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

Jiayi Li, Dennis Shung, Bradly Stadie

March 24, 2026

Abstract

Preprint

1 Introduction

Despite strong progress in applying language models to clinical diagnosis, most existing approaches treat diagnosis as a one-step prediction problem. In these settings, models are given a complete clinical vignette or multiple-choice question and asked to produce a diagnosis in a single pass [7, 5, 1]. While effective on benchmark tasks, this formulation assumes that all relevant information is available upfront, ignoring the process of information acquisition that is central to real-world diagnosis.

Recent work has begun to model diagnosis as a sequential interaction process [4], framing it as a stepwise encounter with gated information release. This more closely reflects how clinicians iteratively gather evidence over time. However, these formulations are primarily designed as evaluation benchmarks rather than training frameworks, and do not provide a mechanism for learning to improve diagnostic behavior through interaction. This raises the question: how can we train models to actively acquire better diagnostic information?

2 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 $\tau = \{(a_0, o_0), \ldots, (a_t, o_t), \hat{d}\}$ 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 3.1. 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 $\pi_\theta(a_t \mid \tau)$. 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.

3 Method

We propose **CRAFT** (Criticality-weighted Reasoning with Atomic Facts), a framework that combines atomic fact extraction to guide diagnostic exploration in the clinical simulator using criticality-aware training signals. CRAFT 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.

3.1 Clinical Simulator Setup

Let M be the patient's full medical file containing all history and diagnostic exams. The ground truth diagnosis is denoted $d^* \in M$. A complete episode of interaction with the simulator is depicted in Algorithm 1. At the start of each episode, the agent receives an initial patient presentation o_0 extracted from M, which initializes the interaction history $\tau \leftarrow \{(\texttt{presentation}, o_0)\}$.

At each timestep t, the agent proposes diagnostic action $a_t \sim \pi_\theta(a_t \mid \tau)$ such as ordering a specific lab test or radiology study. 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 $\hat{d}$.

Algorithm 1 Clinical Simulator μ_θ

Require: Complete ground truth medical record M, simulator μ_θ, policy π_θ, maximum steps T

Initialize: $\tau \leftarrow \emptyset$, $t \leftarrow 0$
$o_0 \leftarrow \textsc{ExtractPresentation}(M)$
$\tau \leftarrow \tau \cup \{(\texttt{presentation}, o_0)\}$

while $t < T$ **do**
  $a_t \sim \pi_\theta(a_t \mid \tau)$
  if a_t is a diagnosis **then**
    $\tau \leftarrow \tau \cup \{\hat{d}\}$
    return τ
  end if
  $o_t = \mu_\theta(M, a_t, \tau)$
  $\tau \leftarrow \tau \cup \{(a_t, o_t)\}$
  $t \leftarrow t + 1$
end while

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.

3.2 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 $F = \phi(M) = \{f_1, f_2, \ldots, f_n\}$, where each f_i is an atomic clinical finding, a single observation such as a lab value, imaging finding, or physical examination result, stated without interpretation or inference. 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:

