
Learning What Matters: Criticality-Guided Reinforcement Learning for Sequential Clinical Diagnosis

Jiayi Li, Dennis Shung, Bradly Stadie

Abstract

Effective clinical diagnosis requires selectively acquiring diagnostically decisive evidence, yet training language model agents to do so remains unsolved. Existing approaches either reduce diagnosis to single-pass inference over fixed inputs, or when modeling multi-turn interaction, rely on terminal accuracy rewards that provide no signal on evidence quality and produce indiscriminate information-gathering instead. We introduce SIFT (Score-based Information-weighted Fact Training), a reinforcement learning framework that trains diagnostic agents without predefined test lists or label spaces using a trajectory-level reward grounded in the diagnostic importance of discovered evidence rather than terminal accuracy alone. A case-level logistic regression confirms that critical information ratio is the only statistically significant predictor of diagnostic success ($p < 0.001$), while evidence coverage and noise ratio are not. On MIMIC-CDM and a rare disease benchmark of gastrointestinal and hepatobiliary conditions, SIFT-trained open-weight model surpasses all frontier baselines, including GPT-5.1.

1 Introduction

A clinician rarely diagnoses from a complete picture. Evidence accumulates across questions asked, tests ordered, and results interpreted, yet most approaches to clinical diagnosis with language models reduce this iterative process to a single inference over a fixed input Singhal et al. [2023], Nori et al. [2023], McDuff et al. [2025], Brodeur et al. [2025]. This presumes that all pertinent information is available at the outset and overlooking the iterative evidence-gathering fundamental to real-world practice.. A more recent line of work lets the agent itself request information from a gatekeeper one step at a time Nori et al. [2025], Li et al. [2024], Hager et al. [2024], but primarily as a benchmark. The closest training effort, Bani-Harouni et al. [2026], trains a two-agent system on MIMIC-CDM with supervised fine-tuning and GRPO, combining terminal accuracy with test-cost penalties. Yet both the action space and the label space remain predefined and closed, and the reward provides no signal on the quality of evidence gathered along the way. This gap leaves open the question of how models can learn to acquire the findings that actually drive diagnosis. We answer this with SIFT, which trains diagnostic agents in a fully open-ended setting to recover high-value clinical evidence, using a trajectory-level reward that scores each finding by its diagnostic importance rather than relying solely on diagnostic accuracy.

Not all information contributes equally to a decision, and clinical evidence is no exception. Some findings are decisive, such as a Doppler ultrasound confirming portal vein aneurysm or lipase elevation ruling in pancreatitis, while others are supportive but non-specific, and many are diagnostically irrelevant. Effective clinicians learn to prioritize decisive findings, selectively pursuing tests that discriminate between competing hypotheses rather than exhaustively gathering all available information Elstein and Schwartz [2002]. Training an agent to do the same requires feedback on whether the actions along a trajectory collectively surfaced decisive evidence. Standard diagnostic accuracy rewards provide signal only at episode termination, telling the agent whether it reached the right

diagnosis but nothing about the quality of the evidence the trajectory gathered. In practice, this produces policies that default to indiscriminate information acquisition, collecting evidence broadly without learning to distinguish critical findings from noise.

Unlike multi-turn RL for search and tool use Jin et al. [2025], Wang et al. [2025], where only sparse terminal rewards are available, clinical diagnosis admits a richer signal: individual findings can be scored for their diagnostic value, enabling direct feedback on evidence quality rather than outcome alone. Our framework, SIFT (Score-based Information-weighted Fact Training), exploits this by extracting atomic clinical facts from each patient record and scoring their diagnostic importance offline. At the end of each training trajectory, SIFT evaluates the full set of discovered evidence, rewarding the agent based on how much diagnostically relevant information its workup recovered. The goal is to train an agent that learns not just which diagnosis to commit to, but which evidence to commit to gathering.

We evaluate SIFT on two clinical diagnostic datasets: MIMIC-CDM Hager et al. [2024], real emergency department cases derived from MIMIC-IV spanning common abdominal pathologies, and a rare disease dataset of 268 cases drawn from published medical case reports spanning a wide range of uncommon conditions. SIFT consistently improves both Qwen3-32B and Kimi-K2.5 across both datasets. Kimi-K2.5 trained with SIFT achieves 0.424 on the rare disease dataset and 0.613 on MIMIC-CDM, surpassing all frontier baselines including GPT-5 and GPT-5.1 on both datasets. Crucially, the gains are driven by evidence quality rather than evidence quantity. A case-level regression shows that recovery of highly critical evidence is the only statistically significant predictor of diagnostic success, not overall coverage. Our main contributions are:

- **Evidence quality as a learning signal.** We introduce a framework that scores atomic clinical facts offline and credits trajectories for the diagnostic value they recover, providing a trajectory-level learning signal grounded in per-fact evidence quality rather than terminal accuracy.
- **Unconstrained action and diagnosis spaces.** We show this signal enables RL training in a fully open-ended setting where the agent freely proposes any diagnostic action and commits to any diagnosis, without the predefined action or label spaces that prior training approaches rely on.
- **Quality not quantity.** Logistic regression shows that recovery of highly critical evidence is the only significant predictor of diagnostic success, not coverage or noise.
- **Strong empirical results across settings.** SIFT improves open-weight models on both MIMIC-CDM and a rare disease dataset, with a SIFT-trained Kimi-k2.5 surpassing all frontier baselines on rare diseases.

2 Related Work

Clinical diagnosis with language models. 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 Singhal et al. [2023], Nori et al. [2023], Liévin et al. [2023], Singhal et al. [2025], complex case vignettes drawn from published case reports Kanjee et al. [2023], McDuff et al. [2025], Brodeur et al. [2025], and controlled trials of model-assisted clinician reasoning Goh et al. [2024]. A more recent line of work moves toward interaction, where the model must request evidence before committing to a diagnosis. Hager et al. Hager et al. [2024] build an interactive simulator from MIMIC-IV. Building on this paradigm, Nori et al. Nori et al. [2025] and Li et al. Li et al. [2024] introduce benchmarks for stepwise evidence acquisition on published case challenges, and Schmidgall et al. ? evaluate LLM agents in simulated clinical environments across multiple specialties. Cabral et al. Cabral et al. [2024] study a related setting where case information is released in staged segments rather than requested by the agent. AMIE Tu et al. [2025] trains a conversational agent to take patient history through self-play dialogues with a simulated patient. Applying RL to clinical diagnosis, Yu et al. Yu et al. [2023] train policies for cost-effective sequential test ordering, and ED-Copilot Sun et al. [2024] demonstrates LM-guided diagnostic ordering in emergency settings. A concurrent effort, LA-CDM Bani-Harouni et al. [2026], introduces hypothesis-driven reasoning on MIMIC-CDM using SFT and GRPO. All three constrain both the action space to a predefined test menu and the diagnosis space to a fixed label set. Our work

operates with open-ended natural language for both actions and diagnoses, and rewards the clinical importance of evidence gathered along the trajectory in addition to final diagnostic accuracy.

