THE ROAD NOT ORDERED: COUNTERFACTUAL STEP PRICING FOR REINFORCEMENT LEARNING IN SEQUENTIAL MEDICAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical diagnosis is a sequential and cost-aware process: physicians order examinations, update their diagnostic reasoning as evidence accumulates, and make a final diagnosis once the evidence is sufficient. Reinforcement learning provides a natural framework for training sequential diagnostic agents. However, reinforcement learning in this setting faces two key challenges: reward blindness, where diagnostic paths with different levels of efficiency can receive the same score if they reach the same conclusion; and credit dilution, where one final score is shared across many steps, making the pivotal examination difficult to distinguish from routine or unnecessary ones. These challenges arise from assigning a single final-outcome reward to the entire trajectory. To address these issues, we propose a counterfactual reinforcement learning approach called Counterfactual Step Pricing (CoSteP). CoSteP focuses on decision-critical states where the policy is uncertain. At each selected state, it compares the chosen action with plausible counterfactual alternatives using short rollouts and a utility that considers both diagnostic accuracy and examination cost. A persistent rollout cache reuses compatible trajectories accumulated across training batches, reducing computation without expert step annotations or a learned critic. Experiments on MIMIC-IV, ClinicalBench, and a private hospital dataset show that CoSteP improves diagnostic accuracy while reducing both the number and cost of examinations.

1 INTRODUCTION

Large language models (LLMs) have made substantial progress in advancing medical AI, achieving strong performance across diverse clinical tasks (Peng et al., 2026; Yan & Pei, 2022). However, existing LLMs are typically trained on structured medical corpora containing complete patient records (Chen et al., 2026; Lacerda et al., 2025). As a result, these models are mainly designed to reason and make diagnoses from fully available clinical information, which frames medical diagnosis as a static prediction problem. In contrast, real-world diagnosis is an interactive, long-term decision-making process under uncertainty, which can be formulated as a sequential decision problem. Clinicians must reason from incomplete evidence, decide which examination to request next, and determine whether the available evidence is sufficient for a diagnosis (Pauker & Kassirer, 1980). They must balance two key goals: reaching the correct diagnosis and doing so efficiently under examination-cost constraints. These requirements make reinforcement learning a natural framework for sequential diagnosis (Qiu et al., 2025; Kazemnejad et al., 2024). By interacting with the diagnostic environment, an agent can learn which examination to request, when to make a diagnosis, and how to balance diagnostic accuracy with examination cost.

While promising, reinforcement learning for sequential diagnosis faces two key challenges. The first is **reward blindness**. Specifically, many existing reinforcement learning methods rely on outcome-only rewards, which evaluate only whether the final diagnosis is correct (Yu et al., 2021; Qiu et al., 2025). Consequently, this single reward cannot distinguish trajectories that share the same endpoint but differ completely along the way. As shown in Figure 1(a), both trajectories produce the correct diagnosis. However, one uses a small number of targeted examinations, whereas the other includes redundant examinations before reaching the same conclusion. Both trajectories receive the same reward, and the policy receives no clear signal to prefer a shorter and less costly diagnostic trajectory.

