rewards across all K trajectories yield $A_k = 0$ and are excluded.

5 Experiments

5.1 Experiment Setup

Datasets. We evaluate on two clinical diagnostic datasets that demand different kinds of reasoning (Table 1). *MIMIC-CDM* [6] contains full hospital records for three related abdominal conditions with distinct diagnostic structures. A few decisive findings must be picked out of a long record of routine measurements, and the conditions must be told apart rather than merely recognized. To ensure reliable evaluation with an LLM-based grader, we restrict to cases with short, well-defined diagnoses. The *Rare Disease* dataset [10] is drawn from published case reports. Every case is a distinct diagnosis, most are uncommon conditions, and the underlying mechanisms span every major category with none dominating. We hold out 75 MIMIC-CDM and 81 Rare Disease cases for testing. Detailed dataset construction and filtering criteria are provided in Appendix B.

Table 1: Disease composition of each dataset.

MIMIC-CDM	
Cholecystitis	34%
Diverticulitis	34%
Pancreatitis	32%
Rare Disease	
Cancer/tumor	31%
Structural/mechanical	16%
Vascular	13%
Infection	12%
Congenital	10%
Toxic/metabolic	8%
Immune/inflammatory	8%
Other	2%

Baselines. We compare against closed-weight frontier LLMs (GPT-5, GPT-5.1, GPT-5.4, Sonnet 4, Sonnet 4.5, Opus 4.7, Gemini 2.5 Pro, Gemini 3.1 Pro) and the untrained base models Qwen3-32B [30] and Kimi-K2.5 [26]. We also include GPT-5.1 with oracle context as an approximate upper bound, where the full patient record is provided upfront, removing the evidence-acquisition burden and bounding what any interactive agent could achieve with complete information. Further implementation details for all methods are in Appendices F and G. We additionally compare against physicians playing the same simulator game on a smaller subset of the held-out test set (see Appendix L).

Evaluation. Models are evaluated on diagnostic accuracy via interaction with the clinical simulator, with Claude Sonnet 4 as both the simulator backbone and the diagnosis grader. The simulator, criticality scorer, and diagnosis grader were each validated against physician judgment (Appendices C.3, D.3, and E.3).

Table 2: Diagnostic accuracy across Rare Disease and MIMIC-CDM datasets. **Bold**: best frontier. Underline: best open-weight.

	Model	Rare Disease	MIMIC-CDM
Frontier LLMs	GPT-5	0.419 ± 0.013	0.567 ± 0.016
	GPT-5.1	0.407 ± 0.025	0.577 ± 0.040
	GPT-5.4	0.374 ± 0.025	0.604 ± 0.046
	Sonnet 4	0.319 ± 0.019	0.519 ± 0.026
	Sonnet 4.5	0.357 ± 0.040	0.595 ± 0.040
	Opus 4.7	$\mathbf{0.456 \pm 0.049}$	$\mathbf{0.653 \pm 0.073}$
	Gemini 2.5 Pro	0.386 ± 0.047	0.573 ± 0.038
	Gemini 3.1 Pro	0.435 ± 0.014	0.560 ± 0.013
Qwen3-32B	Base	0.316 ± 0.028	0.555 ± 0.024
	TraceDx	0.370 ± 0.024	0.600 ± 0.027
Kimi-K2.5	Base	0.378 ± 0.026	0.568 ± 0.028
	TraceDx	$\underline{0.424 \pm 0.026}$	$\underline{0.613 \pm 0.026}$
Oracle	GPT-5.1 (full context)	0.543 ± 0.032	0.680 ± 0.013

5.2 Accuracy Analysis

Table 2 reports diagnostic accuracy across both datasets. TraceDx delivers consistent improvements over the untrained base in every setting we tested. Both base models gain roughly 4.5 to 5.4 percentage points across the two datasets. The gains are nearly identical across base models of different scale and architecture, and across datasets with very different fact distributions. The consistency across architectures and datasets suggests the gains are attributable to the TraceDx framework rather than to the capabilities of any particular base model.

Beyond improving over the base models, TraceDx makes open-weight models competitive with frontier models, indicating that training signal, not scale alone, drives diagnostic performance in this setting. Kimi-K2.5 with TraceDx surpasses most frontier models on both datasets, ranking second on MIMIC-CDM and within the top three on Rare Disease. Qwen3-32B with TraceDx also surpasses most frontier models on MIMIC-CDM, despite its base trailing the majority of them. The remaining gap to the oracle baseline reflects the cost of acquiring evidence interactively.

5.3 Ablation Study

We conduct an ablation study to identify the source of TraceDx's gains. We compare TraceDx against an untrained Base and a Diagnosis-only (Dx) ablation that swaps TraceDx's criticality reward for binary correctness. This comparison separates the effect of reinforcement learning itself from that of explicitly training for evidence acquisition. Diagnosis-only fails to improve over the base model and even slightly degrades performance on MIMIC-CDM, whereas TraceDx performs best on both datasets. A dense fact reward might help regardless of how facts are valued. To test this, we replace clinical criticality with rarity, defining the value of a piece of information by how rarely it is recovered across 10 frontier-model runs on MIMIC-CDM. We run this comparison on MIMIC-CDM because its records are far more information-rich, giving a rarity-based reward its best chance to work. A MIMIC-CDM record averages 131 atomic facts, mostly routine laboratory values and negative imaging findings, whereas a rare-disease case report averages 10 facts already curated to the relevant workup (Table 10, Appendix D.1). Recovery frequency can only separate facts that every agent finds from those that few do when there are many facts to choose among, and the short rare-disease records leave little room for that spread. *Linear IDF* weights each fact by its inverse recovery frequency, and *Band frequency* rewards recall separately within frequency bands (Appendix H.1). Neither improves meaningfully over the base model, and Band frequency even degrades it. A finding that agents rarely recover does not necessarily narrow the differential, so rarity is not a proxy for clinical value. Further ablations are in Appendix H. We next test the premise underlying the criticality-weighted reward, asking whether evidence quality predicts diagnostic success.

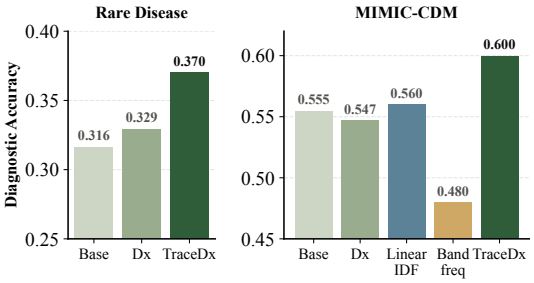

Figure 3: Reward ablation for Qwen3-32B. Linear IDF and Band freq were run on MIMIC-CDM only.

5.4 Evidence Quality Predicts Diagnostic Success

We focus this analysis on the rare disease data, where criticality annotations are well-distributed across all four levels (0–3), enabling reliable discrimination between facts of different diagnostic value. In MIMIC-CDM, more than 80% of facts receive criticality 0, which compresses the signal range and makes case-level estimates of hallmark-evidence concentration too noisy for inference (Figure 7; Appendix D.2). The rare disease set therefore provides a cleaner setting for testing whether evidence quality predicts diagnostic success.

To test this premise, each of the twelve models in Table 2 first plays the simulator game on rare-disease cases to gather its own evidence. We then pass that evidence to a fixed GPT-5.1 reasoner that produces the diagnosis without access to the gold diagnosis. Holding the reasoner constant ensures that any variation in diagnostic success reflects evidence quality, not diagnostic reasoning capability.

We characterize each evidence set along three dimensions: *Crit-3 ratio*, the fraction of retrieved facts with $c(f) = 3$; *noise ratio*, the fraction of retrieved facts with $c(f) = 0$; and *coverage*, the total fraction of available facts retrieved. We fit a logistic regression predicting diagnostic success from these three features, with agent fixed effects and standard errors clustered by case.

Table 3: Logistic regression of diagnostic success on evidence features (N=972, agent fixed effects).

Predictor	**Coef.**	p
Crit-3 Ratio	6.61	**0.002**
Noise Ratio	−0.16	0.803
Coverage	0.11	0.862
Agent Fixed Effects	Included	

Crit-3 ratio is the only significant predictor of diagnostic success, while noise ratio and coverage are both non-significant (Table 3). The fixed effects ensure this reflects within-agent variation in evidence quality. Holding model identity constant, hallmark concentration is what separates correct from incorrect diagnoses. The criticality reward is designed to shift evidence acquisition toward hallmark findings, and the regression shows that hallmark concentration is the strongest predictor of diagnostic success. These findings are consistent with TraceDx's premise that diagnostic success depends on evidence quality, not quantity.

Because criticality is scored with the gold diagnosis, high-scoring facts might look decisive only in hindsight. The outcome here, however, comes from a reasoner that never sees the gold diagnosis, so facts that mattered only in hindsight would not help it succeed. The association also holds within each of the ten agents never trained on our reward (Appendix I), so it does not arise from our training signal.

5.5 Robustness to LLM Components

Criticality scoring agreement. Criticality scores enter the reward directly through criticality recall (Equation 1), so a scorer's idiosyncrasies would carry straight into training. To test whether the scores depend on the scorer, we re-scored the full rare-disease fact set with three scorers from different model families (GPT-5.1, Gemini 2.5 Pro, Claude Sonnet 4.5), plus a second GPT-5.1 run as a within-model reference. Agreement across the three families reaches the conventional threshold for reliable annotation (Krippendorff's $\alpha = 0.80$). Pairwise, two scorers from different families agree nearly as well (quadratic-weighted κ of 0.78–0.82) as one model agrees with its own rerun ($\kappa = 0.85$), so switching scorer families adds little beyond sampling noise. When scorers do disagree, they rarely disagree by much. Every pair places 97–99% of facts within one point of each other, and hallmark facts, which carry the most weight in the reward, receive identical scores 92–96% of the time. A fact's score therefore reflects the finding rather than any one model's judgment (Appendix D.3).

Reward stability. Fact extraction and criticality scoring run once per case before training, but the fact-matching grader (Gemini 3 Flash) runs on every rollout. Given a trajectory and the case's fact list, it decides which facts the agent uncovered, and any inconsistency in that judgment would make the training target shift from one rollout to the next. To measure this, we re-graded the same held-out Rare Disease trajectories five times and compared criticality recall and the final reward across runs. Most cases receive exactly the same value in all five runs, and the spread is negligible (Table 4). The residual variance comes from cases with few facts, where criticality recall is a ratio over a small denominator and a single disputed fact shifts the score. The final reward varies even less, since recall enters it only for correct diagnoses (Appendix E.3).

Table 4: Fact-matching grader across five reruns on Rare Disease ($n = 81$).

Metric	**Std**	**Identical**
Criticality recall	0.0071	71/81
Final reward	0.0034	74/81

Simulator sensitivity. The simulator is held fixed across every model and reward variant, so we also re-evaluate under a different simulator backbone (GPT-5.1) and a paraphrased simulator prompt, changing one at a time. Absolute accuracy shifts slightly, but TraceDx's margin over its untrained base holds in every condition (Appendix C.3).

5.6 Performance by Disease Category

Overall accuracy can obscure important differences in how models behave across diseases with distinct diagnostic structures. We therefore break down performance on MIMIC-CDM by disease category to ask when criticality-guided training helps most.

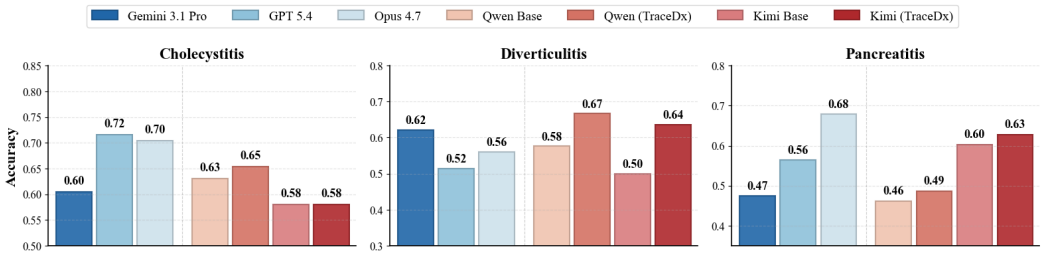

Figure 4: Mean diagnostic accuracy per disease category across models.

Cholecystitis. TraceDx yields a modest improvement for Qwen3-32B but no gain for Kimi-K2.5. Qwen3-32B (TraceDx) surpasses Gemini 3.1 Pro but trails GPT 5.4 and Opus 4.7. Cholecystitis has no single pathognomonic finding. Diagnosis requires satisfying criteria across local signs, systemic markers, and imaging, which may dilute the benefit of criticality-guided training.

Diverticulitis. TraceDx shows the clearest gains here, with both models improving substantially over their respective base models and surpassing all three frontier models shown. This is consistent with the relatively direct diagnostic structure of diverticulitis, where CT imaging provides a clear and decisive signal.

Pancreatitis. Both TraceDx models improve modestly over their base models by a similar margin. Kimi-K2.5 with TraceDx surpasses two of the three frontier models shown (Gemini 3.1 Pro and GPT 5.4), trailing only Opus 4.7. Qwen3-32B Base starts substantially weaker than Kimi-K2.5 Base on pancreatitis, which explains why the TraceDx variants diverge. The gap reflects the base models' relative strengths rather than a differential effect of criticality-guided training.

TraceDx shows the clearest gains on diverticulitis, where both models exceed frontier performance, and limited benefit on cholecystitis. On pancreatitis, both models yield consistent modest gains, with the divergence between the two TraceDx models explained by their starting points.

5.7 Case Study

We examine one case from each dataset to illustrate how TraceDx shapes diagnostic reasoning.

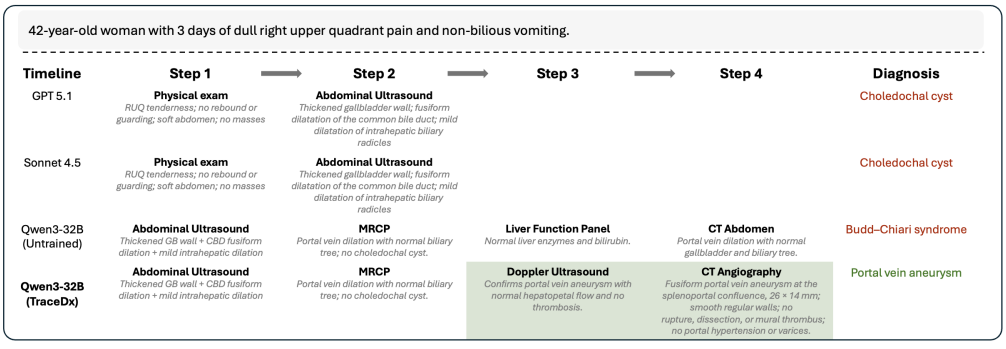

Figure 5: Diagnostic trajectories for portal vein aneurysm. The grey box shows the initial presentation. Each row shows a model's sequence of diagnostic actions and the resulting final diagnosis. Dark green boxes mark the actions that resolve the key diagnostic uncertainty.