Multi-turn RL for LLM agents. A growing body of work trains LLM agents with RL over multi-turn interaction with external environments. Search-R1 Jin et al. [2025] trains an LLM to interleave reasoning with search queries under an exact-match reward, masking retrieved tokens during the policy gradient update. RAGEN Wang et al. [2025] studies training dynamics in multi-turn agent RL and introduces stabilization techniques for symbolic and game-like environments. These methods use terminal rewards based on task success. We extend this paradigm by rewarding not only whether the final diagnosis is correct but how much clinically relevant evidence the agent recovers during the trajectory.

3 Preliminaries

Sequential Diagnostic Reasoning. Medical diagnosis can be framed as a sequential decision-making problem where a diagnostic agent π_θ must iteratively gather clinical evidence through a series of actions a_t , each yielding a clinical observation o_t , reflecting what that action reveals. The agent uses these observations to reduce diagnostic uncertainty and arrive at a final diagnosis \hat{d} . The resulting diagnostic trajectory $\tau = \{(a_0, o_0), \dots, (a_t, o_t), \hat{d}\}$ is an explainable itinerary of the diagnostic process, capturing both the clinical reasoning steps and the evidence gathered along the way.

POMDP Formulation. The diagnostic process is formally represented as a Partially Observable Markov Decision Process (POMDP). The true environment state is the complete patient medical record M , which remains hidden from the agent throughout the episode. The observation space \mathcal{O} consists of all possible clinical observations o_t returned by the simulator μ_θ in response to a_t . The action space \mathcal{A} represents all possible diagnostic actions such as ordering lab tests, imaging studies, or physical examinations. The transition and observation functions are jointly implemented by the simulator μ_θ , described in Section 4.1. Since the agent never directly observes M , it maintains the interaction history τ as a proxy for the underlying state, yielding a history-conditioned policy $\pi_\theta(a_t | \tau)$. The reward function $R(\tau)$ is computed at episode termination and is detailed in Section ???. The episode terminates when the agent outputs a final diagnosis \hat{d} or the maximum number of steps T .

The goal of training is to optimize π_θ to maximize the expected reward $\mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)]$, encouraging the agent to learn diagnostic strategies that surface clinically critical evidence efficiently and arrive at accurate diagnoses.

4 Method

We propose **SIFT** (Score-based Information-weighted Fact Training), a framework that combines atomic fact extraction to guide diagnostic exploration in the clinical simulator using criticality-aware training signals. SIFT consists of three key components: (1) a clinical simulator environment that enables multi-step diagnostic interaction, (2) an offline fact extraction and criticality scoring pipeline, and (3) a criticality-weighted reward used to optimize the policy via GRPO. We now describe each component of the framework.

4.1 Clinical Simulator Setup

Let M be the patient’s full medical file containing all history and diagnostic exams. The ground truth diagnosis is denoted $d^* \in M$. At the start of each episode, the agent receives an initial patient presentation o_0 extracted from M , which initializes the interaction history $\tau \leftarrow \{(\text{presentation}, o_0)\}$.

At each timestep t , the agent proposes diagnostic action $a_t \sim \pi_\theta(a_t | \tau)$ such as ordering a specific lab test or radiology study, with no predefined set of allowed actions. Our simulator μ_θ takes the full medical record, the gold diagnosis, the current action, and the interaction history as input and generates an observation $o_t = \mu_\theta(M, a_t, \tau)$ reflecting what the action would reveal. The action-observation pair (a_t, o_t) is then appended to τ , updating the history available for the next step. This process repeats until the agent determines it has gathered sufficient evidence to commit to a diagnosis \hat{d} , which is similarly unconstrained to any fixed label set.

For example, if at step t , the agent takes the action a_t ordering an abdomen CT, the simulator μ_θ is prompted with M , d^* , and τ to generate a consistent observation o_t that is coherent with the patient’s known history, prior actions, and ground truth diagnosis. An example of the conversation between the agent and the simulator is shown on the left panel in Figure 1.

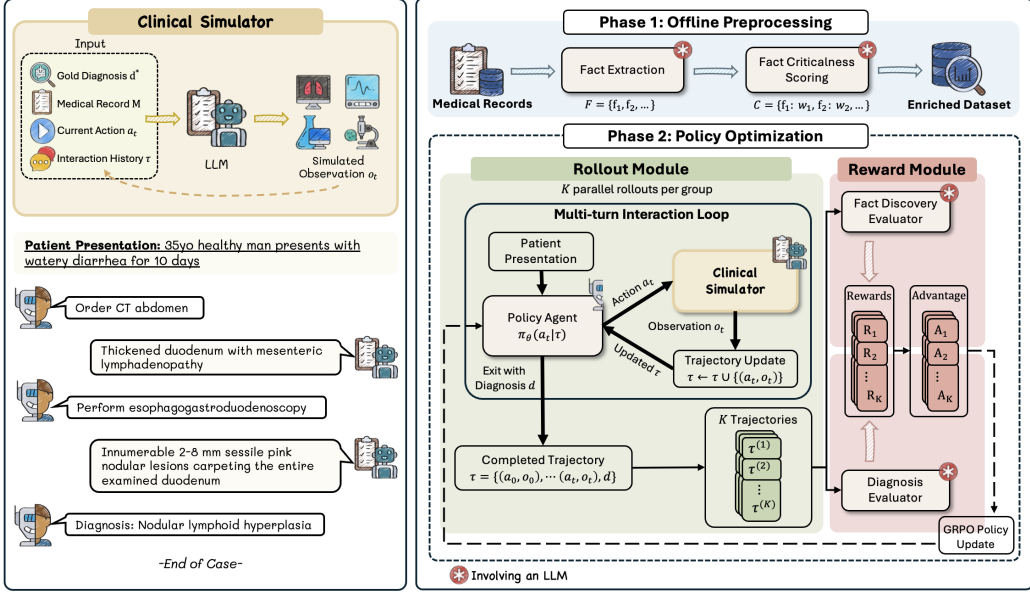


Figure 1: Overview of the SIFT framework. *Left:* The diagnostic agent interacts with the clinical simulator in a POMDP loop: at each step the agent issues a diagnostic action (e.g., ordering a lab test or imaging study), and the simulator generates a grounded observation from the patient record M . *Right:* Offline preprocessing extracts atomic clinical facts from each record and assigns a criticality score $w_f \in \{0, 1, 2, 3\}$ to each fact conditioned on the gold diagnosis d^* . During training, GRPO optimizes the agent policy using the criticality-weighted reward $R(\tau)$, which couples diagnostic accuracy with the fraction of clinically decisive evidence recovered along each trajectory.

4.2 Fact Extraction and Criticality Scoring

Offline Fact Extraction and Criticalness Scoring. 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 $F = \phi(M) = \{f_1, f_2, \dots, f_n\}$, 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. Atomicity is enforced to ensure that each fact can be independently evaluated for criticalness and matched against evidence discovered along the diagnostic trajectory.