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

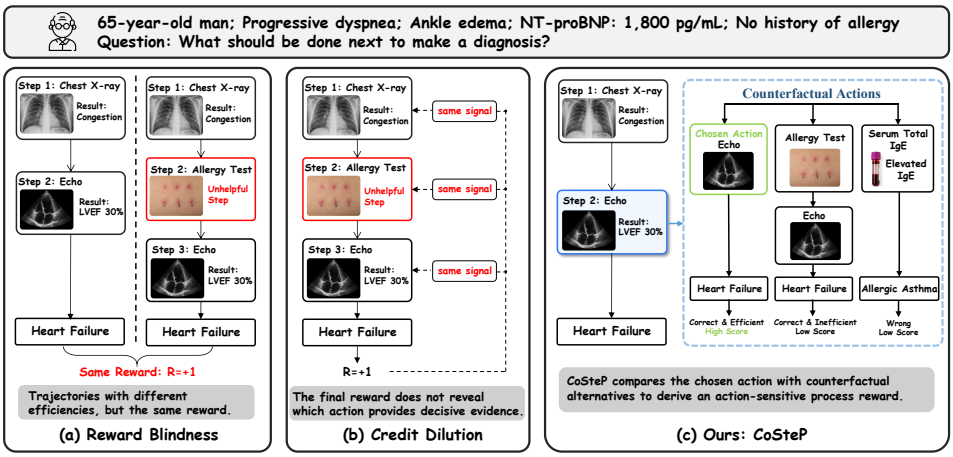


Figure 1: Illustration of reward blindness and credit dilution in outcome-only reinforcement learning. CoSteP mitigates these limitations by comparing the chosen action with counterfactual alternatives to derive action-sensitive process rewards for trajectory-level optimization.

The second challenge is **credit dilution**. In a long diagnostic trajectory, a pivotal action may be surrounded by routine or unhelpful actions. However, a final outcome reward evaluates the trajectory as a whole and provides no direct information about the relative diagnostic value of its individual actions (Feng et al., 2026; Kazemnejad et al., 2024). As illustrated in Figure 1(b), the heart test provides decisive evidence whereas the allergy test is unhelpful, yet an outcome-only score does not distinguish their contributions.

We argue that these challenges arise because an *outcome-only, trajectory-level* reward does not account for examination efficiency or the relative diagnostic value of intermediate decisions. This approach may be effective in short-horizon tasks (Lightman et al., 2024), where a correct answer often follows a clear solution path and credit is easier to assign. However, in clinical diagnosis, the same correct conclusion can be reached through many different examination paths. This makes it difficult to obtain process-level supervision from medical experts, as exhaustively labeling all possible trajectories would require substantial clinical effort. A direct alternative is to estimate action values through Monte Carlo rollouts, but performing rollouts at every state is computationally expensive and often wasteful. Most states are routine, and their next actions are largely determined by the evidence already collected. Therefore, an effective solution should focus evaluation on the most informative decision states while limiting the cost of rollout-based estimation.

To address these issues, we propose a reinforcement learning framework called **CoSteP** (**Co**unterfactual **Ste**p **P**ricing). CoSteP is inspired by counterfactual reasoning employed in real clinical judgment, which distinguishes a good decision from a bad one not by its outcome in isolation but by comparing it against the alternatives that were plausibly available at the moment of choice. CoSteP first identifies decision-critical states where the policy is uncertain about the next action. It measures this uncertainty from the policy's action distribution and performs counterfactual evaluation only at these states. This focuses computation on decisions that are likely to affect the diagnostic trajectory, rather than evaluating every step uniformly. More importantly, CoSteP constructs action-sensitive reward signals through counterfactual comparison. At each selected state, it compares the chosen action with the top-ranked alternatives generated by the policy using counterfactual rollouts and a cost-aware utility. The chosen action receives a positive process reward only when it produces a more accurate or efficient diagnostic outcome than the alternatives. These action-sensitive process rewards are aggregated into the trajectory score, making trajectory-level optimization more responsive to the relative quality of selected intermediate decisions and thereby alleviating credit dilution. As shown in Figure 1(c), the heart test receives the highest reward because it outperforms the competing examinations at that state. Furthermore, CoSteP maintains a persistent rollout cache that accumulates compatible trajectories across training batches. These cached trajectories provide counterfactual samples with little additional computation, reducing the cost of action-value estimation. Together with selective state evaluation and counterfactual reward esti-

mation, this design requires neither expert step annotations nor a learned critic. It enriches sparse outcome rewards with counterfactual signals derived from intermediate decisions that are likely to influence the final diagnosis.