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

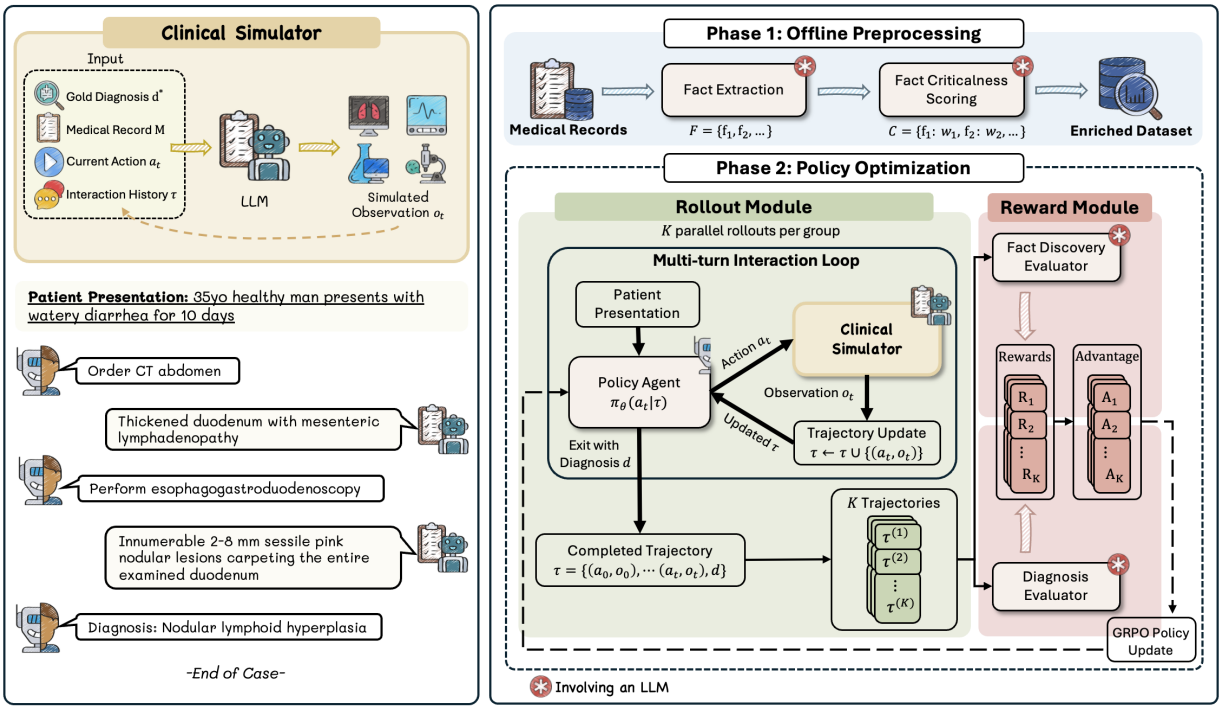

Figure 1

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 $C = \{f_1 : w_{f_1}, f_2 : w_{f_2}, \ldots, f_n : w_{f_n}\}$ This scoring is performed offline prior to training, so the criticality scorer ψ adds no computational overhead during rollout.

3.3 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

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

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 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 (Order preservation under approximate calibration)**.** *Let each fact $f \in F$ have an unknown latent diagnostic relevance $r_f \geq 0$, and suppose the criticality scorer satisfies*

$$w_f = a r_f + \varepsilon_f, \qquad a > 0, \quad |\varepsilon_f| \leq \delta.$$

For any discovered fact sets $D_1, D_2 \subseteq F$, define

$$\mathcal{R}(D) = \sum_{f \in D} r_f.$$

If

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

then

$$\sum_{f \in D_1} w_f > \sum_{f \in D_2} w_f,$$

and consequently

$$CR(D_1) > CR(D_2).$$

This result shows that, up to bounded scoring error, criticality recall preserves the ordering of trajectories by the amount of diagnostically relevant evidence they recover. In particular, trajectories that uncover more clinically informative facts achieve higher criticality recall.

The following corollary shows how this property translates to the reward function.

Corollary 1 (Reward preference at fixed accuracy)**.** *Consider two trajectories τ_1, τ_2 for the same case with identical diagnosis accuracy and identical step count. If $CR(\tau_1) > CR(\tau_2)$, then*

$$R(\tau_1) > R(\tau_2).$$

Thus, among trajectories with the same diagnostic outcome, the proposed reward assigns higher value to those that recover more clinically relevant evidence, encouraging targeted information acquisition rather than indiscriminate exploration.

3.4 Policy Optimization

We optimize the policy π_θ using Group Relative Policy Optimization (GRPO) [6]. For each patient case, we sample K trajectories $\{\tau^{(1)}, \ldots, \tau^{(K)}\}$ 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:

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

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 [3].

4 Experiments

4.1 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 [2], 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 diagnostically challenging cases where decisive evidence is sparse and must be actively identified. 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 frontier LLMs (GPT-5, GPT-5.1, Sonnet 4, Sonnet 4.5, Gemini 2.5) and a base open-weight model (Qwen3-32B). We additionally evaluate hybrid pipelines that combine learned evidence acquisition with a strong diagnosis model. Full implementation details for each baseline are provided in Appendix F.