Case 1 (Portal Vein Aneurysm)

A 42-year-old woman presented with right upper quadrant pain and vomiting (Figure 5). Initial ultrasound showed a dilated common bile duct and intrahepatic biliary radicles, suggesting a choledochal cyst, a congenital bile-duct abnormality. GPT-5.1 and Sonnet 4.5 accepted this framing and concluded choledochal cyst without further workup. Qwen3-32B Base obtained an MRCP that revealed a normal biliary tree but a dilated portal vein, directly contradicting the bile-duct interpretation, but misread this as Budd-Chiari syndrome, a different vascular disorder. TraceDx recognized the contradiction, shifted its hypothesis from biliary to vascular pathology, ordered a Doppler ultrasound confirming a portal vein aneurysm, and used CT angiography to characterize the lesion. TraceDx succeeds not by gathering more evidence, but by being willing to abandon a working hypothesis when the evidence contradicts it.

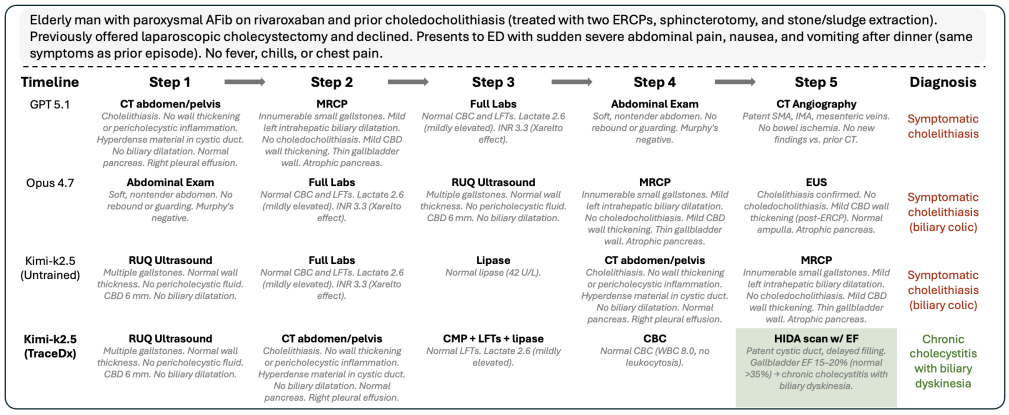

Figure 6: Diagnostic trajectories for chronic cholecystitis. Same layout as Figure 5.

Case 2 (Chronic Cholecystitis)

An elderly man with recurrent biliary pain posed a different challenge (Figure 6). Standard tests came back negative for acute pathology across all models, narrowing the differential to two conditions. Symptomatic cholelithiasis refers to gallstones causing pain in a structurally normal gallbladder. Chronic cholecystitis with biliary dyskinesia presents with the same imaging appearance but reflects a gallbladder that has lost its ability to contract. The two are indistinguishable on ultrasound and CT, and the only differentiating test is a HIDA scan, which measures gallbladder contraction directly. The baselines, including Opus 4.7, stopped at symptomatic cholelithiasis without testing gallbladder function. TraceDx ordered the HIDA scan, and the resulting ejection fraction of 15–20%, well below the normal threshold of 35%, established chronic cholecystitis with biliary dyskinesia. The case shows that TraceDx improves not only by selecting decisive actions, but by recognizing when a standard workup leaves a diagnostically pivotal question unresolved.

6 Conclusion

TraceDx trains diagnostic agents on the evidence they recover, not just the diagnoses they reach. We decompose each patient record into atomic clinical facts and score them by diagnostic criticality, building a fine-grained training signal where prior approaches saw only a final outcome. Agents are rewarded for recovering decisive evidence in an open-ended simulator with no predefined actions or diagnoses, training them toward evidence quality rather than terminal accuracy. A regression shows hallmark concentration as the only significant predictor of diagnostic success, and an ablation traces the gain to the criticality-weighted reward. Across two datasets, TraceDx-trained open-weight models match or exceed multiple frontier baselines. Looking ahead, this fact-atomization and criticality-scoring pipeline could serve as a reusable training and evaluation signal in other expert domains where decisions hinge on evidence quality.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their constructive feedback.

References

[1] David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. Language agents for hypothesis-driven clinical decision making with reinforcement learning. In *International Conference on Learning Representations*, 2026.

[2] Georges Bordage. Why did I miss the diagnosis? some cognitive explanations and educational implications. *Academic Medicine*, 74(10):S138–S143, oct 1999. doi: 10.1097/00001888-199910000-00065. October Supplement.

[3] Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdulnour, Adrian D. Haimovich, Jason A. Freed, Andrew Olson, Daniel J. Morgan, Jason Hom, Robert Gallo, Liam G. McCoy, Haadi Mombini, Christopher Lucas, Misha Fotoohi, Matthew Gwiazdon, Daniele Restifo, Daniel Restrepo, Eric Horvitz, Jonathan Chen, Arjun K. Manrai, and Adam Rodman. Superhuman performance of a large language model on the reasoning tasks of a physician, 2025.

[4] Arthur S Elstein and Alan Schwarz. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*, 324(7339):729–732, March 2002. doi: 10.1136/bmj.324.7339.729.

[5] Alexander V. Eriksen, Sören Möller, and Jesper Ryg. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI*, 1(1):AIp2300031, 2024. doi: 10.1056/AIp2300031.

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

[7] Heyuan Huang, Alexandra DeLucia, Vijay Murari Tiyyala, and Mark Dredze. Medscore: Generalizable factuality evaluation of free-form medical answers by domain-adapted claim decomposition and verification, 2025. URL `https://arxiv.org/abs/2505.18452`.

[8] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL `https://arxiv.org/abs/2503.09516`.

[9] Zahir Kanjee, Byron Crowe, and Adam Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*, 330(1):78–80, 2023. doi: 10.1001/jama.2023.8288.

[10] Youngmin Ko, Navia Yun, Dennis Shung, and Bradly Stadie. Sa2051: Automated prompt optimization strategy improves large language model diagnostic accuracy for complex clinical cases in gastroenterology and hepatology. *Gastroenterology*, 169(1):S–632, 2025.

[11] Thinking Machines Lab. Tinker, 2025. URL `https://thinkingmachines.ai/tinker/`.

[12] Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning, 2024. URL `https://arxiv.org/abs/2406.00922`.

[13] Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen. Flame: Factuality-aware alignment for large language models, 2024. URL `https://arxiv.org/abs/2405.01525`.

[14] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pages 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

[15] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023. URL https://arxiv.org/abs/2311.16452.

[16] Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. From medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond, 2024. URL https://arxiv.org/abs/2411.03590.

[17] Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with language models, 2025. URL https://arxiv.org/abs/2506.22405.

[18] Stephen G. Pauker and Jerome P. Kassirer. The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117, 1980. doi: 10.1056/NEJM198005153022003. URL https://www.nejm.org/doi/full/10.1056/NEJM198005153022003.

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

[20] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[21] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. doi: 10.1038/s41586-023-06291-2.

[22] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H Chen, Nigam H Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera Y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S Sara Mahdavi, Joelle K Barral, Dale R Webster, Greg S Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nat. Med.*, 31(3):943–950, March 2025.

[23] Yixiao Song, Yekyung Kim, and Mohit Iyyer. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation, 2024. URL https://arxiv.org/abs/2406.19276.

[24] Harold C. Sox, Michael C. Higgins, Douglas K. Owens, and Gillian Sanders Schmidler. *Medical Decision Making*. John Wiley & Sons, 2024. ISBN 9781119627876.

[25] Liwen Sun, Abhineet Agarwal, Aaron Kornblith, Bin Yu, and Chenyan Xiong. Ed-copilot: Reduce emergency department wait time with language model diagnostic assistance, 2024. URL https://arxiv.org/abs/2402.13448.

[26] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026. URL `https://arxiv.org/abs/2602.02276`.

[27] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality, 2023. URL `https://arxiv.org/abs/2311.08401`.

[28] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, June 2025.

[29] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models, 2024. URL https://arxiv.org/abs/2403.18802.

[30] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[31] Zheng Yu, Yikuan Li, Joseph Chahn Kim, Kaixuan Huang, Yuan Luo, and Mengdi Wang. Deep reinforcement learning for cost-effective medical diagnosis. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0WVNuEnqVu.

A Appendix Overview

Appendix B: Dataset Details 15
Construction, filtering criteria, and statistics for MIMIC-CDM and Rare Disease.

Appendix C: Clinical Simulator Details 16
Implementation, simulator limitations, and validation.

Appendix D: Fact Extraction and Criticality Scoring 19
Atomic fact extraction, criticality scoring, and scoring validation.

Appendix E: Evaluation Protocol 21
Grader design, dataset-specific grading criteria, and grader validation.

Appendix F: Baseline Implementation Details 24
Diagnostic loop setup for frontier models and oracle context baseline.

Appendix G: Training Details 24
LoRA configurations, GRPO hyperparameters, reward parameters, and inference setup.

Appendix H: Additional Reward Ablations 25
Frequency-based reward baselines and fact-weighting ablations on Qwen3.6-27B.

Appendix I: Per-Agent Evidence-Quality Analysis 26
Within-agent correlation between hallmark evidence and diagnostic success for agents never trained on our reward.

Appendix J: Proof of Proposition 1 27
Formal proof of reward order preservation under approximate criticality scoring.

Appendix K: Failure Case Study 27
Two failure trajectories and the open questions they raise.

Appendix L: Physician Baseline 29
Physician diagnostic accuracy on the simulator game compared to models.

Appendix M: Impact Statement 30
Positive and negative societal impacts of TraceDx, including clinical deployment caveats.

Appendix N: Prompt Templates 30
Full prompts for simulator, fact extraction, criticality scoring, grader, baselines, and GRPO training.

B Dataset Details

Both datasets cover gastrointestinal and hepatobiliary disease, so our evaluation is confined to this region. We study sequential diagnosis, where an episode begins with the presenting complaint and ends when the agent commits to a diagnosis. Patient management and longitudinal care therefore fall outside our scope, and extending TraceDx to them would require longer episodes and datasets that follow care over time. Specialties with more diffuse reasoning, such as psychiatry and chronic disease, remain untested, and cross-specialty validation is still needed to establish how far the method generalizes. The main constraint is data availability. Patient records are rarely released publicly because of privacy and data-use restrictions, and to our knowledge MIMIC-CDM is the only public patient-record dataset suited to open-ended sequential diagnosis. We therefore pair it with the rare disease set, which broadens diagnostic diversity beyond MIMIC-CDM's three conditions, and we expect the approach to extend as suitable datasets become available.

B.1 MIMIC-CDM Dataset

Data source. Cases are drawn from MIMIC-IV [6], a large de-identified electronic health record database. The source pool contains 648 cholecystitis, 257 diverticulitis, and 538 pancreatitis admissions (1,443 total). Each admission includes patient history (free text), physical examination findings, laboratory tests, microbiology results, radiology reports, and discharge diagnosis.

Dataset fields. After processing, each case in the MIMIC-CDM dataset contains the following fields:

- **case_id**: hospital admission identifier from MIMIC-IV.
- **context**: the full patient record from MIMIC-IV that the simulator uses to ground its responses, containing patient history (free-text narrative of presenting complaint and medical, surgical, social, and family history), physical examination (vital signs and exam findings), laboratory tests (mapped from numeric item IDs to human-readable labels), microbiology results (same mapping), and radiology notes (each with modality, region, and exam name). De-identification in MIMIC-IV replaces identifiers such as age, dates, and names with placeholder tokens, preserved as-is.
- **initial_prompt**: the opening patient presentation provided to the agent at the start of each episode, derived from the patient history component of the context.
- **ground_truth**: the discharge diagnosis, used as the target label for evaluation.
- **source**: disease category label (cholecystitis, diverticulitis, or pancreatitis), used for stratified sampling and per-disease analysis.
- **facts**, **fact_criticalness**: atomic clinical facts and their diagnostic importance scores; described in Appendix D.
- **procedures_discharge**: discharge procedures associated with the admission.

Disease selection. The original MIMIC-CDM benchmark includes four abdominal conditions; we exclude appendicitis. Hager et al. [6] report that appendicitis achieves the highest LLM diagnostic accuracy among all four conditions, and attribute this to 82.7% of radiologist reports explicitly stating that the appendix is dilated, enlarged, or fluid-filled—reducing the task to pattern recognition rather than multi-signal reasoning. We focus on the three remaining conditions, which require integrating heterogeneous evidence and are more prone to mutual confusion, making them a more demanding diagnostic reasoning benchmark.

Short-answer filter. To enable reliable LLM judge evaluation, we retain only admissions whose discharge diagnosis contains at most 8 words, filtering the majority of patients that have more than two diagnoses. Long discharge summaries tend to be compound or qualified, making LLM-based grading unreliable.

Sampling and splits. After filtering, we sample 100 cases uniformly at random from each disease stratum (`random_state=42`), yielding 300 balanced cases. We then apply a stratified 70/30 train/test split per disease. To reduce the computational cost of running full multi-turn simulation trajectories, we independently subsample 175 cases from the training partition and 75 from the validation partition (`seed=42`), yielding a final dataset of 250 cases while preserving the 70/30 ratio. The final composition is shown in Table 5. Train and test splits are non-overlapping, enforced by assertion.

Table 5: MIMIC-CDM dataset statistics. Source pool counts reflect primary-diagnosis admissions before filtering. Train and test counts reflect the final 250-case subset after the short-answer filter, stratified sampling, and 70/30 split.

Disease	Source Pool	Train	Test
Cholecystitis	648	58	27
Diverticulitis	257	62	22
Pancreatitis	538	55	26
Total	1,443	175	75

B.2 Rare Disease Dataset

Data source. The dataset comprises 268 cases drawn from published gastrointestinal and hepatobiliary case reports in the clinical literature [10].

Dataset fields. Each case contains the following fields:

- **case_id**: DOI-encoded string identifying the source case report.
- **context**: a structured text document with labeled sections reflecting the sequential disclosure of clinical information as it appears in a case report.
- **initial_prompt**: the opening patient presentation provided to the agent at the start of each episode, derived from the *patient_presentation* section. Mirrors the MIMIC-CDM design.
- **ground_truth**: the diagnosis from the original case report, used as the target label for evaluation.
- **COMMON_UNCOMMON**: rarity label (COMMON or UNCOMMON), used for stratified splitting and per-rarity analysis.
- **facts**, **fact_criticalness**: atomic clinical facts and their diagnostic importance scores; described in Appendix D.

Disease rarity classification. Each case is labeled COMMON (66 cases, 24.6%) or UNCOMMON (202 cases, 75.4%). UNCOMMON cases correspond to rare diagnoses and form the majority of this dataset, making it substantially more challenging than MIMIC-CDM cases.

Splits. All 268 cases are assigned to train or test partitions with no subsampling. Splits are stratified by the COMMON_UNCOMMON label. The final split contains 187 training cases (46 COMMON, 141 UNCOMMON) and 81 test cases (20 COMMON, 61 UNCOMMON).

C Clinical Simulator Details

C.1 Implementation Details

The simulator is implemented using Claude Sonnet 4 as the backbone LLM. It receives the full patient record and the accumulated action–result history as context, and generates a grounded clinical observation in response to each agent action. Separate system prompts are used for the rare disease and MIMIC-CDM datasets to account for dataset-specific formatting and de-identification requirements; both are provided in Appendix N.1.