2 Related Work

Most medical language models perform static diagnosis from complete case information, without actively deciding which evidence to acquire (Peng et al., 2026; Jin et al., 2024; Peng et al., 2024). Recent work has increasingly modeled diagnosis as a sequential decision process. In this setting, evidence is collected step by step, reflecting the dynamic nature of real clinical practice. Multi-agent simulators approximate clinical practice by constructing clinical environments and patient simulators, where agents interact with patients and make diagnostic decisions (Schmidgall et al., 2024; Almansoori et al., 2025; Tang et al., 2024; Kim et al., 2024; Peng et al., 2025; Nori et al., 2025; He et al., 2026; Zheng et al., 2024). Although these methods model diagnosis as a sequential decision process, the agents themselves are typically driven by prompting and rely heavily on backbone capability rather than learned diagnostic policies. More recent works train diagnostic policies with reinforcement learning. For instance, Qiu *et al.* train an agent in a virtual clinical environment with agentic reinforcement learning (Qiu et al., 2025). These methods rely on sparse outcome rewards and therefore suffer from severe credit-assignment difficulty along long diagnostic trajectories. To handle this challenge, recent studies have explored step-level or process-level rewards that assign credit to intermediate actions. For example, ReasonRAG (Zhang et al., 2026), StepSearch (Wang et al., 2025b), and GiGPO (Feng et al., 2026) rely on densely sampling continuations or alternative actions from intermediate states to estimate step-level values without extra annotation. Although these methods provide denser supervision than outcome-only rewards, they may suffer from limited scalability as performing such rollouts at every state is expensive.

3 Methodology

3.1 Task Formulation

Most existing medical diagnosis methods formulate diagnosis as a static task, where the model receives all available examination results of a patient and produces a diagnosis in a single pass. In real clinical practice, evidence is instead acquired progressively. We therefore formulate diagnosis as a sequential decision process, as illustrated in Figure 2(a). At step t, the diagnostic state is

$$s_t = \left(x_0, \{(a_i, o_i)\}_{i=0}^{t-1}\right), \tag{1}$$

where x_0 contains the patient's initial clinical information, such as symptoms, medical history, and available vital signs. In implementation, the diagnostic state s_t is serialized as a prompt that concatenates x_0 with the chronological interaction history. Each a_i records the model's reasoning and structured decision at step i, while o_i records the corresponding environment response. Given the state prompt s_t, the policy generates the next model output:

$$a_t \sim \pi_\theta(\cdot \mid s_t). \tag{2}$$

The model is required to produce a reasoning process and one of two structured actions: `Request_Examination` or `Diagnosis`. `Request_Examination` action queries the environment for the corresponding result and appends both the action and result to the prompt to form s_{t+1}. A `Diagnosis` action terminates the diagnostic process with the final prediction $\hat{d}$. The resulting trajectory is denoted by $\tau = (s_0, a_0, \ldots, s_T, a_T)$, where the terminal action a_T contains the final diagnosis $\hat{d}$. Unlike static diagnosis, sequential diagnosis requires the policy to determine both which examination to request and when the accumulated evidence is sufficient to issue a diagnosis. Given the ground-truth diagnosis $d^\star$, the terminal outcome reward is

$$R_{\text{out}}(\tau) = \mathbf{1}(\hat{d} = d^\star). \tag{3}$$

A desirable diagnostic trajectory should not only reach the correct diagnosis, but should also avoid unnecessary, expensive, or unavailable examinations. We therefore define its cost-aware utility as

$$\begin{aligned} U(\tau) = R_{\text{out}}(\tau) &- \lambda_{\text{test}} N_{\text{tests}}(\tau) \\ &- \lambda_{\text{cost}} C_{\text{total}}(\tau) - \lambda_{\text{na}} N_{\text{na}}(\tau), \end{aligned} \tag{4}$$