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

4.2 Accuracy Analysis

Category	Model	Rare Disease	MIMIC-CDM
Frontier LLMs	GPT 5	$\underline{0.419 \pm 0.013}$	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
Trained Models	Qwen3-32B (Base)	0.283	0.480 ± 0.040
	RL (Crit.)	0.370	0.541 ± 0.040
Hybrid Pipelines	Hybrid (Crit.+ Aux)	**0.481**	**0.653**
	GPT 5.1 (Oracle Context)	0.543 ± 0.032	—

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

Table 1 reports diagnostic accuracy across Rare Disease and MIMIC-CDM for frontier LLMs, our RL-trained information policy (Qwen3-32B), and hybrid pipelines that combine learned evidence acquisition with a strong diagnosis model. Overall, we observe consistent improvements from RL training and further gains from hybrid pipelines across both datasets.

- **RL improves the base policy across datasets.** Applying RL increases Qwen3-32B accuracy from 0.283 to 0.370 on Rare Disease and from 0.480 to 0.541 on MIMIC-CDM, demonstrating that criticality-driven training consistently improves diagnostic performance across both settings.
- **Stronger gains on rare and harder cases.** The improvement is more pronounced on Rare Disease (+8.7% absolute) compared to MIMIC-CDM (+6.1%), suggesting that criticality-aware training is particularly beneficial in settings where diagnostically decisive information is sparse and must be actively prioritized.
- **RL policy is competitive with frontier models.** Despite using a smaller base model, the RL-trained Qwen3-32B approaches or exceeds several frontier LLMs (Sonnet 4/4.5) on

both datasets, indicating that improved information acquisition can partially compensate for model scale.

- **Hybrid pipelines achieve the best performance.** Combining the learned policy with a strong diagnosis model yields the highest accuracy on both datasets (0.481 on Rare Disease, 0.653 on MIMIC-CDM), outperforming all standalone frontier baselines. This demonstrates that the learned policy retrieves higher-quality, diagnosis-relevant information that significantly improves downstream reasoning.
- **GT context is an upper bound.** Providing GPT-5.1 the full ground-truth context achieves 0.543, serving as a reference for policies that must acquire evidence interactively.

These results suggest that the primary gains arise from improved evidence acquisition rather than purely stronger reasoning. We next analyze how different reward formulations influence information gathering behavior during simulation.

4.3 Effect of Reward Design on Evidence Acquisition

To examine the performance of hybrid pipelines in relation to the RL policy, we analyze how different RL training objectives influence evidence collection during simulation for the rare disease dataset. Because models must actively gather information and decide when it's ready to produce a diagnosis, the quality of collected evidence directly affects downstream diagnosis. To isolate evidence quality from reasoning ability, we evaluate diagnostic accuracy using a fixed hybrid pipeline in which GPT-5.1 performs the final diagnosis based on evidence collected by each policy. This allows us to measure the causal impact of reward design on evidence quality.

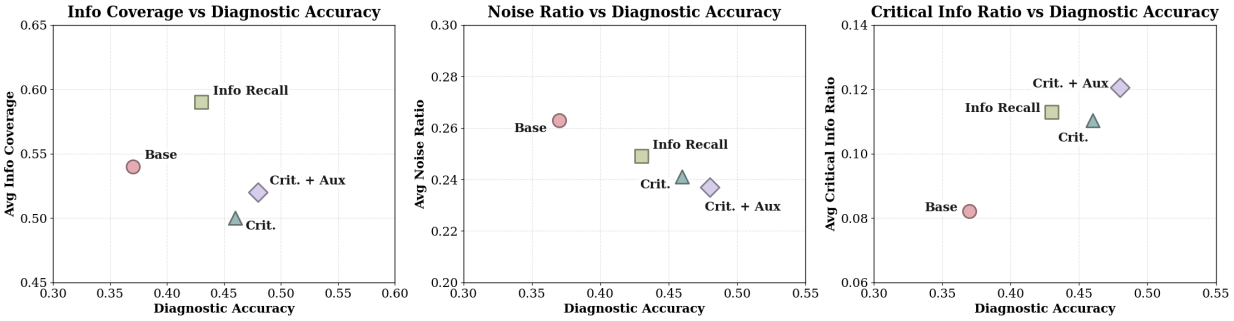

Figure 2: Diagnostic accuracy versus evidence acquisition behavior for different policies. Each point represents a model or policy: the base model is the untrained Qwen3-32B, and other points correspond to RL-trained policies labeled by their reward formulation.