C.2 Simulator Limitations

A recurring source of stochasticity arises when the agent orders a test not documented in the patient record. When the requested test is absent from the patient record, the simulator generates a plausible result consistent with the patient's presentation and past actions. Since no ground truth value exists for that test, the generated result varies across independent runs.

We illustrate this with a rare disease case: a 22-year-old man presenting after elemental mercury ingestion. The patient record documents vital signs, routine labs, abdominal imaging, and a single mercury measurement: 24-hour urine mercury of 44 µg/L. No serum or blood mercury value was

recorded. Across 7 independent runs, the agent ordered serum or blood mercury, a test not present in the record. Table 6 shows a representative sample of the simulator's responses.

Table 6: Selected simulator responses across 7 runs in which the agent ordered serum or blood mercury for the same case. Simulated values span two orders of magnitude, units are inconsistent across runs, and in one run the simulator returns the urine value from the record relabeled as a serum result.

Simulated value	**Simulator interpretation**
(none)	Within normal limits
2.3 mcg/L	Within normal limits; no acute toxicity
12 µg/L	Mildly elevated; consistent with poor GI absorption
15 ng/mL	Mildly elevated *(same number as below; different unit)*
44 µg/L	Elevated *(urine value from record, relabeled as serum)*
125 ng/mL	Markedly elevated; confirms acute exposure
185 ng/mL	Markedly elevated; immediate intervention required

A second case illustrates a related but distinct failure pattern: for the same absent test, the simulator's behavior is inconsistent across runs, sometimes declining to return a result and sometimes simulating one.

The case is a MIMIC-CDM admission with a gold diagnosis of gallstone pancreatitis with choledocholithiasis. The patient record contains an MRCP but no CT with contrast. When the agent ordered CT abdomen with contrast, the simulator refused in one of four runs and simulated a result in the remaining three. Table 7 shows the four runs.

Table 7: Four runs in which the agent ordered CT abdomen with contrast for the same MIMIC-CDM case. In the two runs that simulate without reporting gallstones (rows 2–3), the agent reaches an incorrect diagnosis; the run that simulates inflammatory changes without explicitly excluding gallstones (row 4) produces a correct diagnosis.

Behavior	**Simulated CT finding**
Declined	CT not present in record
Simulated	Gallbladder wall thickening, CBD dilation ∼7–8 mm, PD dilation ∼3–4 mm; *no gallstones or choledocholithiasis identified*
Simulated	Pancreatic parenchyma with decreased enhancement, PD dilation; *no gallstones identified*
Simulated	Gallbladder wall thickening with inflammation, CBD ∼7–8 mm, PD ∼3–4 mm, pancreatic inflammatory changes

Together, these two cases illustrate the main sources of stochasticity when the agent orders a test absent from the record. In the mercury case, the simulated result varies widely across runs. In the CT case, the simulator responds inconsistently to the same request across runs. Both are most pronounced when the record is sparse relative to what a thorough agent might order. The criticality recall reward partially mitigates the downstream effect, since facts are extracted from the patient record and simulated test results do not correspond to any entry in the fact list, yielding no reward signal and limiting the extent to which the agent is reinforced for pursuing absent tests.

We explored an alternative simulator configuration that strictly declines any test not documented in the record, but found that the resulting observations were too sparse to sustain meaningful exploration during training, as agents received little signal to guide subsequent actions. The current design therefore accepts some stochasticity in exchange for richer exploratory signal. This stochasticity is inherent to the open-ended simulation setting. While simulated results for absent tests vary across runs, the validation below shows that they remain clinically reasonable in practice.

C.3 Simulator Validation

We evaluated simulator output quality on the full test set of 156 cases (75 MIMIC-CDM, 81 Rare Disease). Two physician raters reviewed each step of each trajectory and flagged any step meeting one or more of the following criteria:

- **(1) Contradicts established case information** — introduces facts that directly conflict with what has already been presented (e.g., the HPI states no prior surgeries, but the physical exam result mentions a visible laparotomy scar).
- **(2) Clinically impossible or implausible** — values or findings that could not realistically occur (e.g., sodium of 280 mEq/L; pulse oximetry of 15% in an awake, conversational patient).
- **(3) Wrong result for the ordered test** — the result does not correspond to the test ordered (e.g., a CBC was ordered but a lipid panel was returned; an MRI Brain was ordered but an echocardiogram report is returned).
- **(4) Internally inconsistent** — the result contradicts itself within the same response (e.g., "Chest X-ray shows no pneumothorax" in the findings but "large left pneumothorax" in the impression; or "WBC is within normal limits (38,000/μL)").

Table 8: Simulator step validation. A step is flagged if it meets any of the four quality criteria.

Dataset	Cases	Steps	Step-level		Flag type (steps)			
			Flagged	Rate	(1)	(2)	(3)	(4)
MIMIC-CDM	75	155	4	2.6%	0	0	0	4
Rare Disease	81	239	3	1.3%	3	0	0	0
All	156	394	7	1.8%	3	0	0	4

Each case consists of multiple simulator steps (one per diagnostic action taken by the agent), so the step count exceeds the case count. MIMIC-CDM cases average 2.1 steps per case (155 steps across 75 cases) and rare disease cases average 3.0 steps per case (239 steps across 81 cases). Flag types (1)–(4) correspond to the four quality criteria listed above.

Of 394 total steps, 7 were flagged (1.8%). Critically, the simulator never produced clinically impossible values or wrong test results, the failure modes that would corrupt training signal. The 7 flagged steps were minor consistency errors (criteria 1 and 4). These results support the validity of the simulator as a training environment.

Sensitivity to the simulator. The simulator is held fixed across every model and reward variant. To test whether the gain is specific to one simulator configuration, we re-evaluated the Qwen3.6-27B base model and its TraceDx-trained counterpart (Appendix H.2) on Rare Disease under a different simulator backbone and under a paraphrased simulator prompt, changing one variable at a time (Table 9). Absolute accuracies shift slightly with the simulator, but the margin of TraceDx over its untrained base is preserved in every tested condition.

Table 9: Diagnostic accuracy of Qwen3.6-27B on Rare Disease under alternative simulator configurations.

Simulator condition	Base	TraceDx
Original (Sonnet 4, original prompt)	0.333	0.382
Swapped backbone (GPT-5.1)	0.320	0.395
Paraphrased prompt (Sonnet 4)	0.333	0.407

D Fact Extraction and Criticality Scoring

D.1 Atomic Fact Extraction

Facts are extracted offline before any training and stored directly in the dataset, avoiding LLM calls at training time. The two datasets differ substantially in their data structure and information density, which motivates different extraction strategies. Rare disease cases are short narrative case reports with a mean of 9.7 facts per case (median 8.5, std 5.5, range 1–35), making a single prompt sufficient to process the full context. MIMIC-CDM cases are structured records with a mean of 131.1 facts per case (median 122.5, std 46.6, range 41–401). Table 10 summarizes the fact statistics for both datasets.

Table 10: Fact extraction statistics for both datasets. The difference in mean facts per case reflects the difference in source data: rare disease cases are short case reports, while MIMIC-CDM cases are full records with structured laboratory, radiology, and examination data.

Metric	**Rare Disease** (268 cases)	**MIMIC-CDM** (250 cases)
Total facts	2,609	32,772
Mean / case	9.7	131.1
Median / case	8.5	122.5
Std	5.5	46.6
Min / Max	1 / 35	41 / 401

Rare Disease. Each case is processed with a single LLM call that applies a general-purpose extraction prompt to the full context field. The prompt instructs the model to extract objective findings from diagnostic testing (imaging, endoscopy, pathology, laboratory results) and physical examination, while excluding patient initial presentation, demographic details, and interpretations of test results. Multi-fact sentences are split into atomic items, where each item is a single observation stated without inference. The extraction model is Claude Sonnet 4.5 (prompt in Appendix N.2).

MIMIC-CDM. The MIMIC-CDM context is structured as a JSON record with distinct clinical modalities, so extraction uses three separate strategies. The extraction model is Claude Sonnet 4.5 for both physical examination and radiology (prompts in Appendix N.2).

Physical examination. A dedicated physical examination prompt is applied to the examination field of the context. Included findings are vital signs and explicit exam findings from auscultation, palpation, and inspection, along with negative findings when explicitly stated. Demographics, history, symptoms, and interpretive conclusions are excluded.

Radiology. Each radiology report in the case is sent to the LLM independently, one call per report. Isolating reports prevents findings from one imaging study from suppressing or contaminating those of another when multiple studies are present. The radiology prompt extracts both positive and negative findings and quantitative measurements, while excluding interpretive conclusions.

Laboratory tests and microbiology. Lab and microbiology values are stored as structured key-value dictionaries in the MIMIC context and are extracted deterministically without an LLM call.

The three modality fact lists are concatenated in order (physical examination, radiology, laboratory and microbiology) to form the final fact list for each case.

D.2 Fact Criticality Scoring

For both datasets, we use GPT-5.1 to score each fact on a 0–3 integer scale, scoring for how critical the fact is in confirming the gold diagnosis rather than for what diagnosis it suggests. The prompts for both datasets are provided in Appendix N.4 and Appendix N.3.

Rare disease scoring. Rare disease cases use a simpler prompt with no injected diagnostic criteria. At an average of 10 facts per case drawn from curated case report workups, the scoring task is smaller and less ambiguous. Each case is a distinct rare condition with its own unique presentation, so

disease-specific criteria would not generalize across cases. The scoring scale uses general definitions without disease-specific examples.

MIMIC-CDM scoring. MIMIC cases require a more carefully designed prompt because of scale and ambiguity. At an average of 131 facts per case, most of which are routine lab values and negative imaging findings, the model needs specific clinical anchors to distinguish signal from noise. The three diseases also share overlapping acute abdominal presentations, making it easy for the model to apply the wrong scoring logic.

To address this, the MIMIC prompt injects a disease-specific diagnostic criteria block selected by the case's disease label (cholecystitis, diverticulitis, or pancreatitis), drawn from established clinical guidelines. The gold diagnosis takes priority over the injected criteria, which serve as a reference framework rather than a strict checklist. The full criteria blocks are provided in Appendix N.4. Table 11 summarizes the diagnostic structure for each disease category.

Table 11: Comparison of diagnostic structures across disease categories.

Disease	Diagnostic Rule	Clinical Signs	Lab Evidence	Imaging Role
Cholecystitis	Local + systemic signs; definitive diagnosis requires imaging (Tokyo Guidelines)	Murphy sign, right upper quadrant pain/tenderness	Fever, elevated WBC, elevated CRP	Required for definitive diagnosis
Diverticulitis	Clinical suspicion with imaging confirmation	Left lower quadrant pain, tenderness, fever, history	Elevated CRP (supporting)	CT typically used to confirm diagnosis and assess complications
Pancreatitis	Diagnosis requires ≥ 2 of: abdominal pain, enzyme elevation, or imaging findings	Upper abdominal pain	Amylase or lipase $\geq 3\times$ normal	Optional confirmation if diagnosis uncertain

The MIMIC prompt also provides explicit guidance on negative and absent findings, which the radiology extraction pipeline preserves in large volume. Absent findings (e.g., "no free air", "normal appendix") default to score 0, but score 1 when the absence specifically confirms the variant stated in the gold diagnosis. For example, "no free air" scores 1 for uncomplicated diverticulitis (confirms the uncomplicated nature).

Score distributions. Figure 7 shows the observed score distributions. The 88.1% zero rate in MIMIC-CDM reflects that routine laboratory panels and negative CT findings are correctly identified as non-confirmatory. The more balanced rare disease distribution (31% score 0) reflects that case report contexts are already curated to contain the clinically relevant workup.

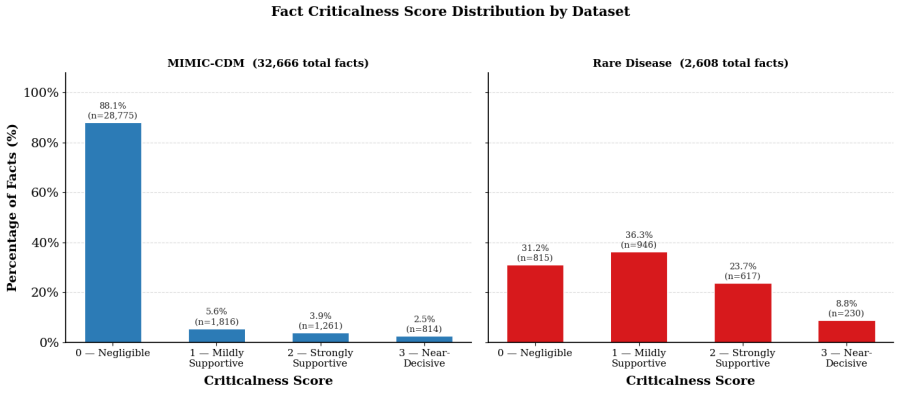

Figure 7: Criticalness score distributions for both datasets. MIMIC-CDM is heavily zero-skewed (88% score 0) because most extracted facts are routine lab values; rare disease cases have a more balanced profile (31% score 0) because case reports contain only the clinically relevant workup.

D.3 Criticality Scorer Validation

We validated the automated criticality scorer against two physician raters on 156 cases (75 MIMIC-CDM, 81 Rare Disease). For each fact in each case, the physicians indicated whether they agreed

with the assigned score. Disagreement rates were 1.0% of facts on MIMIC-CDM and 5.7% on Rare Disease (1.3% overall). The higher rate on Rare Disease partly reflects sample size, making the rate more sensitive to individual disagreements. These results indicate that the automated scorer is well-validated against physician judgment across both datasets.

Agreement across model families. To test whether criticality scores depend on the choice of scorer, we re-scored the entire rare disease fact set with three independent scorers from different model families (GPT-5.1, Gemini 2.5 Pro, and Claude Sonnet 4.5), plus a second GPT-5.1 run as a within-model reference. Both Krippendorff's α and quadratic-weighted κ are at or above conventional thresholds for reliable agreement, and disagreements are almost never larger than one point (Table 12). Agreement across model families (κ 0.78–0.82) is close to a model's agreement with its own rerun (κ 0.85), so changing the scorer family costs little beyond sampling noise.

Table 12: Criticality score agreement across independent scorers on the Rare Disease fact set.

Measure	**Value**
Krippendorff's α (ordinal, 3 independent scorers)	0.80
Pairwise quadratic-weighted κ	0.78–0.85
Scores within ± 1 point	97–99%
Exact agreement on hallmark ($c = 3$) facts	92–96%

Robustness to scorer error. Agreement shows that scorers rarely err, and we also test what happens when they do. Using the Qwen3.6-27B setup of Appendix H.2, we retrain on Rare Disease with ± 1 noise added to the criticality scores of a random 20% of facts. This policy reaches 0.370, retaining most of the gain of the full method (0.382) over the untrained base (0.333), whereas randomizing all scores returns accuracy to base. The reward needs the scores to point at the right facts, not to be exact. Proposition 1 formalizes this, showing that criticality recall preserves the ordering of trajectories under bounded scoring error.