Given the fully extracted fact set F , gold diagnosis d^* , and initial patient presentation o_0 , we define the criticality scorer ψ as an LLM-based scoring function that assigns a criticality weight to each fact conditioned on the initial patient presentation o_0 :

$$w_f = \psi(f | d^*, o_0), \quad w_f \in \{0, 1, 2, 3\}$$

where higher scores indicate greater clinical importance for establishing d^* . Concretely, a score of 0 indicates the fact is clinically irrelevant to the diagnosis, 1 indicates supportive but non-specific evidence, 2 indicates significant evidence, and 3 indicates a pathognomonic or hallmark finding that directly informs the final diagnosis. The fully scored fact list for each case is denoted as $C = \{f_1: w_{f_1}, f_2: w_{f_2}, \dots, f_n: w_{f_n}\}$. This scoring is performed offline prior to training, so the criticality scorer ψ adds no computational overhead during rollout.

4.3 Criticality-Weighted Reward

Starting from a naive formulation, one natural baseline is to optimize the policy using a binary outcome-based reward that depends only on whether the final diagnosis \hat{d} matches the gold diagnosis d^* : $R_{\text{naive}}(\tau) = \text{acc}(\hat{d}, d^*)$, where $\text{acc}(\hat{d}, d^*) = 1$ if $\hat{d} = d^*$ and 0 otherwise. However, this signal is insufficient for diagnostic reasoning, as it is only observed at episode termination and provides no feedback on what actions contributed to a correct diagnosis. In the absence of such a signal, the agent tends to greedily pursue all available facts regardless of their diagnostic relevance, leading to a preference for indiscriminate information gathering. However, not all information is equally informative. Some observations carry far greater diagnostic value than others.

To address this limitation, we define a mechanism to measure how much diagnostically relevant information is recovered along a trajectory. Given the criticality-scored fact set C and a completed trajectory τ , we define a discovered-fact extraction function $\delta(\tau, F) \subseteq F$ which maps the interaction history τ to the subset of facts in F that are identified as having been revealed during the trajectory. We denote the discovered fact set as $D(\tau) = \delta(\tau, F)$. In practice, δ is implemented using an LLM-based evaluator that parses the trajectory and outputs indices of observed facts.

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

$$CR(\tau) = \frac{\sum_{f \in D(\tau)} w_f}{\sum_{f \in F} w_f},$$

which measures the fraction of total diagnostic criticality recovered by the trajectory.

We then define the criticality-weighted reward as

$$R(\tau) = (\alpha \cdot CR(\tau) + \beta) \text{acc}(\hat{d}, d^*) + \eta \cdot CR(\tau) - \lambda \frac{t}{T}.$$

where $\alpha, \eta, \lambda \geq 0$ and $\beta > 0$. This formulation couples diagnostic accuracy with the amount of clinically critical evidence recovered along the trajectory. Trajectories that uncover more diagnostically decisive facts receive higher reward for the same correct diagnosis, while the offset $\beta > 0$ ensures that correct diagnoses remain rewarded even when criticality recall is low. The auxiliary term $\eta \cdot CR(\tau)$ provides partial credit for recovering critical evidence even when the final diagnosis is incorrect, preventing the agent from being penalized indiscriminately on trajectories that surfaced decisive findings but committed to the wrong diagnosis. The step penalty $-\lambda t/T$ discourages unnecessary actions and encourages efficient reasoning. We additionally apply a small penalty to malformed trajectories that fail to produce a valid final diagnosis.

We provide a simple justification showing that the proposed reward favors trajectories that recover more diagnostically relevant information when the criticality scorer is approximately calibrated.

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

$$w_f = ar_f + \varepsilon_f, \quad a > 0, \quad |\varepsilon_f| \leq \sigma.$$

For trajectories τ_1, τ_2 with discovered sets $D_i = D(\tau_i)$, define $\mathcal{R}(D) = \sum_{f \in D} r_f$. If

$$\mathcal{R}(D_1) - \mathcal{R}(D_2) > \frac{\sigma}{a} (|D_1| + |D_2|),$$

then $CR(\tau_1) > CR(\tau_2)$. Furthermore, if $\alpha > 0$ and both trajectories are correct with identical step count, then $R(\tau_1) > R(\tau_2)$.

Thus, up to bounded scoring error, the reward assigns strictly higher value to trajectories that recover more diagnostically relevant evidence: whenever the relevance gap between two trajectories exceeds the accumulated scoring noise, the reward ordering reflects the true clinical importance of the evidence gathered. A formal proof is provided in Appendix H.

4.4 Policy Optimization

We optimize the policy π_θ using Group Relative Policy Optimization (GRPO) Shao et al. [2024]. For each patient case, we sample K trajectories $\{\tau^{(1)}, \dots, \tau^{(K)}\}$ from the same medical record M and gold diagnosis d^* , each receiving a scalar terminal reward $R_k = R(\tau^{(k)})$.

We form a group-relative advantage by normalizing rewards within the group:

$$A_k = \frac{R_k - \text{mean}(R_1, \dots, R_K)}{\text{std}(R_1, \dots, R_K) + \varepsilon},$$

and broadcast A_k to all agent-generated tokens in $\tau^{(k)}$.

The policy is updated using a clipped importance-ratio objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{K} \sum_{k=1}^K \frac{1}{\sum_t m_{k,t}} \sum_t m_{k,t} \min(\rho_{k,t} A_k, \text{clip}(\rho_{k,t}, 1 - \epsilon, 1 + \epsilon) A_k),$$

where $\rho_{k,t} = \frac{\pi_\theta(a_{k,t}|\tau_{k,<t})}{\pi_{\text{old}}(a_{k,t}|\tau_{k,<t})}$ is the importance ratio and $m_{k,t} \in \{0, 1\}$ masks agent-generated tokens.

Simulator-generated observation tokens o_t are treated as environment outputs and masked out with $m_{k,t} = 0$, so gradients are computed only over agent-generated tokens, analogous to retrieved-token masking in Search-R1 Jin et al. [2025].

5 Experiments

5.1 Experiment Setup

Datasets. We evaluate our method on two clinical diagnostic datasets: *MIMIC-CDM* and a *Rare Disease* dataset. The MIMIC-CDM set is constructed by sampling 250 cases from the original MIMIC database Hager et al. [2024], focusing on three conditions with distinct diagnostic structures: diverticulitis, cholecystitis, and pancreatitis. To ensure reliable evaluation with an LLM-based grader, we restrict to cases with short, well-defined diagnoses. The Rare Disease dataset consists of 268 cases drawn from published medical case reports, spanning a wide range of uncommon conditions. We report results on held-out test sets of 75 MIMIC-CDM cases and 81 Rare Disease cases. Detailed dataset construction and filtering criteria are provided in Appendix B.