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. Policy trained with information-recall rewards increase overall coverage but introduce more low-value evidence, whereas criticalness-based rewards (Crit. and Crit.+Aux) recover more diagnostically informative facts with less noise, yielding higher accuracy. The auxiliary objective achieves the best tradeoff between recovering critical evidence and limiting noise.

Predictor	**Coefficient**	**Std. Error**	p**-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	

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 ($N = 324$).

To quantify these relationships, we fit a case-level logistic regression predicting diagnostic success from evidence metrics while controling 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.

4.4 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.

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. RL (Crit.) achieves the highest accuracy (0.74), substantially outperforming all frontier models, with Gemini 2.5 next at 0.64. Diagnosis requires integrating symptoms, inflammatory markers, and confirmatory imaging, making performance sensitive to selecting hypothesis-discriminating evidence. Criticalness-based training appears suitable to identifying these decisive signals.

Diverticulitis. GPT 5.1 achieves the highest accuracy (0.64), with GPT 5 close behind (0.63).

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

Table 3: Comparison of diagnostic structures across disease categories.

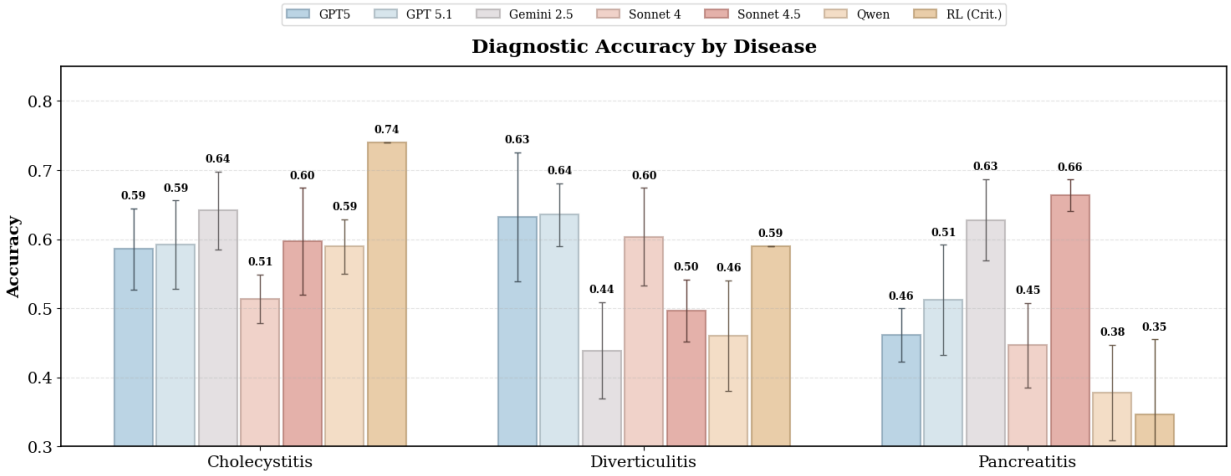

Figure 3: Mean accuracy for each model on cholecystitis, diverticulitis, and pancreatitis cases. Performance varies across diseases, with RL (Crit.) improving performance on cholecystitis while frontier models achieve strong results on diverticulitis.

RL (Crit.) performs slightly lower (0.59), but still provides a clear improvement over the base Qwen model (0.46). iagnosis follows a relatively direct pathway in which CT imaging provides clear confirmation of localized inflammation, and the improvement over the base model shows that the criticalness reward appears to effectively guide acquisition of this confirmatory evidence.