E Evaluation Protocol

E.1 LLM Judges

Two separate LLM graders are used, one for diagnostic accuracy and one for criticality recall.

Diagnosis grader. Final diagnoses are evaluated using Claude Sonnet 4 as the grading model. Given the LLM-generated diagnosis and the ground truth discharge diagnosis, the grader reasons over clinical equivalence and outputs a binary score (1.0 correct, 0.0 incorrect). For MIMIC-CDM, discharge procedures are also provided as supporting evidence for severity and treatment intent.

Fact matching grader. Criticality recall requires determining which atomic facts from the scored fact set were discovered during a trajectory. This matching is performed using google/gemini-3-flash-preview. Given the full interaction history and the list of atomized facts, the grader identifies which facts were revealed during the trajectory and returns their indices, used to compute criticality recall as described in Section 4.

The full prompts for both graders are provided in Appendix N.5.

E.2 Diagnosis Grader Prompt Design

The two datasets require different levels of grader design effort.

Rare Disease grader. Rare disease labels are highly specific and largely unambiguous. Conditions such as leiomyosarcoma or polycystic liver disease admit no clinically meaningful near-miss. The differential diagnosis space is so large that any wrong answer is a complete miss. Grading therefore reduces to a straightforward identity check of whether the model identified the correct rare condition.

That said, rare diseases can carry multiple valid names in the clinical literature, and the grader must still recognize terminological equivalence. For example, *emphysematous gastritis* and *gastric cystic pneumatosis* describe the same condition (gas within the gastric wall), and a prediction of one should be accepted when the ground truth uses the other. The grader prompt is designed to handle this

through semantic reasoning rather than string matching, without requiring the elaborate clinical equivalence rules needed for MIMIC-CDM.

MIMIC-CDM grader. MIMIC-CDM presents a harder grading problem despite its terse labels. The three conditions (cholecystitis, diverticulitis, pancreatitis) are clinically related acute abdominal presentations that can overlap, and the ground truth labels are short phrases that hide substantial clinical variation. Terms such as *acute cholecystitis*, *calculous cholecystitis*, and *acute calculous cholecystitis* all refer to the same condition, while *biliary colic* and *gangrenous cholecystitis* do not, yet all are superficially similar. The grader must reason about lexical variation, specificity, severity, and clinical equivalence before it can answer the basic yes/no question. This is why the MIMIC-CDM grader was designed through physician consultation and validated on an explicit test suite, as described below.

The diagnosis grader prompt was developed through consultation with a physician to resolve ambiguous grading cases that arise from the open-ended label space. The core grading question is whether a clinician acting on the predicted diagnosis would arrive at the same treatment decisions as one acting on the ground truth. The following principles were established:

- **Specificity differences are only meaningful if they change treatment.** A prediction more or less specific than the gold is not automatically incorrect. *Sigmoid diverticulitis* and *acute diverticulitis of the sigmoid colon* are both acceptable for a gold of *acute diverticulitis* since management is identical, but predicting *acute cholecystitis* for a gold of *gangrenous cholecystitis* is incorrect because gangrenous cholecystitis requires urgent surgical intervention.

- **Related diseases are not equivalent.** Conditions that share a mechanism or anatomical region but have distinct management pathways are graded incorrect. *Biliary colic* and *symptomatic cholelithiasis* are not acceptable for a gold of *acute cholecystitis*, nor is *choledocholithiasis with biliary obstruction*, which involves the bile duct rather than the gallbladder.

- **Extra diagnoses are acceptable unless they redirect care.** If a prediction correctly identifies the primary diagnosis and appends additional conditions, it is graded correct provided the added diagnosis does not redirect care to a fundamentally different pathway. Adding a closely related complication within the same organ system is not sufficient grounds to fail a prediction.

- **For multi-diagnosis gold labels, the primary diagnosis must be present.** When the gold contains multiple conditions, the prediction must cover the primary (first-listed) one. Predicting only a secondary condition while missing the primary is incorrect. Procedures are used as soft evidence for severity and treatment intent but do not override gold label ordering to reclassify which diagnosis is primary.

Table 13 illustrates these principles with representative cases.

Table 13: Selected grading cases across five categories (17/17 passed). Grade 1.0 = correct, 0.0 = incorrect. Etiology matters only when it changes the treatment pathway.

Gold	Prediction	Grade	Reasoning
Specificity & Severity			
Gangrenous cholecystitis	Acute cholecystitis	**0.0**	Misses severity; gangrenous carries high mortality and requires urgent surgery
Perforated diverticulitis	Acute sigmoid diverticulitis with mild inflammation	**0.0**	Severity downgrade; perforation confirmed by laparotomy
Diverticulitis	Perforated diverticulitis with pneumoperitoneum	**0.0**	Severity upgrade; implies emergency surgery when gold warrants antibiotics only
Diverticulitis	*[verbose paragraph concluding:]* acute uncomplicated sigmoid diverticulitis	**1.0**	Same diagnosis regardless of narrative framing; management unchanged
Related Diseases			
Acute cholecystitis	Symptomatic cholelithiasis (biliary colic)	**0.0**	Distinct disease; biliary colic is managed expectantly, not surgically
Extra Diagnoses			
Acute cholecystitis	Acute calculous cholecystitis with mild gallstone pancreatitis	**1.0**	Extra complication in same biliary domain; does not redirect care
Acute alcoholic pancreatitis	Acute alcoholic pancreatitis with suspected pancreatic adenocarcinoma	**0.0**	Added cancer diagnosis redirects to fundamentally different clinical pathway
Multi-Diagnosis			
Acute cholecystitis + cholelithiasis	Acute cholecystitis	**1.0**	Primary diagnosis covered; cholelithiasis is implied by calculous cholecystitis and does not change management
Etiology			
Gallstone pancreatitis	Acute pancreatitis	**0.0**	Gallstone etiology omitted; triggers cholecystectomy, changing management
EtOH pancreatitis	Acute pancreatitis	**1.0**	EtOH etiology omitted; acute management identical (bowel rest, IV fluids, NPO)

E.3 Grader Validation

We validated the diagnosis grader against two physician raters on a sample of 143 cases (71 MIMIC-CDM, 72 Rare Disease). Physician availability constrained the sample size. For each case, the physicians reviewed the LLM grader's binary correct/incorrect judgment and indicated whether they agreed with it.

Table 14: Diagnosis grader validation against physician rater.

Dataset	N	Agree	Disagree	Agreement %
MIMIC-CDM	71	64	7	90.1
Rare Disease	72	67	5	93.1
All	143	131	12	91.6

Agreement is 91.6% overall (90.1% on MIMIC-CDM, 93.1% on Rare Disease), indicating that the grader's binary judgments are consistent with physician assessment across both datasets.

Reward stability. Fact extraction and criticality scoring happen once per case before training, but the fact-matching grader runs on every rollout. To check that this does not make the training target noisy, we ran the fact-matching grader five times over the Rare Disease held-out set (Table 4 in Section 5.5). Most cases receive identical criticality recall and final reward across all five runs. The remaining variance comes from cases with few facts, where criticality recall is a ratio over a small denominator.

F Baseline Implementation Details

F.1 Closed-Weight Model Evaluation

Diagnostic loop. Each evaluation episode runs the policy LLM against the clinical simulator (Claude Sonnet 4, held fixed across all experiments) for at most 7 turns. The policy never has direct access to the patient record. It observes only the action–result history that accumulates turn by turn. At each turn it must output exactly one line, either "Diagnosis: <diagnosis>" to commit to a final answer and terminate the episode, or "Next Action: <action>" to request the next clinical test. The action selection prompt is provided in Appendix N.6.

Post-hoc differential. A structured differential diagnosis is generated *once* at loop termination by prompting the policy on the completed history. If the agent has not committed to a diagnosis after 6 actions, the top entry of the differential is taken as the fallback final diagnosis. The differential generation prompt is provided in Appendix N.6.

Models evaluated. The policy is instantiated with each of the following models accessed via OpenRouter: GPT-5, GPT-5.1, GPT-5.4, Claude Sonnet 4, Claude Sonnet 4.5, Claude Opus 4.7, Gemini 2.5 Pro, and Gemini 3.1 Pro. The simulator is always Claude Sonnet 4 regardless of which model serves as the policy.

F.2 Oracle Context

The oracle context baseline evaluates GPT-5.1 in a one-shot setting where the full patient record is provided directly, with no sequential workup and no simulator. It establishes an upper bound on diagnostic accuracy. The model sees the complete information state a clinician would have access to, so any gap relative to the simulation pipeline reflects the cost of having to acquire evidence interactively.

The full patient context field is passed directly to the differential diagnosis prompt (the same Top-K framing used in the action simulation, see Appendix N.6). The model produces a structured differential, and the top entry is taken as the final diagnosis. No simulator is invoked and steps taken is always 0.

G Training Details

All training is run through Tinker [11], the managed training API from Thinking Machines, which handles GPU allocation and orchestration. We therefore report training configuration (LoRA rank, GRPO settings, reward parameters, optimizer) rather than raw hardware specifications.

G.1 Models

We fine-tune two base models, Kimi-K2.5 and Qwen3-32B, both with LoRA rank 32. Both are trained with GRPO on the same reward function. Dataset-specific and model-specific differences are noted below where they arise.

G.2 GRPO Configuration

Each GRPO update step samples 6 patient cases and draws 8 rollouts per case (48 effective rollouts per step). Both models are trained for 2 epochs on MIMIC-CDM and 1 epoch on the rare disease dataset. The agent and simulator prompts used during rollouts are provided in Appendix N.7.

G.3 Reward Parameters

Table 15 lists the concrete parameter values used in the reward function $R(\tau) = (\alpha \cdot CR(\tau) + \beta)\,\mathrm{acc}(\hat{d}, d^*) - \lambda T/T_{\max}$ across all configurations. Diagnosis correctness is graded by Claude Sonnet 4 (binary), and criticality recall is computed by google/gemini-3-flash-preview.

Table 15: Reward function parameter values.

Parameter	Description	Value
α	Criticality recall multiplier on correct diagnosis	1.0
β	Base reward for correct diagnosis	0.5
λ	Step penalty weight	0.05
Malformed penalty	Subtracted for invalid output format	0.3

The step penalty is a regularizer, not a model of clinical cost. It gives the agent a reason to commit once the evidence supports a diagnosis rather than ordering tests until the turn limit, and it charges every action the same amount. Real diagnostic actions differ widely in expense, risk, invasiveness, and turnaround, so a physical exam, a serum lipase, an abdominal CT, and a HIDA scan are far from equivalent. Modeling these costs is a separate question from the one this work addresses, which is what makes a piece of evidence worth acquiring. In an open-ended action space it is also a research problem in its own right, since the agent can name any test in any phrasing and a fixed cost table has no entry to match against. We leave cost-aware evidence acquisition as an open question.

G.4 Optimizer and Generation

Both models use AdamW. Kimi-K2.5 uses a constant learning rate of 5×10^{-6}, and Qwen3-32B uses a constant learning rate of 1×10^{-5}. During training, rollouts are sampled at temperature 0.95, and validation uses greedy decoding. Maximum new tokens per turn is 6,144 for both Kimi-K2.5 and Qwen3-32B. Episodes terminate when the agent issues a diagnosis, produces malformed output, or exceeds the maximum of 7 turns.

G.5 Evaluation

At test time, both RL-trained models are evaluated under greedy decoding (temperature 0.0) with the same simulator, episode structure, and turn limit as during training. The untrained base model (reported as *Base* in the results) is evaluated on the identical setting with the same prompts, simulator, and 7-turn maximum. This ensures that any difference in performance between the base and RL-trained models is attributable solely to the policy update and not to differences in inference configuration.

H Additional Reward Ablations

H.1 Frequency-Based Reward Baselines

The two frequency-based baselines in Section 5.3 replace clinical criticality with rarity, defining the value of a piece of information by how rarely it is recovered. We ran 10 frontier-model trajectories on each MIMIC-CDM training case, two each from Gemini 2.5 Pro, Claude Sonnet 4, Claude Sonnet 4.5, GPT-5, and GPT-5.1, and counted how many runs recovered each fact.

Linear IDF weights each fact by its inverse recovery frequency, so facts that frontier models rarely recover count most. This weighted fact recall takes the place of criticality recall in the reward.

Band frequency groups facts into high-, mid-, and low-frequency bands and rewards recall within each band separately, alongside diagnostic accuracy and the step penalty.

H.2 Fact-Weighting Ablations on Qwen3.6-27B

To separate the contribution of graded criticality from that of dense fact-level feedback, we train additional reward variants on a third base model, Qwen3.6-27B. All variants share the same base model, simulator, fact pipeline, and reward equation, and differ only in how facts are weighted. *Dx-only* removes the fact term entirely. *Random* assigns each fact a random criticality score. *Unweighted* gives every fact equal weight. *Binary* collapses the scale to critical or non-critical.

TraceDx is the strongest variant on both datasets (Figure 8). Unweighted fact recall recovers about half of the TraceDx gain on Rare Disease and a fifth on MIMIC-CDM. Random gives no consistent benefit, and Binary helps on MIMIC-CDM but falls below the base model on Rare Disease. Dense fact supervision alone therefore does not explain the gain. The graded structure of the criticality scores contributes beyond it.

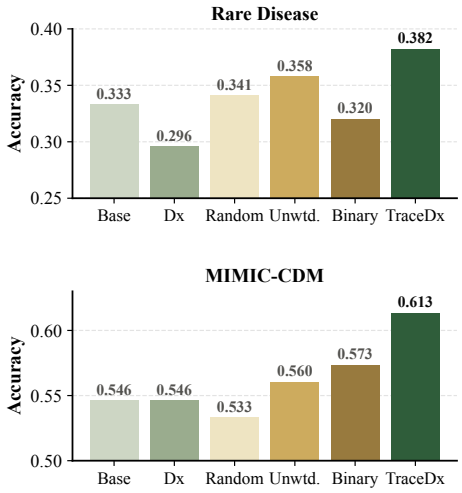

Figure 8: Fact-weighting ablation for Qwen3.6-27B (Unwtd. = Unweighted).

We do not include an information-gain baseline. Information gain measures how much an action reduces uncertainty over the diagnosis, which requires an explicit belief distribution over candidate diagnoses. Our environment has no such distribution, since actions and diagnoses are free text with no predefined label space, so information gain is not directly defined without an auxiliary model to maintain beliefs. We leave this comparison to future work.

I Per-Agent Evidence-Quality Analysis

The regression in Section 5.4 pools all twelve agents, including the two trained with TraceDx. To check that the association between hallmark evidence and diagnostic success does not depend on our reward, we compute it separately within each of the ten agents never trained on it. For each agent, we take the Spearman correlation between Crit-3 ratio and diagnostic success across its 81 rare-disease cases, with the diagnosis again produced by the fixed GPT-5.1 reasoner. The association is positive and significant in all ten agents (Table 16). None of these agents has ever seen a criticality score, so the link between hallmark evidence and diagnostic success holds independently of our training signal.

Table 16: Within-agent Spearman correlation between Crit-3 ratio and diagnostic success on Rare Disease ($n = 81$ per agent), for agents never trained on the TraceDx reward.