Baselines. We compare against closed-weight frontier LLMs (GPT-5, GPT-5.1, Sonnet 4, Sonnet 4.5, Gemini 2.5; see Appendix F.1) and the untrained base models Qwen3-32B and Kimi-k2.5 run with direct inference under the same episode setting as the RL-trained models (see Appendix G.5).

Evaluation. Models interact with the clinical simulator to collect evidence and produce a final diagnosis. Performance is measured by diagnostic accuracy using Claude Sonnet 4 as the LLM-based evaluator. Additional details on evaluation protocols are deferred to Appendix E.

5.2 Accuracy Analysis

Table 1: Diagnostic accuracy across Rare Disease and MIMIC-CDM datasets.

| Model / Method | 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 |
| Sonnet 4 | 0.319 ± 0.019 | 0.519 ± 0.026 |
| Sonnet 4.5 | 0.357 ± 0.040 | 0.595 ± 0.040 |
| Gemini 2.5 Pro | 0.386 ± 0.047 | 0.573 ± 0.038 |
| Qwen3-32B | | |
| Direct Inference | 0.349 ± 0.018 | 0.480 ± 0.040 |
| SIFT | 0.395 ± 0.021 | 0.541 ± 0.040 |
| Kimi-k2.5 | | |
| Direct Inference | 0.378 ± 0.026 | 0.568 ± 0.028 |
| SIFT | 0.424 ± 0.026 | 0.613 ± 0.026 |
| Oracle | | |
| GPT 5.1 (Oracle Context) | 0.543 ± 0.032 | 0.680 ± 0.013 |

Table 1 reports diagnostic accuracy across both datasets for frontier LLMs, direct inference baselines, and SIFT-trained models. SIFT consistently improves over direct inference for both base models and achieves the strongest non-oracle performance overall.

- **SIFT improves both models across datasets.** SIFT increases Qwen3-32B accuracy from 0.349 to 0.395 on Rare Disease (+4.6%) and from 0.480 to 0.541 on MIMIC-CDM (+6.1%). For Kimi-K2.5, accuracy improves from 0.378 to 0.424 on Rare Disease (+4.6%) and from 0.568 to 0.613 on MIMIC-CDM (+4.5%), demonstrating that criticality-driven training consistently yields gains regardless of the base model.
- **SIFT outperforms all frontier LLMs.** Kimi-K2.5 trained with SIFT achieves the best non-oracle accuracy on both datasets (0.424 on Rare Disease, 0.613 on MIMIC-CDM), surpassing all closed-weight frontier models including GPT-5 and GPT-5.1. Qwen3-32B with SIFT also exceeds Sonnet 4 and Sonnet 4.5 on Rare Disease, demonstrating that targeted evidence acquisition can compensate for differences in model scale.
- **Oracle context is an upper bound.** Providing GPT-5.1 the full patient record without any interactive workup achieves 0.543 on Rare Disease and 0.680 on MIMIC-CDM, establishing the ceiling for models that must acquire evidence interactively.

We next analyze how different reward formulations influence information gathering behavior during simulation.

5.3 Evidence Quality over Volume

Diagnostic accuracy depends on both the quality of evidence the agent gathers and the reasoning ability it applies to reach a diagnosis. To isolate the effect of reward design on evidence quality alone, we hold reasoning fixed: each policy collects evidence through the simulator, and a single strong diagnoser (GPT-5.1) produces the final diagnosis from the resulting trajectory. Under this setup, accuracy differences across policies can only arise from differences in what they choose to collect, providing a direct measure of how reward formulation shapes evidence-gathering behavior. We conduct this analysis on the Rare Disease dataset, where the diversity of conditions and the absence of a narrow set of expected tests makes evidence selection the primary challenge, since there is no obvious default workup and what the policy chooses to gather is more consequential.

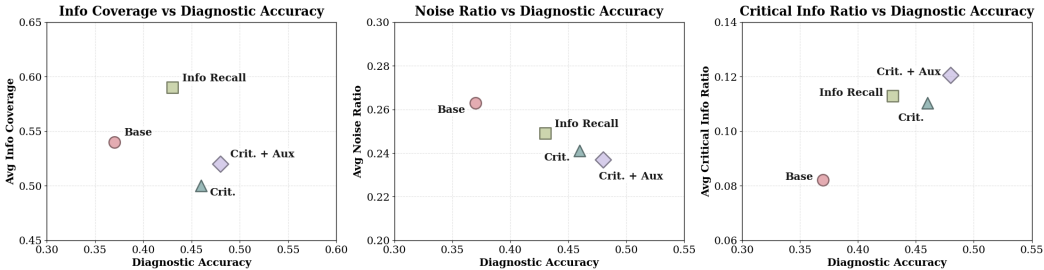


Figure 2: Diagnostic accuracy versus evidence acquisition behavior for different policies. Each point represents a model or policy: the base model is the untrained Qwen3-32B, SIFT ($\eta=0$) and SIFT ($\eta=0.2$) are criticality-weighted reward variants, and Info Recall rewards the fraction of all facts discovered regardless of criticality.

Figure 2 relates diagnostic accuracy to evidence acquisition behavior using three metrics. Let $w_f \in \{0, 1, 2, 3\}$ denote the criticalness score of a fact f :

- **Info Coverage:** proportion of total information discovered.
- **Noise Ratio:** proportion of discovered facts with $w_f = 0$.
- **Critical Info Ratio:** proportion of discovered facts with $w_f = 3$.

Evidence coverage alone does not explain performance differences. Instead, higher critical info ratio and lower noise ratio are consistently associated with higher accuracy. Policies trained with info-recall rewards (which reward the fraction of all facts discovered, regardless of their criticality) increase

overall coverage but introduce more low-value evidence, whereas SIFT recovers more diagnostically informative facts with less noise, yielding higher accuracy.

Table 2: logistic regression that tests whether evidence quality (critical information, noise, and coverage) predicts diagnostic success, while controlling for differences between models and accounting for repeated evaluations on the same cases ($N = 324$).

| Predictor | Coefficient | Std. Error | <i>p</i> -value |
|---------------------|-------------|------------|-----------------|
| Critical info ratio | 9.155 | 2.077 | < 0.001 |
| Noise ratio | -1.706 | 1.058 | 0.107 |
| Info coverage | 1.479 | 0.838 | 0.077 |
| Model fixed effects | Included | | |

To quantify these relationships, we fit a case-level logistic regression predicting diagnostic success from evidence metrics while controlling for model differences and repeated evaluations on the same case (Table 2). Recovery of highly critical evidence is the strongest and only statistically significant predictor. On the other hand, Noise ratio and evidence coverage exhibit weaker, non-significant effects. These results indicate that diagnostic success depends primarily on recovering diagnostically decisive information rather than collecting more evidence overall.

These results show that reward design improves diagnostic performance primarily by improving evidence quality.

5.4 Performance by Disease Category

To understand how diagnostic structure affects model behavior, we analyze accuracy across disease categories with different evidence requirements (Table 3) for the MIMIC-CDM data.