Pancreatitis. RL (Crit.) performs worst across all models (0.346), while Sonnet 4.5 (0.663) and Gemini 2.5 (0.628) achieve the highest accuracy. Unlike the other two diseases, pancreatitis diagnosis requires satisfying at least two of three heterogeneous criteria—symptoms, enzyme elevation, or imaging—with no single decisive test. The criticalness reward, which incentivizes targeting individually critical evidence, appears poorly suited to this multi-signal threshold structure, leading to a marked degradation in performance relative to frontier models, and even underperforming the base Qwen model (0.378).

The results suggest that criticalness-based rewards provide the largest gains when diagnosis depends on targeted confirmatory evidence, as in cholecystitis and diverticulitis. When no single finding is decisive and diagnosis emerges from the conjunction of heterogeneous signals, RL policy

performance deteriorates.

4.5 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)

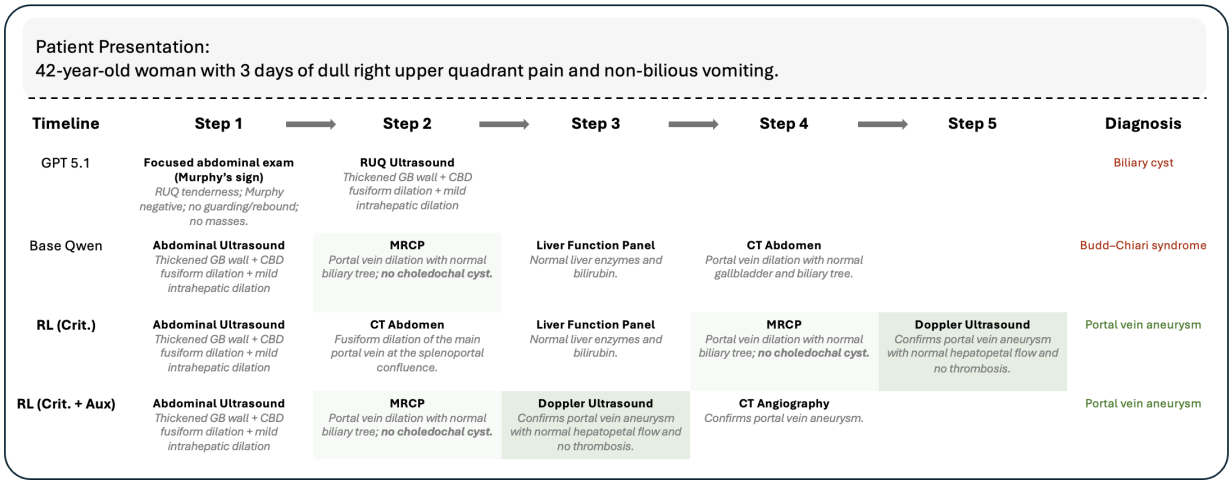

Figure 4: Case study: diagnostic trajectories for portal vein aneurysm. RL (Crit.) and RL (Crit.+ Aux) acquire definitive vascular evidence (Doppler ultrasound), while baseline models rely on non-specific findings and reach incorrect diagnoses.

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)