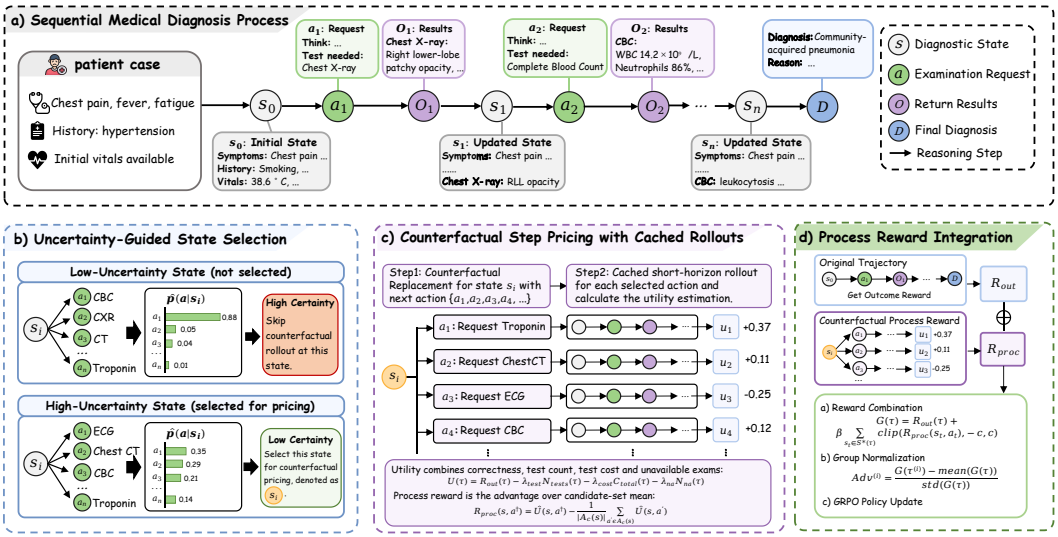


Figure 2: The overview of the framework. (a) We reformulate medical diagnosis as a sequential decision process. (b) For each trajectory, we identify decision-critical states based on action uncertainty. (c) On the selected states, we estimate a counterfactual process reward by rolling out alternative actions. (d) The process rewards are aggregated with the outcome reward to form the trajectory score for GRPO optimization.

where $N_{\text{tests}}(\tau)$ is the number of examination requests in trajectory τ, $C_{\text{total}}(\tau)$ is the sum of examination costs, and $N_{\text{na}}(\tau)$ is the number of requested examinations that are unavailable in the current case. The non-negative coefficients λ_{test}, λ_{cost}, and λ_{na} control the penalties for examination quantity, cost, and unavailable requests, respectively.

3.2 Core Principle: Counterfactual Step Pricing

When a trajectory reaches the correct diagnosis, another action available at the same state may have reached the same conclusion with fewer examinations or a lower monetary cost. CoSteP therefore evaluates an action relative to other plausible actions. We define the ideal full-horizon cost-aware value of the state–action pair (s, a) as

$$Q_U^{\pi}(s, a) = \mathbb{E}_{\tau_{s,a} \sim \pi_\theta(\cdot|s,a)} \left[U(\tau_{s,a}) \right]. \tag{5}$$

s denotes the current state. The action a is the decision taken by the model at state s. The continuation $\tau_{s,a}$ starts from state s, executes action a, and follows policy π_θ until a final diagnosis is produced. The expectation $\mathbb{E}[\cdot]$ is taken over full downstream continuations induced by the current policy. In practice, CoSteP uses a computationally tractable bounded-horizon surrogate of this ideal quantity. To evaluate an action counterfactually, we keep the diagnostic state s and its preceding interaction history fixed, replace the action taken in the original trajectory with different candidate actions, and generate a downstream continuation for each candidate under the same policy. Let $\mathcal{A}_c(s)$ denote a finite candidate set containing the canonical action $a^\dagger$ taken in the original trajectory and other distinct, policy-generated actions that are plausible under the same state. Each $a' \in \mathcal{A}_c(s)$ denotes a candidate action used as a counterfactual alternative to the original action. The ideal counterfactual step price of $a^\dagger$ is defined as

$$P^{\pi}(s, a^\dagger) = Q_U^{\pi}(s, a^\dagger) - \frac{1}{|\mathcal{A}_c(s)|} \sum_{a' \in \mathcal{A}_c(s)} Q_U^{\pi}(s, a'). \tag{6}$$

The first term measures the expected utility obtained by selecting the original action $a^\dagger$, while the second term is the average expected utility of the candidate actions considered under the same diagnostic state. A positive price indicates that $a^\dagger$ performs above the candidate-set average, whereas

a negative price indicates that more effective alternatives are available. Here, pricing denotes the relative utility assigned to an individual diagnostic step.