Table 3: 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 |

These differences provide a natural testbed for evaluating how reward design affects evidence acquisition under varying diagnostic requirements. Figure 3 summarizes accuracy by disease and model.

Cholecystitis. Qwen3-32B SIFT achieves the highest accuracy (0.74), substantially outperforming all frontier models including Gemini 2.5 (0.64). Cholecystitis diagnosis requires integrating local signs, inflammatory markers, and confirmatory imaging under the Tokyo Guidelines, making targeted evidence selection particularly valuable. Kimi-K2.5 SIFT (0.58) does not show the same advantage, suggesting the gain is specific to how criticality-guided training interacts with the Qwen base model.

Diverticulitis. Kimi-K2.5 SIFT ties GPT-5.1 for the highest accuracy (0.64), while Qwen3-32B SIFT performs slightly lower (0.59). Diverticulitis follows a relatively direct diagnostic pathway in which CT imaging provides clear confirmation of localized inflammation, and both SIFT models remain competitive with the strongest frontier models.

Pancreatitis. The two SIFT models diverge sharply. Qwen3-32B SIFT achieves the lowest accuracy of any model (0.35), while Kimi-K2.5 SIFT (0.63) matches Gemini 2.5 and trails only Sonnet 4.5 (0.66). Pancreatitis diagnosis requires satisfying at least two of three heterogeneous criteria, abdominal pain, enzyme elevation, or imaging findings, with no single decisive test. The collapse of Qwen3-32B SIFT on this disease suggests that criticality-guided training can misfire when diagnostic

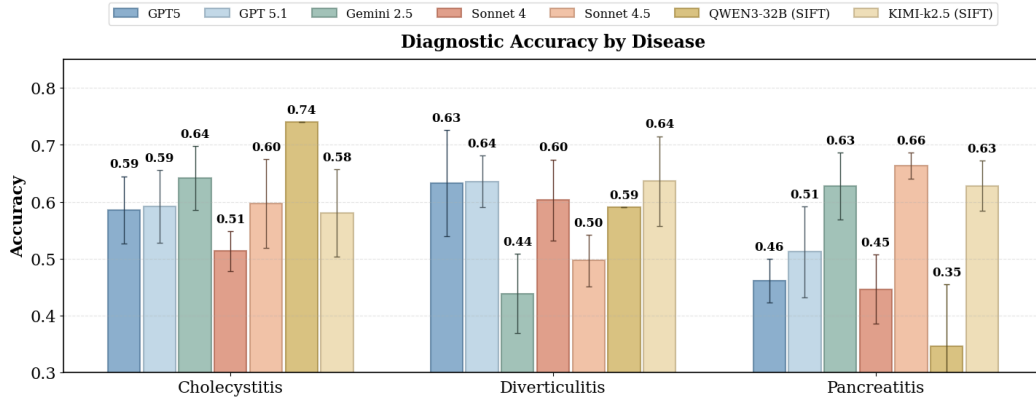


Figure 3: Mean accuracy per disease category for each model. Qwen3-32B SIFT dominates on cholecystitis, where hallmark imaging findings are decisive. Both SIFT models remain competitive with frontier baselines on diverticulitis. On pancreatitis, the two SIFT models diverge sharply: Qwen3-32B SIFT achieves the lowest accuracy across all models, while Kimi-K2.5 SIFT remains competitive.

value is distributed across conditionally important signals rather than concentrated in individually hallmark findings.

Taken together, the disease-level results indicate that the benefit of criticality-guided training depends on whether decisive evidence exists to be targeted. Qwen3-32B SIFT shows the sharpest trade-off: dominant on cholecystitis where hallmark findings exist, and weakest on pancreatitis where they do not. Kimi-K2.5 SIFT is more uniformly competitive, suggesting that base model capacity moderates how strongly the reward shapes exploration behavior.

5.5 Case Study

To understand how reward design shapes diagnostic behavior, we analyze representative cases where policies trained with criticality-based rewards differ from baseline and frontier models. These cases illustrate how learned policies improve diagnostic performance by prioritizing hypothesis-discriminating evidence, avoiding diagnostic anchoring, and obtaining definitive confirmation.

Case 1 (Portal vein aneurysm)

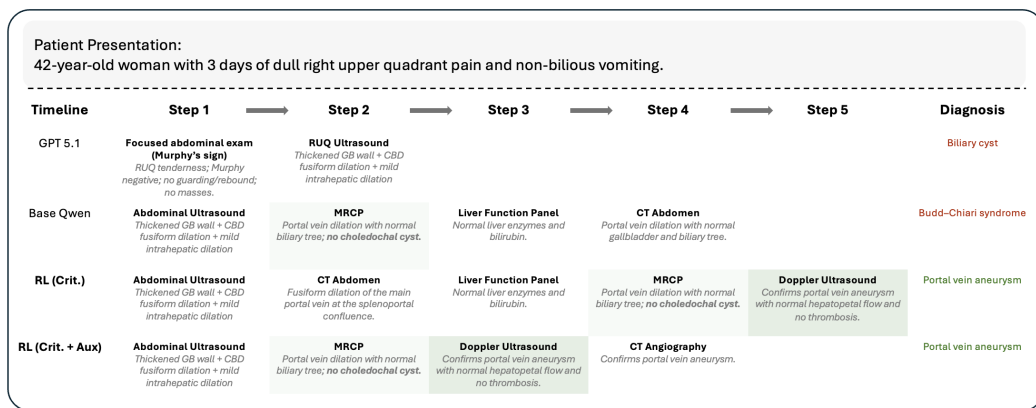


Figure 4: Case study: diagnostic trajectories for portal vein aneurysm. SIFT acquires definitive vascular evidence (Doppler ultrasound), while baseline models rely on non-specific findings and reach incorrect diagnoses.

A 42-year-old woman presented with right upper quadrant pain and non-bilious vomiting with normal laboratory findings. Initial imaging suggested biliary dilation, while subsequent studies revealed fusiform dilation of the main portal vein. Doppler ultrasound confirmed portal vein aneurysm.

In this case (Figure 4), baseline models anchor on hepatobiliary obstruction and fail to pursue vascular confirmation. GPT-5.1 interprets biliary dilation as a biliary cyst and does not obtain Doppler imaging, while the base Qwen policy attributes the finding to hepatic outflow obstruction despite normal liver function tests.

In contrast, RL policies trained with criticality-based rewards identify portal vein dilation as hypothesis-discriminating evidence and prioritize confirmatory testing. They obtain Doppler ultrasound to directly confirm aneurysmal dilation, leading to the correct diagnosis. This demonstrates that the learned policy selectively targets decisive evidence to resolve diagnostic uncertainty.