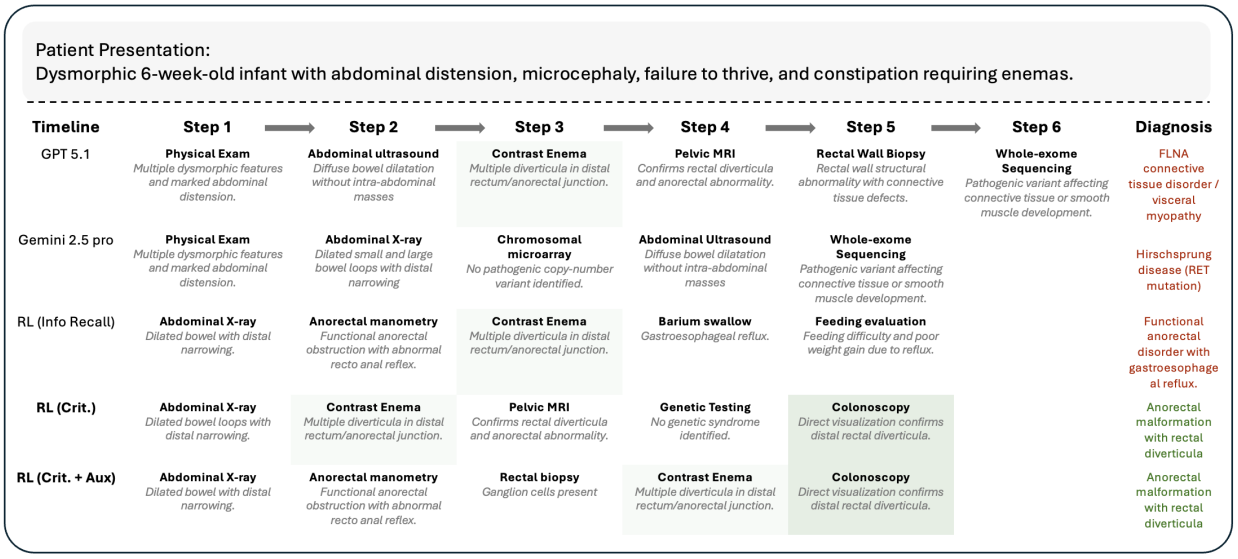

Figure 5: Case study: diagnostic trajectories for distal rectal diverticula. RL (Crit.) and RL (Crit.+ Aux) identify disease-defining structural evidence (contrast enema and colonoscopy) and reach the correct diagnosis, while other models either miss, ignore, or over-interpret the same findings, leading 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.

Case 3 (Sigmoid Diverticulitis)

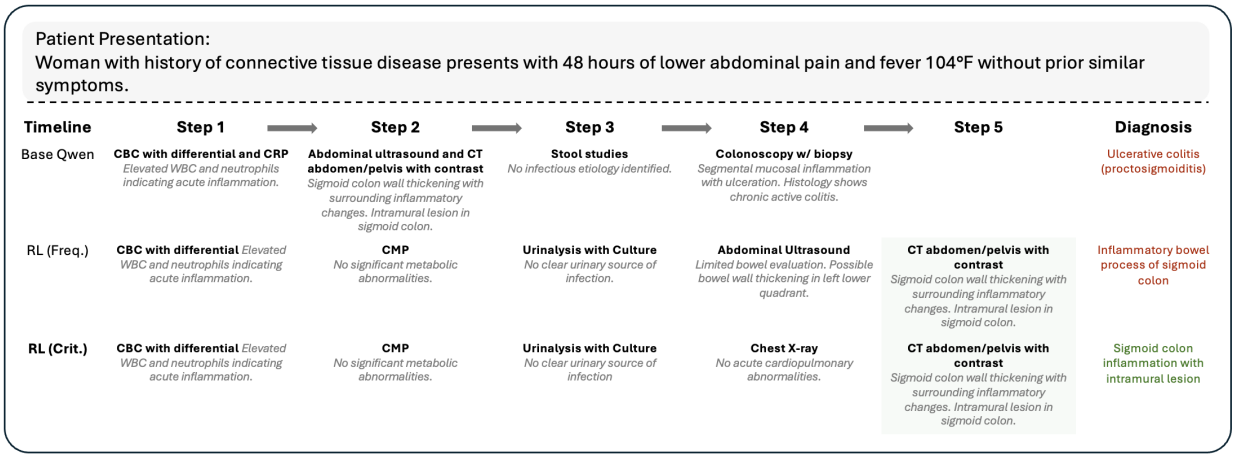

Figure 6: Case study: diagnostic trajectories for sigmoid diverticulitis. RL policies identify localized sigmoid inflammation on CT imaging, while the baseline model overinterprets colonoscopy findings and incorrectly diagnoses ulcerative colitis, highlighting differences in evidence interpretation.

A patient presented with acute lower abdominal pain and fever with laboratory evidence of inflammation. CT imaging revealed sigmoid colon wall thickening with surrounding inflammatory changes, creating diagnostic uncertainty between diverticulitis and inflammatory bowel disease.