Agent	**Spearman** ρ	p
Kimi-K2.5 (Base)	$+0.55$	< 0.001
GPT-5.1	$+0.52$	< 0.001
GPT-5	$+0.51$	< 0.001
Gemini 2.5 Pro	$+0.51$	< 0.001
Sonnet 4.5	$+0.50$	< 0.001
Sonnet 4	$+0.48$	< 0.001
Gemini 3.1 Pro	$+0.47$	< 0.001
Opus 4.7	$+0.40$	< 0.001
Qwen3-32B (Base)	$+0.39$	< 0.001
GPT-5.4	$+0.38$	< 0.001

J Proof of Proposition 1

Proposition (Proposition 1, restated)**.** *Assume $\sum_{f\in F} c(f) > 0$. Let each fact $f \in F$ have latent relevance $r_f \geq 0$, and suppose the scorer satisfies*

$$c(f) = \kappa r_f + \varepsilon_f, \quad \kappa > 0,\ |\varepsilon_f| \leq \sigma,$$

where $\sigma \geq 0$ is the worst-case per-fact scoring error. For two trajectories $\tau^{(1)}, \tau^{(2)}$ from the same patient case (sharing fact set F) with discovered sets $D_i = D(\tau^{(i)})$, define $\nu(D) = \sum_{f\in D} r_f$. If

$$\nu(D_1) - \nu(D_2) > \frac{\sigma}{\kappa}(|D_1| + |D_2|),$$

then $CR(\tau^{(1)}) > CR(\tau^{(2)})$. Furthermore, if $\alpha > 0$, both trajectories are correct, and $CR(\tau^{(1)}) - CR(\tau^{(2)}) > \frac{\lambda}{\alpha} \cdot \frac{T_1 - T_2}{T_{\max}}$, where T_i is the number of actions in $\tau^{(i)}$, then $R(\tau^{(1)}) > R(\tau^{(2)})$.

Proof. **Part 1: $CR(\tau^{(1)}) > CR(\tau^{(2)})$.** Expanding the weighted sums using $c(f) = \kappa r_f + \varepsilon_f$:

$$\sum_{f\in D_1} c(f) - \sum_{f\in D_2} c(f) = \kappa\big(\nu(D_1) - \nu(D_2)\big) + \sum_{f\in D_1} \varepsilon_f - \sum_{f\in D_2} \varepsilon_f.$$

Since $|\varepsilon_f| \leq \sigma$, the error terms satisfy $\sum_{f\in D_1} \varepsilon_f \geq -\sigma|D_1|$ and $-\sum_{f\in D_2} \varepsilon_f \geq -\sigma|D_2|$, so

$$\sum_{f\in D_1} c(f) - \sum_{f\in D_2} c(f) \ \geq\ \kappa\big(\nu(D_1) - \nu(D_2)\big) - \sigma(|D_1| + |D_2|).$$

Since $\kappa > 0$, the hypothesis gives $\kappa\big(\nu(D_1) - \nu(D_2)\big) > \sigma(|D_1| + |D_2|)$, so the right-hand side is strictly positive. Since $CR(\tau^{(i)}) = \frac{\sum_{f\in D_i} c(f)}{\sum_{f\in F} c(f)}$ and the denominator is shared and positive by assumption, it follows that $CR(\tau^{(1)}) > CR(\tau^{(2)})$.

Part 2: $R(\tau^{(1)}) > R(\tau^{(2)})$. For correct trajectories ($\mathrm{acc}(\hat{d}_i, d^*) = 1$) with action counts T_1 and T_2, expanding the reward (noting that β cancels) gives

$$R(\tau^{(1)}) - R(\tau^{(2)}) = \alpha\big(CR(\tau^{(1)}) - CR(\tau^{(2)})\big) - \lambda\,\frac{T_1 - T_2}{T_{\max}}.$$

Since $\alpha > 0$ and by hypothesis $CR(\tau^{(1)}) - CR(\tau^{(2)}) > \frac{\lambda}{\alpha} \cdot \frac{T_1 - T_2}{T_{\max}}$, it follows that $R(\tau^{(1)}) - R(\tau^{(2)}) > 0$. In particular, when Part 1 holds the CR gap is positive, so Part 2 requires only that this gap not be overwhelmed by the step-penalty difference. □

Remark. Two aspects of the proposition deserve comment. First, on the step-penalty term: with $\alpha = 1.0$, $\lambda = 0.05$, and $T_{\max} = 7$, the maximum value of $\frac{\lambda}{\alpha} \cdot \frac{T_1 - T_2}{T_{\max}}$ is 0.05, so the reward ordering is overturned only when two correct trajectories differ in criticality recall by less than five percentage points. Second, on the scoring noise: since $c(f) \in \{0, 1, 2, 3\}$, the worst-case per-fact error is $\sigma \leq 3$ (a scorer that assigns 0 to every hallmark fact, or 3 to every irrelevant one), and the Part 1 sufficient condition becomes harder to satisfy as σ grows. The bound is vacuous at $\sigma = 3$ and most informative when σ is well below this maximum, for instance $\sigma = 1$ (the scorer is at most one level off on any fact). Empirical validation of the scorer is reported in Appendix D.2.

K Failure Case Study

We examine two cases where baselines reach the correct diagnosis but TraceDx does not. Together they mark the boundary of what a criticality reward supervises.

Case: Gastric metastasis from uveal melanoma

A 55-year-old man with a history of uveal melanoma treated 14 years prior presented for routine follow-up. Imaging revealed a T1-hyperintense gastric wall lesion, and endoscopy with biopsy confirmed melanoma. Uveal melanoma is known to metastasize years or even decades after treatment,

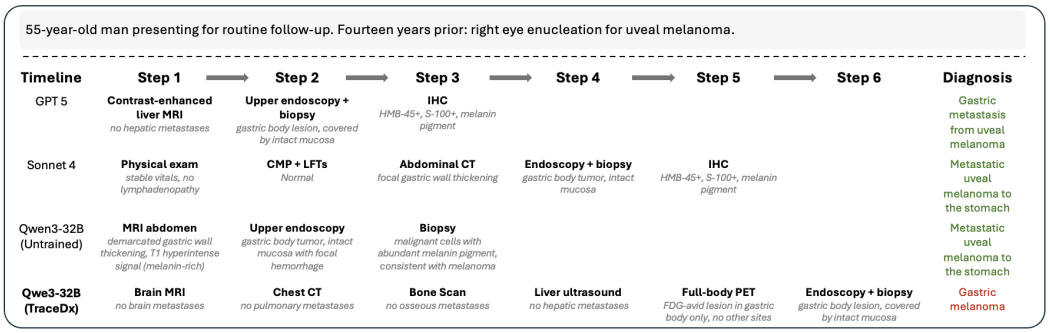

Figure 9: Failure case: diagnostic trajectories for gastric metastasis from uveal melanoma. The grey box (top) shows the initial patient presentation; trajectories below show actions and results. Baseline models correctly anchor on the patient's prior uveal melanoma and reach the correct diagnosis. TraceDx initiates a broad staging workup, finds no distant metastases, and incorrectly concludes primary gastric melanoma, failing to integrate the patient's history as a prior constraint.

most commonly to the liver but also to the gastrointestinal tract. The central diagnostic question was therefore whether this lesion represented delayed metastatic disease rather than a new primary tumor.

Baseline models correctly anchored on the patient's history and interpreted the gastric lesion as metastatic uveal melanoma. They used imaging and biopsy to confirm the presence of melanoma but consistently treated the prior malignancy as the most likely source, leading to the correct diagnosis.

In contrast, the trained model initiated a comprehensive staging workup, including brain MRI, chest CT, bone scan, liver imaging, and PET. These tests showed no evidence of disease outside the stomach. Based on this, the model concluded that the gastric lesion represented a primary melanoma. However, primary gastric melanoma is extraordinarily rare and is a diagnosis of exclusion that requires ruling out all known primary sites, including prior uveal melanoma.

This failure reflects a breakdown in information integration rather than knowledge or execution. The model correctly identified melanoma on biopsy and performed an appropriate staging protocol, but did not incorporate the patient's prior uveal melanoma into its reasoning. Instead, it treated the case as a new malignancy and used the absence of widespread disease to infer a primary origin.

Notably, the error was driven by otherwise correct actions. The staging workup was clinically appropriate, but by prioritizing it over the patient's history, the model allowed the absence of metastases to override a far stronger prior. In this case, more comprehensive testing led to a less accurate conclusion. The model did not fail to gather information. It failed to use the most important information it already had.

Case: Acute gallstone pancreatitis

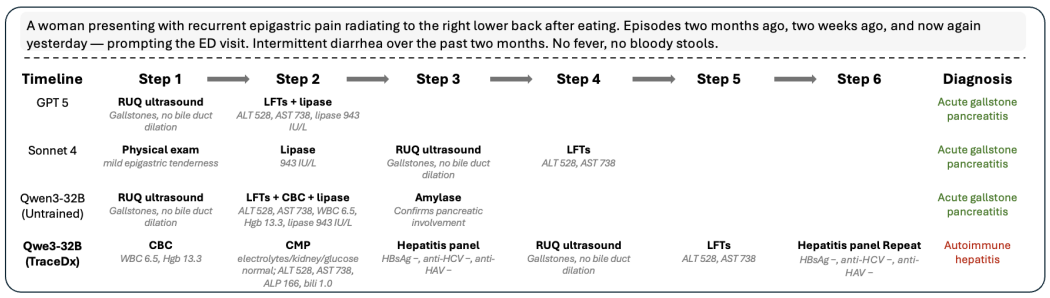

Figure 10: Failure case: diagnostic trajectories for acute gallstone pancreatitis. The grey box (top) shows the initial patient presentation; trajectories below show actions and results. Baseline models rapidly identify gallstones and measure lipase, correctly diagnosing gallstone pancreatitis. In contrast, the trained model follows a hepatology-focused workup, repeatedly evaluating for hepatitis and never measuring lipase, leading to an incorrect diagnosis of autoimmune hepatitis.

A woman presented with recurrent post-prandial epigastric pain radiating to the right lower back, a pattern suggestive of biliary disease. Imaging demonstrated gallstones, and laboratory results showed markedly elevated transaminases (ALT 528 IU/L, AST 738 IU/L). The central diagnostic question was whether these abnormalities reflected primary liver disease or secondary effects of biliary obstruction involving the pancreas.

Baseline models followed a short and targeted reasoning chain. After identifying gallstones, they measured serum lipase, which was markedly elevated (943 IU/L, nearly 16× the upper limit of normal), confirming pancreatic involvement and leading directly to the diagnosis of acute gallstone pancreatitis. In each case, the models prioritized the pancreatic hypothesis early and selected the test that could resolve it.

The trained model instead pursued a hepatology-focused workup. Markedly elevated transaminases alongside gallstones are a legitimate reason to investigate primary liver disease, and ruling out viral hepatitis is a clinically reasonable step. The model did this correctly. But it never closed the loop on pancreatic involvement, and without a lipase measurement, the elevated liver enzymes became the dominant signal, leading to a conclusion of autoimmune hepatitis.

The failure is therefore one of completeness rather than reasoning. The hepatology workup was not wrong, but it was pursued to the exclusion of a simpler test that would have resolved the case. In gallstone pancreatitis, elevated transaminases are a secondary effect of transient biliary obstruction, and measuring lipase is the step that distinguishes this from primary liver disease. The trained model never took that step.

The two failures differ in kind. In the melanoma case, the model gathered the right evidence and drew the wrong inference from it, which is a failure of integration. In the pancreatitis case, it never ordered lipase, so the failure arose during evidence acquisition. Acquisition errors are the class TraceDx is designed to reduce, and this case shows that it reduces them without eliminating them. Integration errors lie outside what the reward supervises. Extending process supervision to hypothesis maintenance and evidence integration remains an open question that these cases motivate.

L Physician Baseline

To contextualize model performance, we evaluated physicians playing the same simulator game as the models, subject to the same turn limit and interface. Due to physician time constraints, evaluation was conducted on a smaller subset of the held-out test set. Physicians interacted with the clinical simulator by issuing free-text diagnostic actions and committing to a final diagnosis, with no access to the patient record.

Table 17: Physician and model accuracy on randomly sampled subsets of each test set under the same simulator game. Rare Disease: 60 cases, 2 physicians (30 each). MIMIC-CDM: 30 cases, 1 physician.

Model	Rare Disease	MIMIC-CDM
Physician	31.7%	76.7%
Opus 4.7	41.7%	67.8%
GPT-5.4	33.3%	64.4%
Kimi-K2.5 TraceDx	37.8%	65.6%
Qwen3-32B TraceDx	33.3%	62.2%

Rare Disease. Two physicians each evaluated 30 cases randomly sampled from the Rare Disease test set (60 cases total). Physician accuracy was 31.7%, below all models in the comparison. This reflects the intrinsic difficulty of the rare disease task, as the conditions are uncommon enough that even experienced clinicians may not recognize the hallmark findings without the breadth of training data available to large language models.

MIMIC-CDM. One physician evaluated 30 cases randomly sampled from the MIMIC-CDM test set (10 per disease category). Physician accuracy was 76.7%, exceeding all models. On familiar acute abdominal presentations, clinicians apply well-practiced diagnostic heuristics that current models have not yet matched in the interactive setting.

The two datasets present opposite orderings, with physicians trailing models on rare diseases but leading on common abdominal conditions. On Rare Disease, both TraceDx variants match or exceed physician performance, with Kimi-K2.5 TraceDx leading by 6.1 percentage points. On MIMIC-CDM, TraceDx trails the physician but remains competitive with frontier models, with Kimi-K2.5 TraceDx surpassing GPT-5.4.

M Impact Statement

The proposed framework, TraceDx, advances sequential clinical diagnosis by training agents to recover the evidence that drives diagnosis. By decomposing patient records into atomic clinical findings and rewarding agents for retrieving high-criticality evidence, TraceDx provides a fine-grained training signal that aligns with how clinicians weigh evidence in practice. This contribution is particularly impactful in domains where decisions hinge on identifying decisive findings within heterogeneous, noisy patient data, and demonstrates that open-weight models can match or exceed multiple frontier baselines when supervised by an evidence-aware reward.

The primary benefit of this work is methodological. TraceDx shows that fact-level diagnostic criticality is a tractable and effective training signal for sequential clinical reasoning, enabling open-weight diagnostic agents that match or exceed multiple frontier baselines on both common and rare conditions. The fact-atomization and criticality-scoring pipeline also extends naturally to other expert domains where decisions hinge on evidence quality, such as legal reasoning, scientific investigation, or technical troubleshooting.

As with any machine learning method applied to healthcare, careful evaluation precedes any deployment. TraceDx is developed and evaluated within a clinical simulator grounded in real patient records and published case reports, providing a reproducible and validated training environment (Appendix C.3). Performance in this setting reflects diagnostic reasoning under faithful evidence, and transfer to live clinical practice would benefit from prospective validation, regulatory review, and clinician oversight following the standard pathway for medical AI. The training corpora also span a specific patient population (MIMIC-IV draws from a single US tertiary care system) and a specific disease area (abdominal pathologies in MIMIC-CDM, gastrointestinal and hepatobiliary conditions in the rare disease set), and broader clinical generalization will benefit from continued evaluation across additional sites, populations, and disease categories. We position TraceDx as a research artifact intended to advance the methodology of clinical diagnostic AI.