Case 2 (Anorectal malformation with rectal diverticula)

| Patient Presentation: Dysmorphic 6-week-old infant with abdominal distension, microcephaly, failure to thrive, and constipation requiring enemas. | | | | | | | |
|--|---|---|--|--|---|---|--|
| Timeline | Step 1 | Step 2 | Step 3 | Step 4 | Step 5 | Step 6 | Diagnosis |
| GPT 5.1 | Physical Exam Multiple dysmorphic features and marked abdominal distension. | Abdominal ultrasound Diffuse bowel dilatation without intra-abdominal masses | Contrast Enema Multiple diverticula in distal rectum/anorectal junction. | Pelvic MRI Confirms rectal diverticula and anorectal abnormality. | Rectal Wall Biopsy Rectal wall structural abnormality with connective tissue defects. | Whole-exome Sequencing Pathogenic variant affecting connective tissue or smooth muscle development. | FLNA connective tissue disorder / visceral myopathy |
| Gemini 2.5 pro | Physical Exam Multiple dysmorphic features and marked abdominal distension. | Abdominal X-ray Dilated small and large bowel loops with distal narrowing | Chromosomal microarray No pathogenic copy-number variant identified. | Abdominal Ultrasound Diffuse bowel dilatation without intra-abdominal masses | Whole-exome Sequencing Pathogenic variant affecting connective tissue or smooth muscle development. | | Hirschsprung disease (RET mutation) |
| RL (Info Recall) | Abdominal X-ray Dilated bowel with distal narrowing. | Anorectal manometry Functional anorectal obstruction with abnormal recto anal reflex. | Contrast Enema Multiple diverticula in distal rectum/anorectal junction. | Barium swallow Gastroesophageal reflux. | Feeding evaluation Feeding difficulty and poor weight gain due to reflux. | | Functional anorectal disorder with gastroesophageal reflux. |
| RL (Crit.) | Abdominal X-ray Dilated bowel loops with distal narrowing. | Contrast Enema Multiple diverticula in distal rectum/anorectal junction. | Pelvic MRI Confirms rectal diverticula and anorectal abnormality. | Genetic Testing No genetic syndrome identified. | Colonoscopy Direct visualization confirms distal rectal diverticula. | | Anorectal malformation with rectal diverticula |
| RL (Crit. + Aux) | Abdominal X-ray Dilated bowel with distal narrowing. | Anorectal manometry Functional anorectal obstruction with abnormal recto anal reflex. | Rectal biopsy Ganglion cells present | Contrast Enema Multiple diverticula in distal rectum/anorectal junction. | Colonoscopy Direct visualization confirms distal rectal diverticula. | | Anorectal malformation with rectal diverticula |

Figure 5: Case study: diagnostic trajectories for distal rectal diverticula. SIFT identifies disease-defining structural evidence (contrast enema and colonoscopy) and reaches the correct diagnosis, while other models either miss, ignore, or over-interpret the same findings, leading to incorrect conclusions.

A dysmorphic six-week-old infant presented with abdominal distension, microcephaly, and failure to thrive requiring enemas despite a normal rectal biopsy. Imaging revealed bowel dilation with anal narrowing and multiple diverticula arising from the distal rectum and anorectal junction, indicating a structural anorectal abnormality.

In this case (Figure 5), baseline models fail to prioritize the structural signal and instead pursue broad or unrelated investigations. GPT-5.1 anchors on a genetic or connective tissue etiology and proceeds to biopsy and whole-exome sequencing rather than confirming the structural defect. Gemini 2.5 Pro similarly prioritizes genetic testing despite inconclusive findings, while the RL information-recall policy emphasizes functional abnormalities such as reflux. In all baseline trajectories, the disease-defining structural evidence is not treated as the primary diagnostic hypothesis.

The syndromic features triggered genetic reasoning in frontier models, but the RL policy learned to prioritize the most diagnostically decisive evidence first, so it confirmed the structural abnormality before pursuing broader explanations. After detecting bowel dilation and anorectal abnormalities, the RL critical reward variants prioritize contrast imaging and colonoscopy to directly confirm distal rectal diverticula, yielding the correct diagnosis. The learned policy identifies disease-defining evidence early and prioritizes confirmatory tests, preventing diagnostic drift toward unrelated systemic or functional explanations.

Together, these cases demonstrate that criticalness-based rewards improve diagnostic accuracy by learning targeted evidence acquisition strategies that prioritize hypothesis discriminating findings and definitive confirmation.

References

- 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.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.
- Daniel McDuff, Mike Schaeckermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models. *Nature*, 642:451–457, June 2025. doi: 10.1038/s41586-025-08869-4. URL <https://doi.org/10.1038/s41586-025-08869-4>.
- Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdunour, 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.
- 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>.
- 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>.
- Paul Hager, Friederike Jungmann, and Daniel Rueckert. MIMIC-IV-Ext Clinical Decision Making: A MIMIC-IV Derived Dataset for Evaluation of Large Language Models on the Task of Clinical Decision Making for Abdominal Pathologies. *PhysioNet*, July 2024. doi: 10.13026/ztyg-ah64. URL <https://doi.org/10.13026/ztyg-ah64>. Version 1.1.
- David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Nassir Navab, and Matthias Keicher. Language agents for hypothesis-driven clinical decision making with reinforcement learning, 2026. URL <https://arxiv.org/abs/2506.13474>.
- Arthur S Elstein and Alan Schwartz. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*, 324(7339):729–732, March 2002.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Serkan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.
- Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning, 2025. URL <https://arxiv.org/abs/2504.20073>.
- Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions?, 2023. URL <https://arxiv.org/abs/2207.08143>.
- 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 Schaeckermann, 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.
- Zain Kanjee, Brian 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.
- Eric Goh, Robert Gallo, Jason Hom, et al. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969, 2024. doi: 10.1001/jamanetworkopen.2024.40969.

- Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdounour, and Adam Rodman. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Internal Medicine*, 184(5):581–583, 2024. doi: 10.1001/jamainternmed.2024.0295.
- 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.
- Zheng Yu, Yikuan Li, Joseph Kim, Kaixuan Huang, Yuan Luo, and Mengdi Wang. Deep reinforcement learning for cost-effective medical diagnosis, 2023. URL <https://arxiv.org/abs/2302.10261>.
- 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>.
- 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>.

A Appendix Overview

B Dataset Details

B.1 MIMIC-CDM Dataset

Data source. Cases are drawn from MIMIC-IV Hager et al. [2024], a large de-identified electronic health record database. We restrict to hospital admissions where the target condition is the primary (first-listed) ICD diagnosis. After this filter, 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 (mapped to human-readable labels), microbiology results, radiology reports, and discharge diagnosis (used as ground truth).

Dataset fields. Each case in the MIMIC-CDM dataset contains the following fields:

- **case_id:** hospital admission identifier from MIMIC-IV.
- **context:** the full structured clinical record, comprising five components used as the oracle queried by the simulator: patient history (free-text narrative of the 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. [2024] 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 has more than two diagnosis. 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. On the rare disease dataset, Kimi-K2.5 trained with SIFT reaches 0.424 accuracy, surpassing all frontier baselines including GPT-5 and GPT-5.1. On MIMIC-CDM, Kimi-K2.5 SIFT achieves 0.613, again outperforming all frontier models while preserving the 70/30 ratio. The final composition is shown in Table 4. Train and test splits are non-overlapping, enforced by assertion.