Across models, diagnostic differences arise from how imaging evidence of sigmoid inflammation is interpreted. Although all models obtain CT imaging, they differ in how decisively the findings are interpreted. The base model allows additional downstream testing to dilute the diagnostic signal, while the RL policy trained with criticalness reward treats localized sigmoid inflammation as hypothesis-discriminating and commits accordingly. Although both RL variants obtain CT imaging, they differ in how the findings are interpreted. The criticalness-based policy treats localized sigmoid inflammation as disease-defining for diverticulitis, whereas the other policy treats the same evidence as non-specific and selects an incorrect diagnosis.

Case 4 (Acute Pancreatitis)

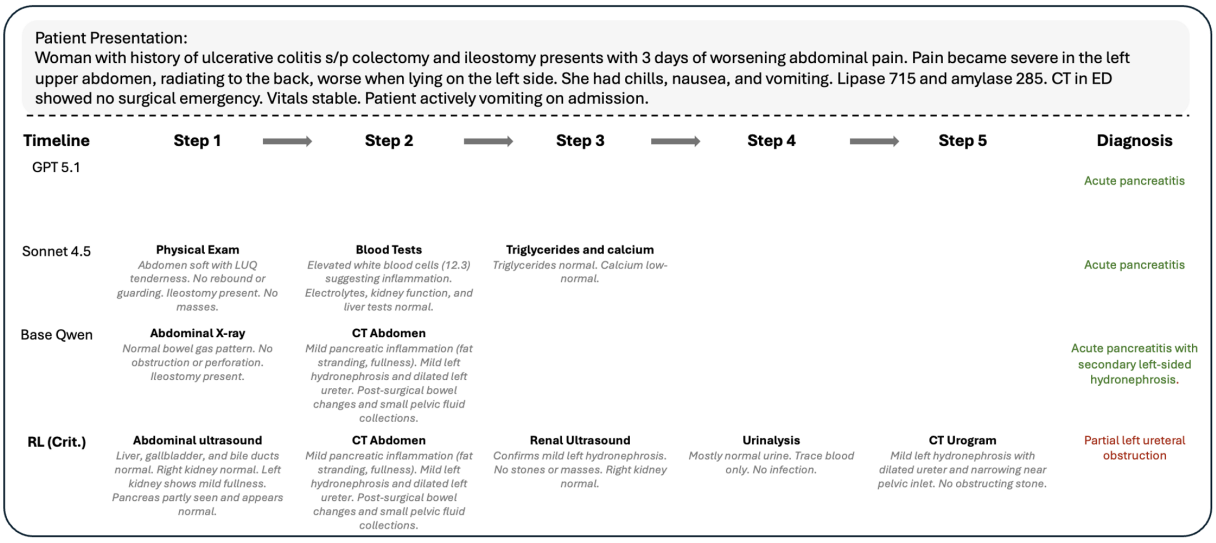

Figure 7: Case study: diagnostic trajectories for appendicitis. The RL (Crit.) policy over-focuses on secondary findings and pursues extended urologic evaluation, leading to an incorrect diagnosis of ureteral obstruction, while frontier and base models correctly identify acute pancreatitis from minimal diagnostic evidence.

A patient presented with severe abdominal pain radiating to the back, elevated lipase and amylase, and stable vitals

Frontier models correctly identify acute pancreatitis using minimal but sufficient evidence. The base Qwen model also reaches the correct diagnosis while noting incidental findings.

In contrast, the RL (Crit.) policy deviates by over-prioritizing secondary findings. As shown in Figure X, the policy follows a trajectory of imaging and urologic evaluation (renal ultrasound, CT urogram), focusing on mild hydronephrosis and ultimately diagnosing partial ureteral obstruction instead of pancreatitis.

This failure highlights a limitation of the criticality-based reward: when multiple moderate-signal findings are present, the policy may over-focus on individually salient but non-decisive evidence, rather than satisfying the minimal diagnostic criteria. In such settings, diagnosis depends on meeting a threshold of heterogeneous signals rather than identifying a single critical feature, which is not well captured by the current reward formulation.

References

[1] Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdulnour, and Adam Rodman. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Internal Medicine*, 184(5):581–583, 2024.

[2] 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. Version 1.1.

[3] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[4] 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.

[5] 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.

[6] 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.

[7] 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.