N Prompt Templates

N.1 Simulator Prompts

Simulator System Prompt for Rare Disease Dataset

```
We are interested in simulating a medical diagnosis pipeline.

Specifically, I am going to give you prior steps that have been taken towards diagnosis, and also the
most recent action that the doctor has taken, which might be a physical exam or medical imaging.

You will also be given access to the ground truth patient file.
This data is the result of the true medical diagnosis pipeline that occured at a real hospital.
The doctor should not have access to this data.
You only have it to help inform your simulation.
We are trying to train new doctors. And part of this is simulating results of medical exams and tests.

Your job is to look at the ground truth patient file, along with the doctor's past actions and current
action, and output what information would likely be revealed by the doctor's current action.
Please ground your simulation in the patient's ground truth data, as that document is a ground truth
record of what actually happened.
Do not mention this ground truth data in your simulation.
The doctor does not have access to this data and does not know it exists. It is a secret.

Here is an example of a good simulation:

<begin simulation example>

{simulation_example}
```

```
<end simulation example>

Here is the ground truth patient file for simulation assistance:

<begin ground truth patient file for simulation assistance>

{full_patient_file}

<end ground truth patient file for simulation assistance>

Please only use the ground truth patient file to simulate the next step.
Note that we are trying to test the doctor's ability to treat,
so you don't want to leak anything from this ground truth file.
This file was only collected at the very end of a medical case.
It is not the patient's current file or medical history.
It is a tool to help you with simulation.
If you leak information from this file, you will ruin the simulation
and doctors will not learn anything from this exercise.
Think about what information would be revealed by the doctor's action, and don't disclose any more than
 that information.

Here is the list of prior actions the doctor has taken and their results:

<begin prior actions>

{action_history}

<end prior actions>

Here is the most recent action the doctor has taken:

<begin most recent action>

{action}

<end most recent action>

Please use the patient's full medical history, along with the doctor's action, to simulate what
information will be revealed by the doctor's action.

Please output only the results of your simulation. Keep things concise.

Simulation Results:
```

Simulator System Prompt for MIMIC-CDM

```
We are interested in simulating a medical diagnosis pipeline.

Specifically, I am going to give you prior steps that have been taken towards diagnosis, and also the
most recent action that the doctor has taken, which might be a physical exam or medical imaging.

You will also be given access to the ground truth patient file.
This data is the result of the true medical diagnosis pipeline that occured at a real hospital.
The doctor should not have access to this data.
You only have it to help inform your simulation.
We are trying to train new doctors. And part of this is simulating results of medical exams and tests.

Any underlines or blanks (e.g., "___", "____") in the patient case file indicate intentionally redacted
 information.
These redactions are normal and reflect either sensitive information or details that would directly
reveal the ultimate diagnosis.
Do NOT attempt to infer, redacted information. It is acceptable for the simulation output to include
underlines or blanks (e.g., "___", "____") when the corresponding information is redacted.
Do NOT attempt to replace or resolve redacted content.

Your job is to look at the ground truth patient file, along with the doctor's past actions and current
action, and output what information would likely be revealed by the doctor's current action.
Please ground your simulation in the patient's ground truth data, as that document is a ground truth
record of what actually happened.
Do not mention this ground truth data in your simulation.
The doctor does not have access to this data and does not know it exists. It is a secret.

Here is an example of a good simulation:

<begin simulation example>

{simulation_example}
```

```
<end simulation example>

Here is the ground truth patient file for simulation assistance:

<begin ground truth patient file for simulation assistance>

{full_patient_file}

<end ground truth patient file for simulation assistance>

Please only use the ground truth patient file to simulate the next step.
Note that we are trying to test the doctor's ability to treat,
so you don't want to leak anything from this ground truth file.
This file was only collected at the very end of a medical case.
It is not the patient's current file or medical history.
It is a tool to help you with simulation.
If you leak information from this file, you will ruin the simulation
and doctors will not learn anything from this exercise.
Think about what information would be revealed by the doctor's action, and don't disclose any more than
 that information.

Here is the list of prior actions the doctor has taken and their results:

<begin prior actions>

{action_history}

<end prior actions>

Here is the most recent action the doctor has taken:

<begin most recent action>

{action}

<end most recent action>
Please use the patient's full medical history, along with the doctor's action, to simulate what
information will be revealed by the doctor's action.

Please output only the results of your simulation. Keep things concise.

Simulation Results:
```

Simulation Example Insert

```
Prior Steps Taken:
- Patient presented with 3-day history of right lower abdominal pain
- Obtained medical history showing 16-year history of Crohn's disease, currently in remission
- No NSAIDs usage reported
- Previous colonoscopy and MR enterography were unremarkable

Doctor's Action:
Physical examination of the abdomen

Simulation Results:
Examination reveals notable tenderness in the right lower quadrant of the abdomen. No rebound
tenderness or guarding is present. The rest of the abdominal examination is unremarkable. No masses are
 palpated. The tenderness is localized and does not appear to be spreading across the abdomen.
```

N.2 Fact Extraction Prompts

Rare Disease

```
You are a medical expert. Extract ONLY the atomic clinical facts from the patient case.

OUTPUT FORMAT:
- Respond with a JSON list: ["fact 1", "fact 2", ...]
- No numbering, no labels, no explanations.
- JSON list ONLY. No prose before or after.

INCLUDE:
- Objective findings obtained from diagnostic testing (e.g., imaging, endoscopy, pathology, labs).
- Objective findings from physical examination.
- Results of prior diagnostic evaluations.

EXCLUDE:
```

```
- Patient-reported symptoms or subjective complaints.
- Demographic details.
- Any interpretations, summaries, or reasoning not explicitly stated.
- Historical background information that is not an objective test or exam result.
- Any invented content or unstated inference.

RULES:
- Each item must contain exactly ONE atomic fact.
- If a sentence contains multiple facts, split them into separate items.
- Only include facts directly and explicitly stated in the case.

Your output must be a valid JSON array.

CASE:

{case}
```

MIMIC-CDM Physical Exam

```
You are a medical expert. Extract ONLY atomic clinical facts from the physical examination text below.

OUTPUT:
- A JSON list: ["fact 1", "fact 2", ...]
- JSON list ONLY. No numbering, labels, or explanations.

INCLUDE:
- Explicit physical exam findings (vital signs, auscultation, palpation, inspection findings)
- Negative or absent findings when explicitly stated
  (e.g., "no rebound tenderness", "no guarding", "no edema")

EXCLUDE:
- Demographics, history, symptoms reported by the patient
- Interpretive conclusions (e.g., "consistent with peritonitis")
- Administrative notes
- Any content not explicitly stated in the text

RULES:
- One atomic fact per item
- Use wording as close to the original as possible
- Do NOT fill in "____", keep the original text as is
- Split multiple findings into separate items
- Do NOT rephrase, normalize, or invent information

TEXT:
{case}
```

MIMIC-CDM Radiology

```
You are a medical expert. Extract ONLY atomic clinical facts from the radiology report text below.

OUTPUT:
- A JSON list: ["fact 1", "fact 2", ...]
- JSON list ONLY. No numbering, labels, or explanations.

INCLUDE:
- Explicit radiology findings (both positive and negative)
- Negative or absent findings when explicitly stated
  (e.g., "no free air", "no abscess", "appendix is normal",
  "no free gas in the pelvis", "no pericholecystic fluid")
- Observational findings that precede or accompany interpretive conclusions
  (e.g., extract "fat stranding surrounding the sigmoid colon" and
  "surrounding diverticula" even if the sentence continues with
  "suggestive of ____")
- Measurements and quantitative findings
  (e.g., "gallbladder wall thickened measuring up to 5 mm",
  "common bile duct measures 5 mm")

EXCLUDE:
- Technique descriptions and procedural boilerplate
  (e.g., "axial CT images were acquired", "grayscale and Doppler
  ultrasound images were acquired", "coronal reformats reviewed")
- Interpretive conclusions and diagnostic inferences
  (e.g., "suggestive of diverticulitis", "likely representing X",
  "consistent with cholecystitis") -- exclude the conclusion but
  keep the underlying observation
- Administrative or specimen handling notes
- Any content not explicitly stated in the text
```

```
RULES:
- One atomic fact per item
- Use wording as close to the original as possible
- Do NOT fill in "____", keep the original text as is
- Split multiple findings into separate items
- Do NOT rephrase, normalize, or invent information
- Do NOT invent facts not explicitly stated in the text

TEXT:
{case}
"""
```

N.3 Rare Disease Fact Criticality Scoring Prompts

Rare Disease Fact Criticality Scoring

```
You are a clinical evidence rater.

You will be given:
1) The GOLD diagnosis (ground-truth label).
2) The patient's INITIAL PRESENTATION (history + symptoms + key context).
3) A FACT LIST (individual findings, labs, imaging statements, exam findings).
4) NUM_TOTAL_FACTS: The total number of facts in FACT_LIST.

Your job:
Assign a criticalness score (0-3) to EACH fact for how important it is for
CONFIRMING the GOLD diagnosis, using ONLY the initial presentation + the content
of the fact itself.

IMPORTANT: This is GOLD-CONDITIONED scoring.
- We are NOT asking what diagnosis the fact suggests.
- We are asking: "If the GOLD diagnosis is true, how critical is this fact
  for establishing/confirming that diagnosis in this case?"

Scoring scale (gold-conditioned):
- 0 = negligible / background:
    * Non-specific, common, incidental, administrative, or normal findings
      that do not meaningfully help confirm the GOLD diagnosis.
    * Facts that are irrelevant or only weakly related.
    * If a fact would strongly contradict or rule out the GOLD diagnosis,
      STILL assign 0 (we do not reward contradictions).

- 1 = mildly supportive / contextual:
    * Provides some helpful context consistent with the GOLD diagnosis but is
      not specific and would not change a clinician's confidence much alone.
    * Examples: mild leukocytosis, vague abdominal tenderness, general nausea.

- 2 = strongly supportive for confirming the GOLD diagnosis:
    * A strong, diagnosis-relevant finding that substantially increases
      confidence in the GOLD diagnosis, but is not by itself nearly sufficient.
    * Typically a key imaging/lab/exam feature characteristic of the condition
      (but not a single "smoking gun" that almost forces the diagnosis).

- 3 = near-decisive for confirming the GOLD diagnosis:
    * A hallmark / defining finding that, together with the initial presentation,
      would make the GOLD diagnosis highly likely ("almost forced").
    * Examples (conceptual, not exhaustive): definitive imaging sign of the
      disease, pathognomonic lab pattern, explicit radiology impression matching
      the diagnosis, etc.

Guidelines:
- Use the initial presentation to interpret relevance (e.g., timing, location of pain).
- Do NOT reward extra details that are correct but not decision-relevant (e.g., long lists
  of normal organs on CT).
- If multiple facts repeat the same idea (duplicates), score each independently based on
  how important it is;
- When uncertain between two scores, choose the LOWER score. Be conservative.
- Keep scoring consistent across the list.
- Make sure to score all facts in the fact list.

Output format (STRICT):
Return ONLY a JSON object (dictionary) where each key is EXACTLY one fact string
from FACT LIST and each value is an integer 0-3 (the criticalness score).
The JSON object MUST contain exactly NUM_TOTAL_FACTS key-value pairs.
Each key MUST be copied VERBATIM from FACT LIST (character-for-character, including punctuation,
spacing, and units).
```

```
Do not add or omit any keys. Use the fact strings exactly as given.


Example output format :
{{
  "Fact 0 text here": 0,
  "Fact 1 text here": 1,
  "Fact 2 text here": 3,
  "Fact 3 text here": 0,
  "Fact 4 text here": 2
}}

Do not include explanations, comments, or any additional text.

GOLD DIAGNOSIS:
{gold_diagnosis}

INITIAL PATIENT PRESENTATION:
{initial_presentation}

FACT LIST:
{facts_block}

NUM_TOTAL_FACTS:
{num_total_facts}
```

N.4 MIMIC-CDM Fact Criticality Scoring Prompts

MIMIC-CDM Fact Criticality Scoring

```
You are a clinical evidence rater.

You will be given:
1) The GOLD DIAGNOSIS (ground-truth label).
2) The patient's INITIAL PRESENTATION (history + symptoms + key context).
3) DIAGNOSTIC CRITERIA (general diagnostic framework for the relevant disease categories).
4) A FACT LIST (individual findings, labs, imaging statements, exam findings).
5) NUM_TOTAL_FACTS: The total number of facts in FACT_LIST.

Your job:
Assign a criticalness score (0-3) to EACH fact for how important it is for
CONFIRMING the GOLD DIAGNOSIS, using the initial presentation and the fact itself.

IMPORTANT -- PRIORITY RULE:
- The GOLD DIAGNOSIS is the authoritative ground truth and takes precedence.
- Score facts based on how important they are for confirming what is explicitly
  stated in the GOLD DIAGNOSIS.
- The DIAGNOSTIC CRITERIA are provided as a general reference framework only.
- Do NOT restrict scoring only to items explicitly listed in the criteria.
- If a fact is clearly critical for confirming any component of the GOLD DIAGNOSIS
  (including Primary or Secondary diagnoses), it must be scored appropriately
  even if it is not explicitly emphasized in the criteria.

This is GOLD-CONDITIONED scoring:
- We are NOT asking what diagnosis the fact suggests.
- We are asking: "If the GOLD DIAGNOSIS is true, how critical is this fact
  for establishing or confirming it in this specific case?".

Scoring scale (gold-conditioned):

- 0 = negligible / background
    * Non-specific, incidental, administrative, or normal findings
      that do not meaningfully increase confidence in the GOLD DIAGNOSIS.
    * Findings unrelated or only very weakly related.
    * Absent or negative findings (e.g., "no free air", "normal appendix")
      score 0 UNLESS the absence itself is diagnostically meaningful for
      the specific GOLD DIAGNOSIS (see score=1 below).
    * If a fact directly contradicts or rules out the GOLD DIAGNOSIS,
      assign 0 (do not reward contradictions).

- 1 = mildly supportive / contextual
    * A POSITIVE finding that is relevant and consistent with the GOLD
      DIAGNOSIS, but non-specific and would not meaningfully change
      confidence alone.
    * Negative/absent findings score 1 ONLY when the absence meaningfully
      confirms the specific variant stated in the GOLD DIAGNOSIS.
      Example: "no free air" in UNCOMPLICATED diverticulitis is score=1
```

```
      because it confirms the uncomplicated nature. "No free air" in
      PERFORATED diverticulitis is score=0 because it contradicts it.

- 2 = strongly supportive
    * A positive finding that substantially increases confidence in the
      GOLD DIAGNOSIS, but is not by itself nearly sufficient.
    * Typically a key but non-defining feature -- a finding a clinician
      would specifically look for and weight heavily when working up this
      diagnosis, even if it appears in other conditions too.
    * Examples by disease:
      - Pancreatitis: epigastric pain radiating to the back; elevated WBC
        or CRP; peripancreatic fat stranding without necrosis; gallstones
        on imaging when etiology is unspecified.
      - Diverticulitis: focal LLQ tenderness at the correct site; elevated
        WBC or CRP; prior history of diverticulitis.
      - Cholecystitis: RUQ tenderness or guarding; fever with elevated WBC;
        positive Murphy sign; gallbladder wall thickening without a
        confirmatory stone.

- 3 = near-decisive / hallmark
    * A defining or characteristic finding that directly matches a core
      diagnostic feature of the GOLD DIAGNOSIS.
    * Together with the initial presentation, this would make the GOLD
      DIAGNOSIS highly likely or almost forced.
    * Examples by disease:
      - Pancreatitis: lipase or amylase ≥3× ULN; pancreatic necrosis or
        edema on CT; gallstones with dilated CBD when GT states gallstone
        pancreatitis.
      - Diverticulitis: CT showing wall thickening + fat stranding +
        diverticula at the correct segment; extraluminal air when GT states
        perforated; pericolic abscess when GT states abscess.
      - Cholecystitis: gallbladder wall thickening with impacted stone;
        pericholecystic fluid; interrupted rim sign when GT states
        gangrenous.

Guidelines:
- Use the initial presentation to interpret relevance.
- Use the diagnostic criteria as guidance, not as a strict checklist.
- Ensure that ALL components of the GOLD DIAGNOSIS are considered.
- If multiple diagnoses are listed in GOLD DIAGNOSIS, a fact should
  receive a high score if it is critical for confirming ANY of them.
- When uncertain between two scores, choose the LOWER score.
- Score all facts.

Output format (STRICT):
Return ONLY a JSON object (dictionary) where each key is EXACTLY one fact
string from FACT LIST and each value is an integer 0-3 (the criticalness
score). The JSON object MUST contain exactly NUM_TOTAL_FACTS key-value pairs.
Each key MUST be copied VERBATIM from FACT LIST (character-for-character,
including punctuation, spacing, and units). Do not add or omit any keys.
Use the fact strings exactly as given.

Example output format:
{{
  "Fact 0 text here": 0,
  "Fact 1 text here": 1,
  "Fact 2 text here": 3,
  "Fact 3 text here": 0,
  "Fact 4 text here": 2
}}

Do not include explanations, comments, or any additional text.

GOLD DIAGNOSIS:
{gold_diagnosis}

INITIAL PATIENT PRESENTATION:
{initial_presentation}

DIAGNOSTIC CRITERIA:
{diagnostic_criteria}

FACT LIST:
{facts_block}

NUM_TOTAL_FACTS:
{num_total_facts}
```