Computing Eq. (6) exactly is impractical for two reasons. First, not every intermediate state provides useful supervision. Counterfactual evaluation is more informative at high-uncertainty states (Kazemnejad et al., 2024; Cao et al., 2026). Second, estimating $Q_U^\pi(s, a)$ with full-horizon rollouts for every candidate action is computationally expensive. CoSteP therefore evaluates only high-uncertainty states and constructs a practical utility surrogate using cached rollouts.

3.3 Practical CoSteP

3.3.1 Decision-Critical State Identification

We identify decision-critical states from the uncertainty of the current policy's next-action distribution, as illustrated in Figure 2(b). For each intermediate state s_t in a sampled trajectory, we independently sample n_s next actions from $\pi_\theta(\cdot \mid s_t)$.

Because the policy generates free-form text, we instruct the model to follow a predefined output template. We then use regular-expression rules to extract the structured action from each sampled output and normalize semantically equivalent actions into the same canonical form, denoted by $\bar{a}^{(j)}$. The empirical probability of an action a is defined by its normalized frequency among the n_s samples: $\hat{p}(a \mid s_t) = \frac{1}{n_s} \sum_{j=1}^{n_s} \mathbf{1}(\bar{a}^{(j)} = a)$.

We quantify the local action uncertainty using empirical entropy:

$$\mathcal{H}(s_t) = - \sum_{a \in \tilde{\mathcal{A}}(s_t)} \hat{p}(a \mid s_t) \log \hat{p}(a \mid s_t), \tag{7}$$

where $\tilde{\mathcal{A}}(s_t)$ denotes the set of distinct canonical actions extracted from the samples. A low value of $\mathcal{H}(s_t)$ indicates a stable next decision, whereas a high value indicates hesitation among multiple examinations or between requesting further evidence and issuing a diagnosis.

Given an uncertainty threshold η, we select

$$\mathcal{S}^\star(\tau) = \{s_t \in \tau : \mathcal{H}(s_t) \geq \eta\}. \tag{8}$$

Counterfactual pricing is performed only for states in $\mathcal{S}^\star(\tau)$.

3.3.2 Counterfactual Step Pricing with Cached Rollouts

For each selected state, CoSteP constructs policy-relative counterfactual actions, estimates their downstream utilities, and computes the step price of the original action, as illustrated in Figure 2(c). For a selected state s, we rank the canonical actions in descending order of their empirical probabilities $\hat{p}(a \mid s)$. The top-M actions, together with the action $a^\dagger$ taken in the original trajectory, form the counterfactual candidate set:

$$\mathcal{A}_c(s) = \text{Top-}M\,(\hat{p}(\cdot \mid s)) \cup \{a^\dagger\}, \tag{9}$$

where including $a^\dagger$ ensures that the original action is evaluated against the alternatives most frequently considered by the current policy under the same diagnostic evidence.

For each candidate action $a \in \mathcal{A}_c(s)$, we force the policy to execute a at state s and continue the diagnostic process. The resulting downstream trajectory is denoted by $\tau_{s,a}$, and its k-th sample by $\tau_{s,a}^{(k)}$. We obtain K continuations for each state–action pair (s, a) and use them to estimate a bounded-horizon surrogate of its downstream utility. To reduce redundant generation, we introduce a rollout cache. We align trajectories of the same patient case that share the same initial presentation and examination history but take different next actions. For each action a, the downstream suffixes following that action are stored as cached continuations $\mathcal{T}_{s,a}^{\text{cache}}$. When constructing the K continuations for (s, a), CoSteP first reuses available cached continuations and generates additional rollouts only when fewer than K samples are available. Each continuation is truncated after at most H decision steps. It either reaches a final diagnosis within this horizon or is required to produce a diagnosis using the evidence collected so far. This bounded horizon trades exact full-horizon evaluation for tractable process-reward estimation. Each continuation is evaluated using the trajectory

utility in Eq. (4). We define the resulting bounded-horizon utility estimate as

$$\hat{U}(s,a) = \frac{1}{K}\sum_{k=1}^{K} U\left(\tau_{s,a}^{(k)}\right). \tag{10}$$