Reproducibility. All random operations use `random_state=42` or `seed=42`.

Table 4: 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.

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 are the primary focus of this dataset. The rarity label is preserved through all downstream processing and splits.

Splits. All 268 cases are assigned to train or test partitions with no subsampling. Splits are stratified by the COMMON_UNCOMMON label, preserving the approximately 25/75 rarity ratio in both partitions. The final split contains 187 training cases (46 COMMON, 141 UNCOMMON) and 81 test cases (20 COMMON, 61 UNCOMMON). All splits use `seed=42` and non-overlap is enforced by assertion.

B.3 Dataset Statistics

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, the gold diagnosis, 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 J.1.

C.2 Failure Case Study

A recurring failure mode occurs when the agent orders a test absent from the patient record. Rather than reporting the test as unavailable, the simulator fabricates a plausible-sounding result, and the fabricated value 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 contains a single mercury measurement: 24-hour urine mercury of 44 µg/L. Across 16 independent runs, the agent ordered serum or blood mercury, a test not present in the record. Table 5 shows a representative sample of the simulator’s responses.

Table 5: Selected simulator responses across 16 runs in which the agent ordered serum or blood mercury for the same case. The patient record documents only a 24-hour urine mercury value (44 µg/L). Fabricated 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 (<i>same number as below; different unit</i>) |
| 44 µg/L | Elevated (<i>urine value from record, relabeled as serum</i>) |
| 125 ng/mL | Markedly elevated; confirms acute exposure |
| 185 ng/mL | Markedly elevated; immediate intervention required |

A second case illustrates a different failure pattern: non-deterministic refusal. Here the simulator sometimes correctly declines to return a result, and sometimes fabricates one for the same absent test.

The case is a MIMIC-CDM admission with a gold diagnosis of gallstone pancreatitis with cholelithiasis. 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 fabricated a result in the remaining three. Table 6 shows the four runs.

Table 6: Four runs in which the agent ordered CT abdomen with contrast for the same MIMIC-CDM case. The record contains only an MRCP; no CT was performed. In one run the simulator correctly declines. In the two runs that fabricate without reporting gallstones (rows 2–3), the agent reaches an incorrect diagnosis; the run that fabricates inflammatory changes without explicitly excluding gallstones (row 4) produces a correct diagnosis.

| Behavior | Simulated CT finding |
|------------|--|
| Declined | CT not present in record |
| Fabricated | Gallbladder wall thickening, CBD dilation $\sim 7\text{--}8$ mm, PD dilation $\sim 3\text{--}4$ mm; <i>no gallstones or choledocholithiasis identified</i> |
| Fabricated | Pancreatic parenchyma with decreased enhancement, PD dilation; <i>no gallstones identified</i> |
| Fabricated | Gallbladder wall thickening with inflammation, CBD $\sim 7\text{--}8$ mm, PD $\sim 3\text{--}4$ mm, pancreatic inflammatory changes |

Together, these two cases illustrate the main failure modes when the agent orders a test absent from the record: fabricating a result with high variance (mercury case) and inconsistently declining the request (CT case). 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 fabricated 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.

C.3 Simulator Calibration

D Fact Extraction and Criticality Scoring

D.1 Atomic Fact Extraction

Facts are extracted offline before any simulation runs 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), where laboratory results, radiology reports, and physical examination findings are stored in separate modality-specific fields. At this scale, a single prompt would conflate modalities with different extraction rules, so MIMIC uses three separate strategies matched to each field. Table 7 summarizes the fact statistics for both datasets.

Table 7: 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-reported symptoms, demographic details, interpretations not explicitly stated in the record, and invented content. 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 J.2).

MIMIC-CDM. The MIMIC 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 the physical examination and radiology prompts (prompts in Appendix J.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 technique boilerplate and 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

Both datasets use the same model (GPT-5.1), the same 0–3 integer scoring scale, and the same gold-conditioned framing: score each fact for how critical it is in confirming the gold diagnosis, not for what diagnosis it suggests. The prompts for both datasets are provided in Appendix J.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 J.3.

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) but scores 0 for perforated diverticulitis (contradicts it).

Score distributions. Figure 6 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.

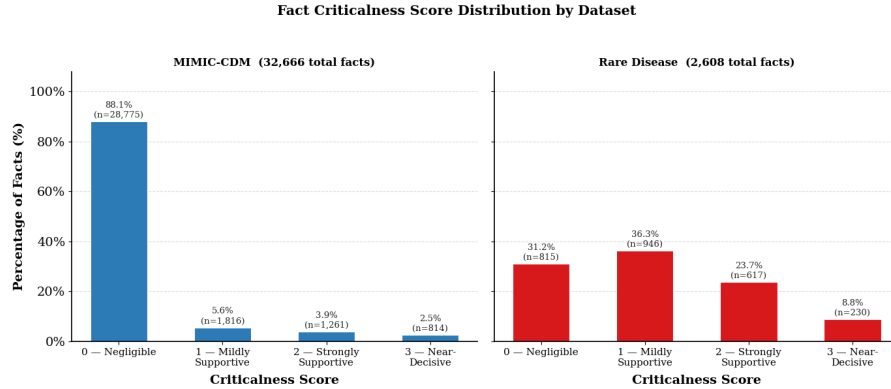


Figure 6: Criticalness score distributions for both datasets. MIMIC-CDM is heavily zero-skewed (88% score 0) because most extracted facts are routine lab values; rare disease cases have a more balanced profile (31% score 0) because case reports contain only the clinically relevant workup.

D.3 Examples

Table 8 shows the extracted facts and criticalness scores for a representative rare disease case. The case illustrates how scores span all four levels from a sequential workup: the initial presentation is minimal (abdominal pain, nausea, bilious emesis), and the diagnostic picture assembles gradually through labs, CT, and endoscopy.

Table 8: Extracted facts and gold-conditioned criticalness scores for a Bouveret syndrome case. Score 0: incidental or non-confirmatory findings. Score 1: consistent with obstruction but non-specific. Score 2: strongly suggestive features (pneumobilia, gastric distention). Score 3: hallmark findings directly confirming the cholecystoduodenal fistula and impacted gallstone.

| Patient profile: A 76-year-old man presented with 1 week of abdominal pain, nausea, and bilious emesis. Liver chemistry values were normal but lipase was elevated at 394 U/L (ref. 16–63 U/L). CT showed pneumobilia, common bile duct measuring 7 mm, decompressed gallbladder with a fistulous tract to the second portion of the duodenum, and a low-density laminated mass lesion in the third part of the duodenum with marked gastric and esophageal distention. EGD showed a massive 8-cm impacted gallstone in the third portion of the duodenum with surrounding circumferential ulceration. Despite extensive electrohydraulic lithotripsy, the stone was unable to be fragmented. | |
|--|--|
| Ground truth: Bouveret syndrome (gastric outlet obstruction via cholecystoduodenal fistula). | |
| Score | Extracted fact |
| 0 | Liver chemistry values were normal |
| 0 | Lipase level was 394 U/L |
| 0 | Common bile duct measured 7 mm on computed tomography |
| 0 | The stone was unable to be fragmented despite extensive electrohydraulic lithotripsy |
| 1 | CT showed decompressed gallbladder |
| 1 | CT showed marked esophageal distention |
| 1 | EGD showed surrounding circumferential ulceration around the gallstone |
| 2 | CT showed pneumobilia |
| 2 | CT showed marked gastric distention |
| 3 | CT showed fistulous tract from gallbladder to the second portion of the duodenum |
| 3 | CT showed a low-density laminated mass lesion in the third part of the duodenum |
| 3 | EGD showed an 8-cm impacted gallstone in the third portion of the duodenum |

D.4 Fact Criticality Scoring Calibration

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 extracted 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 J.5.