Pancreatitis Diagnostic Criteria

```
DIAGNOSTIC CRITERIA -- ACUTE PANCREATITIS
(Revised Atlanta Classification 2012)

Diagnosis requires at least 2 of the following 3 criteria:
1. Characteristic abdominal pain: epigastric, often radiating to the back.
2. Serum lipase OR amylase ≥3× the upper limit of normal (approximately
   ≥150 IU/L for both). Lipase and amylase are equally valid for satisfying
   this criterion and should be scored equivalently. Note: enzyme level
   magnitude does not predict severity -- it only confirms diagnosis.
3. Characteristic cross-sectional imaging findings: pancreatic edema or
   enlargement, peripancreatic fat stranding, fluid collections, or
   pancreatic necrosis (non-enhancing parenchyma on CT).

SEVERITY -- when the GOLD DIAGNOSIS explicitly states a severe or necrotizing
variant, the following are defining features: pancreatic necrosis on imaging,
organized peripancreatic fluid collections (pseudocyst, walled-off necrosis),
infected necrosis (gas in collection or positive culture), and persistent
organ failure (respiratory, cardiovascular, or renal).

ETIOLOGY -- when the GOLD DIAGNOSIS explicitly states an etiology, findings
that confirm that etiology are hallmark:
- Gallstone pancreatitis: gallstones or biliary sludge, dilated common bile
  duct, choledocholithiasis, and elevated liver chemistries (bilirubin, ALT,
  AST, alkaline phosphatase) at presentation.
- Alcoholic/EtOH pancreatitis: a history of significant, long-term alcohol
  use is the key etiologic fact. Binge or social drinking alone is not
  sufficient to establish this etiology.
- Post-ERCP pancreatitis: onset of characteristic pain and enzyme elevation
  within 24 hours of the procedure is the defining feature.
- Hypertriglyceridemia pancreatitis: markedly elevated triglycerides are
  the defining etiologic finding.
```

Diverticulitis Diagnostic Criteria

```
DIAGNOSTIC CRITERIA -- ACUTE DIVERTICULITIS
(WSES 2020 Guidelines and standard clinical practice)

Diagnosis is based on abdominal pain (typically lower-left quadrant), focal
tenderness, low-grade fever, and elevated inflammatory markers (WBC, CRP),
confirmed by CT showing colonic wall thickening, pericolonic fat stranding,
and diverticula at the affected segment.

ANATOMICAL SPECIFICITY -- this is the most important scoring principle:
Score findings at the anatomical site that matches the GOLD DIAGNOSIS. A
finding is only hallmark if it is at the correct location for the stated
diagnosis. For example:
- Sigmoid diverticulitis: sigmoid and descending colon findings are hallmark;
  cecal or right-sided findings are not directly confirmatory.
- Cecal diverticulitis: cecal and right colon findings are hallmark; sigmoid
  findings are not directly confirmatory.
- Small bowel, transverse colon, ampullary, or duodenal diverticulitis:
  findings at those specific segments are hallmark; standard sigmoid/colonic
  findings are not.

UNCOMPLICATED DIVERTICULITIS: wall thickening, fat stranding, and diverticula
at the correct segment are the defining CT findings. Meaningful negative
findings ("no free air," "no abscess") confirm the uncomplicated nature and
are mildly supportive (score=1) when the GOLD DIAGNOSIS states uncomplicated
or simple diverticulitis.

COMPLICATED DIVERTICULITIS -- when the GOLD DIAGNOSIS explicitly states a
complicated variant, the following are its defining features:
- Perforated diverticulitis: pneumoperitoneum or any extraluminal air,
  including tiny locules adjacent to the colon or tracking into surrounding
  structures. Even a small amount of extraluminal air is hallmark for
  perforation and microperforation variants alike.
- Abscess: a pericolic or intra-abdominal fluid collection (with or without
  rim enhancement). An inflammatory phlegmon without a drainable collection
  is also a complication marker.
- Fistula: air in the bladder in the absence of prior instrumentation,
  or an identifiable fistula tract on imaging, are hallmark for colovesical
  or colovaginal fistula variants.
- Obstruction: transition point at the affected colonic segment with proximal
  dilation is hallmark for the obstructing variant.
```

Cholecystitis Diagnostic Criteria

```
DIAGNOSTIC CRITERIA -- ACUTE CHOLECYSTITIS
(Tokyo Guidelines 2018 / TG18, adopted from TG13)

Diagnosis requires at least one local sign of inflammation, one systemic
sign of inflammation, and is confirmed by characteristic imaging findings.

Local signs of inflammation:
- Positive Murphy sign (clinical or sonographic)
- Right upper quadrant mass, pain, or tenderness

Systemic signs of inflammation:
- Fever
- Elevated CRP
- Elevated WBC

Characteristic imaging findings:
- Gallbladder wall thickening (>4 mm)
- Pericholecystic fluid
- Gallbladder distension
- Gallstones or biliary sludge, especially a stone impacted in the neck
  or cystic duct
- Sonographic Murphy sign

COMPLICATED SUBTYPES -- when the GOLD DIAGNOSIS explicitly states a
complicated variant, the following are its defining features:
- Gangrenous cholecystitis: irregular or asymmetric wall thickening,
  interrupted rim sign (loss of normal mural enhancement on contrast CT),
  intraluminal membranes or flaps, gas in the gallbladder lumen or wall,
  and pericholecystic abscess. These are hallmark findings for gangrenous
  variants.
- Emphysematous cholecystitis: gas within the gallbladder wall or lumen
  on CT is the defining finding.
- Perforated cholecystitis: a visible defect in the gallbladder wall,
  pericholecystic abscess, or bile peritonitis are hallmark.
- Empyema of the gallbladder: markedly distended gallbladder with purulent
  contents; positive bile cultures are relevant.
- Acalculous cholecystitis: typical cholecystitis findings in the absence
  of gallstones. "No gallstones" is a meaningful finding that confirms this
  variant (score=1) rather than a contradiction.
- Hydrops of the gallbladder: marked distension with an obstructing stone
  at the neck or cystic duct, without necessarily fulfilling all
  inflammatory criteria.

SEVERITY -- when the GOLD DIAGNOSIS states severe cholecystitis or organ
dysfunction, findings of organ failure (cardiovascular, respiratory, renal,
hepatic, neurological, or hematological) are hallmark for that severity grade.
Markedly elevated WBC is a key marker for moderate (Grade II) severity.

ASSOCIATED SECONDARY DIAGNOSES -- when the GOLD DIAGNOSIS lists
choledocholithiasis, cholangitis, or biliary obstruction as secondary
diagnoses, findings confirming those (CBD dilation, ductal stones, elevated
bilirubin and liver enzymes, jaundice) should be scored at their appropriate
level for confirming those secondary diagnoses.
```

N.5 Grader Prompts

MIMIC-CDM Final Diagnosis Grader

```
We are given an LLM-generated medical diagnosis and a ground truth diagnosis. Optionally,
the procedures performed on the patient at discharge may also be provided. We want to judge
whether the LLM diagnosis is clinically correct.

This is a nuanced grading task. Two diagnoses can be worded differently and still be
clinically equivalent, or they can sound related but be meaningfully distinct. The key
question to ask is: would a clinician acting on the LLM's diagnosis arrive at the same
treatment decisions as one acting on the ground truth?

=== GRADING PRINCIPLES ===

1. SPECIFICITY DIFFERENCES ARE ONLY MEANINGFUL IF THEY CHANGE TREATMENT.
   A prediction that is more or less specific than the gold is not automatically correct
   or incorrect -- what matters is whether the difference in specificity would lead a
   clinician to a meaningfully different treatment decision. If the level of detail added
   or omitted does not change management, the prediction can be considered equivalent.
   If it does change management (e.g., a missing severity marker like "perforated" or
```

```
   "gangrenous" that triggers a different intervention), the prediction is INCORRECT.

2. RELATED DISEASES ARE NOT EQUIVALENT.
   Diseases that share a mechanism or anatomical region but have distinct diagnoses and
   treatment pathways should be graded INCORRECT. Proximity on a disease spectrum does
   not make two diagnoses interchangeable.

3. EXTRA DIAGNOSES ARE ACCEPTABLE UNLESS THEY REDIRECT TO A FUNDAMENTALLY DIFFERENT
   CLINICAL PATHWAY.
   If the prediction correctly identifies the primary diagnosis and appends additional
   conditions, grade as CORRECT. The severity concern applies when the prediction
   DOWNGRADES the gold (misses a dangerous complication). It does not apply when the
   prediction ADDS a related complication on top of a correct primary -- that is an
   enrichment, not a contradiction. Only mark INCORRECT if the added diagnosis would
   redirect care so substantially that the gold's treatment plan would be abandoned or
   seriously delayed -- for example, adding a cancer diagnosis to a benign condition.
   Adding a closely related complication within the same organ system (e.g., mild
   pancreatitis alongside cholecystitis) is not sufficient grounds to fail the prediction.
   When in doubt, lean toward CORRECT if the primary is right.

4. FOR MULTI-DIAGNOSIS GOLD LABELS, THE PRIMARY DIAGNOSIS MUST BE PRESENT.
   If the gold contains multiple conditions, the prediction must cover the primary one.
   Missing a secondary condition alone is acceptable. Predicting only a secondary condition
   while missing the primary is INCORRECT.
   IMPORTANT: Do not use procedures to reclassify a secondary gold diagnosis as primary.
   The ordering of diagnoses in the gold label determines what is primary. Procedures
   reflect what was done, not necessarily which diagnosis was listed first.

5. USE PROCEDURES AS A SOFT ANCHOR FOR SEVERITY ONLY -- NOT FOR DIAGNOSIS CLASSIFICATION.
   Procedures can help confirm whether the gold diagnosis was severe enough to require
   surgery versus medical management. Use them only for this purpose. Do NOT use
   procedures to argue that a secondary diagnosis in the gold should have been predicted,
   or to fail a prediction that correctly identifies the primary. Procedure records may
   also be incomplete, so their absence does not mean no intervention occurred.

=== WORKED EXAMPLES ===

--- EXAMPLE A: CORRECT -- valid specificity refinement ---

LLM diagnosis:
"Acute sigmoid diverticulitis with small pericolic abscess."

Ground truth:
"Acute diverticulitis"

Procedures: None documented.

Thought: The gold is general. The LLM adds anatomical specificity (sigmoid) and a complication
(small abscess), but both diagnoses point to the same management pathway: IV antibiotics and
bowel rest. The prediction is a valid refinement of the gold, not a contradiction of it.

Action: 1.0

---

--- EXAMPLE B: CORRECT -- primary diagnosis covered, secondary missed ---

LLM diagnosis:
"Acute gallstone pancreatitis."

Ground truth:
"Gallstone pancreatitis
Choledocholithiasis"

Procedures: ["ERCP with sphincterotomy and stone removal", "Laparoscopic cholecystectomy"]

Thought: The gold lists two diagnoses; gallstone pancreatitis is first and therefore primary.
The LLM correctly identifies it and misses the secondary (choledocholithiasis). Per principle
4, the primary being covered is sufficient. The procedures confirm surgical management
occurred, consistent with gallstone pancreatitis. Importantly, procedures are not used to
reclassify choledocholithiasis as primary -- the gold label ordering governs that.

Action: 1.0

---

--- EXAMPLE C: CORRECT -- extra diagnosis in same clinical domain ---

LLM diagnosis:
```

```
"Acute calculous cholecystitis with mild gallstone pancreatitis."

Ground truth:
"Acute cholecystitis"

Procedures: None documented.

Thought: The gold is acute cholecystitis. The LLM adds "mild gallstone pancreatitis" on top
of the correct primary diagnosis. Both conditions arise from the same biliary pathology and
are managed within the same clinical domain -- IV fluids, analgesia, NPO, and cholecystectomy
planning. The added pancreatitis is a closely related complication, not a diagnosis that would
redirect care to a fundamentally different pathway. Per principle 3, this is acceptable.

Action: 1.0

---

--- EXAMPLE D: INCORRECT -- related but distinct disease ---

LLM diagnosis:
"Symptomatic cholelithiasis (biliary colic)."

Ground truth:
"Acute cholecystitis"

Procedures: ["Laparoscopic cholecystectomy"]

Thought: Biliary colic and cholecystitis both involve gallstones but are distinct diagnoses
with different urgency and management. Biliary colic is transient and managed expectantly;
cholecystitis involves active gallbladder inflammation requiring urgent intervention, confirmed
here by the cholecystectomy. Acting on biliary colic would delay necessary surgical care.
Per principle 2, proximity on the gallstone spectrum does not make these equivalent.

Action: 0.0

---

--- EXAMPLE E: INCORRECT -- severity downgrade that changes management ---

LLM diagnosis:
"Acute diverticulitis."

Ground truth:
"Perforated diverticulitis"

Procedures: ["Exploratory laparotomy", "Sigmoid resection"]

Thought: The gold specifies perforation, a complication that mandates emergency surgery,
as confirmed by the laparotomy and sigmoid resection. The LLM predicts uncomplicated
diverticulitis, which would be managed with antibiotics alone. Per principle 1, when
the gold specifies a severity that changes management, a general prediction is incorrect.

Action: 0.0

---

Now grade the following:

<begin LLM diagnosis>

{llm_diagnosis}

<end LLM diagnosis>

<begin ground truth diagnosis>

{ground_truth_diagnosis}

<end ground truth diagnosis>

<begin procedures performed>

{procedures}

<end procedures performed>

Use the Thought, Action paradigm. In your Thought, reason over whether the LLM diagnosis
is clinically equivalent to the ground truth by asking: would a clinician acting on this
prediction arrive at the same treatment decisions? Apply the principles above. If procedures
```

```
are provided, use them as supporting evidence for severity and treatment intent, but treat
them as potentially incomplete rather than definitive. If no procedures are listed, rely
on clinical reasoning alone.

Thought: <insert thought here>

Then output an Action: 1.0 for correct, or 0.0 for incorrect.

Action: 0.0 or 1.0

Do not output anything after the Action line.
```