The counterfactual process reward of the action $a^\dagger$ taken in the original trajectory is defined as its advantage over the average utility of the candidate set:

$$R_{\text{proc}}(s,a^\dagger) = \hat{U}(s,a^\dagger) - \frac{1}{|\mathcal{A}_c(s)|}\sum_{a'\in\mathcal{A}_c(s)} \hat{U}(s,a'). \tag{11}$$

This construction provides a practical bounded-horizon approximation to the relative utility comparison in Eq. (6). The reward is positive when the selected action yields higher utility than its policy-relative alternatives.

3.3.3 Reward Integration and Policy Optimization

We optimize π_θ using Group Relative Policy Optimization (GRPO). The terminal outcome reward preserves final diagnostic correctness as the global objective, while counterfactual process rewards inject action-sensitive information from selected decision-critical states into the trajectory-level optimization signal, as illustrated in Figure 2(d).

We combine the terminal outcome reward with the clipped process rewards at selected states to form the trajectory score:

$$G(\tau) = R_{\text{out}}(\tau) + \beta \sum_{s_t\in\mathcal{S}^\star(\tau)} \text{clip}\left(R_{\text{proc}}(s_t,a_t), -c, c\right), \tag{12}$$

where $\beta \geq 0$ controls the contribution of the process reward, and $c > 0$ denotes the clipping threshold. The clipping function is defined as $\text{clip}(x,-c,c) = \min\{c, \max\{-c,x\}\}$, which constrains each process reward to the interval $[-c,c]$ and prevents unusually large counterfactual estimates from destabilizing policy optimization. For each patient case, GRPO samples a group of n trajectories and normalizes their trajectory scores within the group to obtain relative advantages. These advantages are then used to update π_θ with the standard GRPO objective.

4 Experiments

4.1 Experimental Setup

Datasets. We train on 23,377 **MIMIC-IV** cases (Qiu et al., 2025) and evaluate on 750 held-out MIMIC-IV cases, 1,500 **ClinicalBench** (Yan et al., 2026) cases, and 952 **Private** cases from a partner hospital. ClinicalBench and the Private Dataset are used only for out-of-domain evaluation.

Metrics. We report following metrics: (1) **Diagnostic Accuracy (Acc.)** is the percentage of cases whose final diagnosis matches the ground truth. (2) **Average Number of Examinations (AEN)** is the average number of examinations issued per case and reflects diagnostic efficiency. (3) **Average Exam Cost (AEC)** is the average total monetary cost of examinations per case and reflects economic efficiency. Higher Acc. and lower AEN/AEC are better. (4) **Expert rating of intermediate decisions**: three senior physicians independently score randomly sampled intermediate decision nodes on a 1–5 Likert scale along five dimensions, which are described in the appendix.

Implementation Details. We use Qwen3-4B-Instruct as backbone and GPT-5.4 as the judge. Examination actions are normalized and matched to the case-specific examination set, with unrecognized actions treated as unavailable. Full implementation details are provided in the appendix.

4.2 Main Results

Table 1 compares diagnostic accuracy and examination efficiency on the in-domain MIMIC-IV test set and two out-of-domain benchmarks, ClinicalBench and Private. We report Diagnostic Accuracy (Acc., %), Average Number of Examinations (AEN), and Average Exam Cost (AEC, USD). The main findings are as follows.

Table 1: Main results on the MIMIC-IV test set and two out-of-domain benchmarks (ClinicalBench and Private). We report Diagnostic Accuracy (Acc., %), Average Number of Examinations (AEN), and Average Exam Cost (AEC, USD). All results are averaged over three independent runs with different random seeds. Among all non-closed-source methods, the best results are shown in **bold** and the second-best results are underlined. Closed-source LLMs are reported in gray for reference.