E.2 Diagnosis Grader Prompt Design

The two datasets require different levels of grader design effort, for structural reasons.

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: did the model identify 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 9 illustrates these principles with representative cases.

Table 9: 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 |
|--|---|------------|---|
| <i>Specificity & Severity</i> | | | |
| 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 | <i>[verbose paragraph concluding:]</i> acute uncomplicated sigmoid diverticulitis | 1.0 | Same diagnosis regardless of narrative framing; management unchanged |
| <i>Related Diseases</i> | | | |
| Acute cholecystitis | Symptomatic cholelithiasis (biliary colic) | 0.0 | Distinct disease; biliary colic is managed expectantly, not surgically |
| <i>Extra Diagnoses</i> | | | |
| 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 |
| <i>Multi-Diagnosis</i> | | | |
| Gallstone pancreatitis + choledocholithiasis | Acute gallstone pancreatitis | 1.0 | Primary diagnosis covered; secondary missed but acceptable |
| <i>Etiology</i> | | | |
| 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 Calibration

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: `Diagnosis: <diagnosis>` to commit to a final answer and terminate the episode, or `Next Action: <action>` to request the next clinical test. Requesting multiple simultaneous actions or producing empty output is detected and skipped without advancing the turn. The action selection prompt is provided in Appendix J.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 J.6.

Models evaluated. The policy is instantiated with each of the following models accessed via OpenRouter: GPT-5, GPT-5.1, Claude Sonnet 4, Claude Sonnet 4.5, Gemini 2.5 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.

One-Shot evaluation. 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 J.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

G.1 Models

Two base models are fine-tuned: Kimi-K2.5 with LoRA rank 32, and Qwen3-32B with LoRA rank 8 in 4-bit quantization (bitsandbytes). 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 J.7.

G.3 Reward Parameters

Table 10 lists the concrete parameter values used in the reward function $R(\tau) = (\alpha \cdot CR(\tau) + \beta) \text{acc}(\hat{d}, d^*) + \eta \cdot CR(\tau) - \lambda t/T$ across all configurations. Diagnosis correctness is graded by Claude Sonnet 4 (binary); criticality recall is computed by google/gemini-3-flash-preview.

Table 10: Reward function parameter values. $\eta = 0.2$ only for Qwen3-32B on the rare disease dataset; all other configurations use $\eta = 0.0$.

| 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 |
| η | Auxiliary criticality recall bonus | 0.0 (0.2 for Qwen3-32B on rare disease) |
| Malformed penalty | Subtracted for invalid output format | 0.3 |

G.4 Optimizer and Generation

Both models use AdamW. Kimi-K2.5 uses a constant learning rate of 5×10^{-6} ; Qwen3-32B uses a constant learning rate of 1×10^{-5} . During training, rollouts are sampled at temperature 0.95 (top- $p = 1.0$); validation uses greedy decoding. Maximum new tokens per turn is 6,144 for Kimi-K2.5 and 512 for 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 the *Direct Inference* condition in the results) is evaluated on the identical setting with the same prompts, same simulator, same maximum of 7 turns. 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 Proof of Proposition 1

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

$$w_f = ar_f + \varepsilon_f, \quad a > 0, \quad |\varepsilon_f| \leq \sigma.$$

For trajectories τ_1, τ_2 with discovered sets $D_i = D(\tau_i)$, define $\mathcal{R}(D) = \sum_{f \in D} r_f$. If

$$\mathcal{R}(D_1) - \mathcal{R}(D_2) > \frac{\sigma}{a}(|D_1| + |D_2|),$$

then $CR(\tau_1) > CR(\tau_2)$. Furthermore, if $\alpha > 0$ and both trajectories are correct with identical step count, then $R(\tau_1) > R(\tau_2)$.

Proof. Part 1: $CR(\tau_1) > CR(\tau_2)$. Expanding the weighted sums using $w_f = ar_f + \varepsilon_f$:

$$\sum_{f \in D_1} w_f - \sum_{f \in D_2} w_f = a(\mathcal{R}(D_1) - \mathcal{R}(D_2)) + \sum_{f \in D_1} \varepsilon_f - \sum_{f \in D_2} \varepsilon_f.$$

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

$$\sum_{f \in D_1} w_f - \sum_{f \in D_2} w_f \geq a(\mathcal{R}(D_1) - \mathcal{R}(D_2)) - \sigma(|D_1| + |D_2|).$$

By hypothesis $a(\mathcal{R}(D_1) - \mathcal{R}(D_2)) > \sigma(|D_1| + |D_2|)$, so the right-hand side is strictly positive. Since $CR(\tau_i) = \frac{\sum_{f \in D_i} w_f}{\sum_{f \in F} w_f}$ and the denominator is shared and positive, it follows that $CR(\tau_1) > CR(\tau_2)$.

Part 2: $R(\tau_1) > R(\tau_2)$. For correct trajectories ($\text{acc}(\hat{d}_i, d^*) = 1$) with identical step count t , the reward reduces to

$$R(\tau_i) = (\alpha + \eta)CR(\tau_i) + \beta - \lambda \frac{t}{T}.$$

Since $\alpha > 0$ and $\eta \geq 0$, we have $\alpha + \eta > 0$. Combined with $CR(\tau_1) > CR(\tau_2)$ from Part 1:

$$R(\tau_1) - R(\tau_2) = (\alpha + \eta)(CR(\tau_1) - CR(\tau_2)) > 0. \quad \square$$

I Additional Case Studies

I.1 Successful Cases

I.2 Failure Cases

J Prompt Templates

J.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, including the final diagnosis. This data is the result of the true medical diagnosis pipeline that occurred 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, including the final diagnosis. This data is the result of the true medical diagnosis pipeline that occurred 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.

J.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}
"""

J.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}

J.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 $\geq 3 \times$ 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 $\geq 3\times$ 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.

J.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
Cholelithiasis"

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 (cholelithiasis). 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 cholelithiasis 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 satisfied.

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}

J.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.

J.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.