Rare Disease Final Diagnosis Grader

```
We are given a medical diagnosis and a ground truth diagnosis.
We want to judge whether the diagnosis is correct.
This is actually a hard task for LLMs. Let me show you why.

The diagnosis:

"
Based on the provided information, the patient is a 47-year-old man with a history of cocaine addiction
 who was hospitalized after a cardiac arrest. He presented with post-anoxic encephalopathy,
hyperpyrexia, disseminated intravascular coagulation (DIC), and multi-organ failure. The drug screening
 showed the presence of benzoylecgonine (a cocaine metabolite), ephedrine, MDA, MDMA, delta-9-
tetrahydrocannabinol, and morphine.

The esophagogastroduodenoscopy revealed patchy areas of intense inflammation and necrosis in the
gastric fundus, likely due to vasoconstriction and ischemia. The CT scan showed gastric intramural gas
extending from the fundus along the greater curvature, which is indicative of emphysematous gastritis.

Emphysematous gastritis is a rare and severe form of gastritis characterized by the presence of gas
within the wall of the stomach, often due to ischemia, infection, or a combination of factors, and is
associated with a high mortality rate. In this patient, the combination of drug use, particularly
cocaine, which is a potent vasoconstrictor, along with ephedrine, could have led to severe
vasoconstriction and subsequent ischemia of the gastric wall, resulting in necrosis and gas formation.

Final Diagnosis: Emphysematous gastritis due to ischemia and necrosis from vasoconstriction associated
with cocaine and ephedrine use.

"

And

"
Based on these findings, a diagnosis of gastric cystic pneumatosis was made.

"

Are actually the same diagnosis. Or at least, they're close enough that physicians are satisifed.

However, consider this LLM diagnosis:

<begin LLM diagnosis>
Based on the information provided, the patient is a 34-year-old woman who presents with falls, weakness
, new-onset jaundice, mild encephalopathy, and scleral icterus. The laboratory results indicate
elevated liver enzymes, with a disproportionately higher AST compared to ALT, significantly elevated
total bilirubin, and an increased INR.

This clinical picture suggests acute liver dysfunction with possible hepatic encephalopathy. The
combination of neurological symptoms (falls, weakness, encephalopathy), jaundice, and the specific
pattern of liver enzyme elevation (AST > ALT) is highly suggestive of Wilson's disease, particularly in
 a young adult. Wilson's disease is a genetic disorder leading to copper accumulation, which can cause
liver damage and neurological symptoms.

Final Diagnosis: Wilson's disease

<end LLM diagnosis>

Meanwhile, the ground truth diagnosis is:

<begin ground truth diagnosis>

The patient was diagnosed with acute on chronic liver disease secondary to alcoholic hepatitis (AH).

<end ground truth diagnosis>

In this case, the LLM diagnosis is incorrect. And we need to mark it as 0.0 or serious harm may occur.
```

```
Grading with that sort of nuance is tricky.

Using these examples as your motivation, consider the following LLM diagnosis and ground truth
diagnosis:

<begin LLM diagnosis>

{llm_diagnosis}

<end LLM diagnosis>

<begin ground truth diagnosis>

{ground_truth_diagnosis}

<end ground truth diagnosis>

Please use the Thought, Action paradigm to grade the LLM's diagnosis against the ground truth diagnosis
.

First output a Thought where you reason over the LLM's diagnosis, and if it is close to the ground
truth. Keep your Thought to under 50 words.

Thought: <insert thought here>

Then output an Action, which either 1.0 for correct, or 0.0 for incorrect.

Action: 0.0 or 1.0

Do not output anything else for the action, since we will rely on regex parsing downstream.
```

Fact Matching Grader

```
You are a strict evaluator.

You are given:
1. ACTION_HISTORY: A chronological list of diagnostic actions and simulator results.
2. FACT_LIST: A list of target facts.

Task:
Determine which facts in FACT_LIST are supported by ACTION_HISTORY.

Definitions:
- A fact counts as "discovered" ONLY IF the simulator RESULT explicitly states it,
  or directly entails it with no outside medical knowledge.

Rules:
- Do NOT use outside medical knowledge.
- Do NOT infer unstated facts.
- Do NOT invent facts not in FACT_LIST.
- Match facts semantically (paraphrases count; trivial formatting differences like 1200 vs 1,200 count)
.
- Count each fact at most once.
- Be conservative: if unsure, do NOT count.
- Important: Follow these rules strictly so the same rubric is applied to all runs. Same inputs must
yield the same list of indices.

Indexing rules (CRITICAL):
- FACT_LIST is indexed starting from 0.
- Indices refer to the exact position of each fact in FACT_LIST as provided.
- Do NOT renumber, reorder, or skip indices.
- Output indices in strictly increasing order.
- Each index may appear at most once.

Output format (EXACTLY ONE LINE, nothing else):
discovered_fact_indices: <JSON list of 0-based indices>

Rules for output:
- The list must be valid JSON.
- Indices must be integers.
- If no facts are discovered, output an empty list: []

--------------------
EXAMPLE (illustrative only)
--------------------

ACTION_HISTORY:
```

```
[
  {
    "action": "Order comprehensive metabolic panel",
    "result": "Lipase returned at 1200 U/L. ALT is 85 U/L."
  },
  {
    "action": "Ask about alcohol use",
    "result": "Patient denies alcohol use."
  }
]

FACT_LIST:
[
  "Lipase 1,200 U/L",
  "ALT 85 U/L",
  "AST 72 U/L",
  "No alcohol use"
]

Output:
discovered_fact_indices: [0,1,3]

--------------------
NOW EVALUATE THE FOLLOWING CASE
--------------------

ACTION_HISTORY:
{action_history}

FACT_LIST:
{fact_list}
```

N.6 Baselines Diagnostic Loop Prompts

Action Prompt

```
You are given a patient's medical history.

Your job is to determine the next action that we should take
in trying to diagnose the patient as quickly as possible.

You should favor actions that will help you narrow down the diagnosis.

Please keep in mind that doctors often (but not always) do physical exams.
In the past, you've almost never done physical exams.
Sometimes imaging is done before a physical exam, but not always.

It is also customary to only do a SINGLE ACTION at a time.
For example, one physical exam, one set of medical imaging, one set of X-rays, etc.
You are not allowed to ask for multiple actions at once. If you do, you will waste your turn and get a
score of 0.0

Any underlines or blanks (e.g., "___", "____") in the medical history indicate intentionally redacted
information.
It is acceptable to base your decision on information that includes such redactions.

With all of that in mind, your job is to read the medical history below and propose an action.

<begin medical history>

{medical_history}

<end medical history>

Now please think of the next action that we should take to minimize the patient's time to diagnosis.

If you are ready to make a final diagnosis, please output
Diagnosis: <insert diagnosis here>

Otherwise, please output the next action that we should take.

Next Action: <insert action here>

Our goal is to make a final diagnosis as quickly as possible.
Your reward will be determined by how quickly you can make a diagnosis.
However, if you guess incorrectly, you get zero reward. So please balance accordingly.
A typical accuracy rate is around 60 percent for a skilled physician.
```

```
Output format rules (must follow exactly):
- If making a final diagnosis, output exactly one line:
  Diagnosis: <diagnosis>
- Otherwise, output exactly one line:
  Next Action: <action>
Do not include any other text, explanations, bullet points, or reasoning.
```

Final Disease Prompt

```
Have you ever seen ImageNet? This model is evaluated by considering Top-K
How accurate are the models K most likely image classes?

We want to consider a similar Top K scenario for the problem of medical diagnosis.

Specifically, you are given a medical report.
You need to think about the medical report, and then output the top 3 most likely
diseases or pathologies that are consistent with the data.
In addition to the top 3 most likely diseases, you also need to output 2 diseases that are consistent
with the data but less likely due to the rarity of the diagnosis.
Think about this as sampling from the tails of the distribution. The diseases you pick should still be
entirely plausible given the evidence. They should just be less likely because the underlying pathology
 is itself rare. Think Bayes rule.


Here is the medical report, which may be in progress

<begin medical report>

{medical_report}

<end medical report>

Your output should be a valid JSON string in the following format:
{{
    "most_likely_diseases": [
        {{
            "disease": "<insert disease one>",
            "rationale": "<your rationale here>"
        }},
        {{
            "disease": "<insert disease two>",
            "rationale": "<your rationale here>"
        }},
        {{
            "disease": "<insert disease three>",
            "rationale": "<your rationale here>"
        }}
    ],
    "rare_possibilities": [
        {{
            "disease": "<insert rare disease one>",
            "rationale": "<your rationale here>"
        }},
        {{
            "disease": "<insert rare disease two>",
            "rationale": "<your rationale here>"
        }}
    ]
}}

Please ensure the output is a properly formatted JSON string with no additional text before or after.
```

N.7 GRPO Prompts

Rare Disease System Prompt

```
You are a medical diagnostic agent interacting with a simulator.

Your goal is to gather sufficient diagnostic information and then make a correct diagnosis.
You must balance information gathering with timely decision-making.

At each step, output EXACTLY ONE LINE in one of the following two formats:

A: <one diagnostic action>
D: <one diagnosis label>
```

```
Rules:
- The first character of your output MUST be either "A" or "D".
- Do NOT output anything before "A:" or "D:".
- Do NOT explain your reasoning.
- Do NOT include multiple actions or diagnoses in one step.
- Do NOT ask questions directly to the patient.
- Actions must be diagnostic in nature, and must not be therapeutic or treatment-related.
- Do NOT repeat an action that has already been performed unless it is expected to yield new
information.
- When you choose "A:", you will receive new information from the simulator, the result of the action.
- When you choose "D:", you are making your final diagnosis and the episode will end.

Guidelines:
- Use "A:" to gather information when the diagnosis is uncertain.
- Use "D:" only when you are sufficiently confident in the diagnosis.
- Prefer actions that reveal decisive or high-yield clinical information.
- Avoid unnecessary or repetitive actions.

You will be evaluated on:
- Diagnostic correctness
- Coverage of clinically important facts
- Appropriateness of when you decide to diagnose
- Efficiency (number of steps taken)

Output only the action or diagnosis line. Nothing else.
```

MIMIC-CDM System Prompt

```
You are a medical diagnostic agent interacting with a clinical simulator. Your goal is to gather
diagnostic information through strategic actions and make a correct diagnosis.

## How the Interaction Works

This is a multi-turn diagnostic exercise. On each turn:
1. Analyze the information you have so far
2. Decide whether to take another diagnostic action OR make your final diagnosis
3. Output exactly ONE line in the required format (see below)

## Output Format - CRITICAL
Your output must be a single line in one of these two formats:
A: <one diagnostic action>
D: <one diagnosis label>

After you output an action, the simulator will provide results. You then analyze the updated
information and output your next action or final diagnosis. This continues until you output a diagnosis
, which ends the episode.

**Requirements:**
- Output ONLY a single line starting with "A:" or "D:"
- Do NOT add explanations, commentary, or additional text after the action/diagnosis
- Do NOT output multiple actions or diagnoses
- Do NOT output both an action and a diagnosis in the same turn

## Guidelines for Diagnostic Actions (A:)

When you output an action, it should be:
- **Diagnostic in nature**: Gather information to support diagnosis (not treatment)
- **High-yield**: Likely to provide important clinical information
- **Specific**: Be precise about what test or examination you want
- **Non-repetitive**: Don't repeat actions already performed

Do NOT output:
- Therapeutic actions (medications, treatments, procedures)
- Direct questions to the patient (use examination or test orders instead)
- Vague requests (be specific about tests/examinations)

Note: After you output an action, any underlines or blanks (e.g., "___") in the simulator's results
indicate intentionally redacted information.

## Guidelines for Making a Diagnosis (D:)

When you output a diagnosis, the episode immediately ends. Only diagnose when:
- You have gathered sufficient information to support your conclusion
- You have reasonable confidence in your diagnosis
- Additional testing is unlikely to significantly change your assessment

Balance thoroughness (gathering adequate information) with timely decision-making.
```

```
## Evaluation

You will be evaluated on:
- **Diagnostic correctness**: Accuracy of your final diagnosis
- **Information coverage**: Whether you gathered clinically important information
- **Efficiency**: Number of steps taken to reach the correct diagnosis
- **Timing**: Whether you diagnosed at an appropriate point

Output only the action or diagnosis line. Nothing else.
```

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe our method in detail in Section 4 and support each claim with experiments in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss our limitations in Appendix C.2.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We state Proposition 1 with all assumptions in Section 4.3 and provide a complete proof in Appendix J.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose all information needed for reproduction in Appendix G and Appendix F, and submit our code in the supplementary materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We open source our training pipeline, model weights, and training hyperparameters. The MIMIC-CDM data must be obtained via PhysioNet under its credentialed access requirements.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental details in Appendix G and dataset details in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars in Table 2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All training is run through the Tinker managed training API (Appendix G); we report training configuration (LoRA rank, GRPO settings, optimizer) rather than raw hardware since GPU allocation is handled by Tinker.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and have adhered to it throughout our research.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both positive and negative societal impacts of our work in Appendix M.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We discuss the safeguards for responsible research release in Appendix M.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite each dataset and pretrained model used in our work in Section 5 and respect their license and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code, model weights, and training pipeline will be shared publicly to the extent allowed by the training datasets' licensing terms.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We did not conduct crowdsourcing experiments. The physician evaluations described in Appendix C.3, Appendix D.3, Appendix E.3, and Appendix L were small-scale expert reviews involving 5 distinct physicians in total, not formal human-subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work did not involve human-subjects research. Training data are sourced from de-identified MIMIC-IV (accessed via PhysioNet credentialed access) and published case reports.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This work proposes a reinforcement learning framework for training large language models on sequential clinical diagnosis. We describe the approach in detail in Section 4 and Section 5.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.