Method	MIMIC-IV (In-domain)			ClinicalBench (OOD)			Private (OOD)			Macro Average		
	Acc.	AEN	AEC	Acc.	AEN	AEC	Acc.	AEN	AEC	Acc.	AEN	AEC
Closed-Source LLMs												
GPT-5.4	42.93	2.20	62.80	56.00	1.59	18.46	36.97	1.17	67.39	45.30	1.65	49.55
Claude-Sonnet-4.6	38.93	4.65	84.13	54.27	4.88	44.05	32.98	2.61	74.57	42.06	4.05	67.58
Gemini-3.1-Flash	34.93	2.37	54.86	51.67	1.90	15.13	32.98	1.78	55.21	39.86	2.02	41.73
DeepSeek-V4	40.00	3.76	71.90	54.00	2.69	25.85	29.94	2.66	68.23	41.31	3.04	55.33
Open-Source LLMs												
Qwen3-4B (Yang et al., 2025)	32.40	**1.38**	39.80	44.87	**1.13**	20.32	25.95	**1.08**	35.34	34.41	**1.20**	31.82
Qwen3-8B (Yang et al., 2025)	32.53	2.84	75.29	47.80	1.97	28.10	32.88	1.99	55.52	37.74	2.27	52.97
Qwen3-32B (Yang et al., 2025)	37.73	3.12	97.81	55.60	1.96	32.02	33.40	2.01	60.20	42.24	2.36	63.34
Ministral-3-8B (Liu et al., 2026)	32.40	2.47	64.59	46.80	1.66	18.19	19.64	1.90	36.17	32.95	2.01	39.65
Med-Gemma-27B (Sellergren et al., 2025)	35.60	3.35	106.41	54.80	1.80	28.59	32.25	1.92	59.68	40.88	2.36	64.89
Huatuo-o1-8B (Chen et al., 2024)	12.40	8.43	108.80	28.53	7.58	74.95	16.28	7.40	82.26	19.07	7.80	88.67
Baichuan-M1-32B (Wang et al., 2025a)	19.20	8.49	81.96	36.40	6.20	28.30	16.50	6.30	58.20	24.03	7.00	56.15
Medical Agents												
MedAgent (Tang et al., 2024)	33.60	2.40	70.22	49.40	1.70	22.19	30.78	2.02	60.50	37.93	2.04	50.97
MDAgent (Kim et al., 2024)	34.53	2.65	76.88	49.93	1.93	24.30	30.36	2.29	65.92	38.27	2.29	55.70
ToR (Peng et al., 2025)	38.27	2.95	85.34	50.73	2.10	28.13	30.25	2.51	72.43	39.75	2.52	61.97
Reinforcement Variants												
GRPO-4B	47.33	2.55	52.30	51.67	3.65	34.10	29.41	1.50	32.80	42.80	2.57	39.73
GRPO-4B + $U(\tau)$	48.18	2.17	43.59	52.32	3.12	31.55	29.89	1.51	33.01	43.46	2.27	36.05
GRPO-4B + Non-CF Step Reward	49.26	2.15	41.72	53.47	3.08	28.94	29.95	1.42	27.86	44.23	2.22	32.84
DiagAgent-8B (Qiu et al., 2025)	37.47	4.42	231.46	41.07	4.04	231.99	32.04	3.75	318.07	36.86	4.07	260.51
DiagAgent-14B (Qiu et al., 2025)	41.33	6.16	191.81	50.73	5.52	192.11	**34.24**	5.32	263.17	42.10	5.67	215.70
CoSteP-4B (Ours)	**53.07**	2.10	**36.65**	**59.13**	3.02	**17.33**	30.06	1.24	**16.58**	**47.42**	2.12	**23.52**

(1) CoSteP-4B achieves the highest average accuracy and the lowest average examination cost. Across the three datasets, CoSteP-4B obtains an average Acc of 47.42%, outperforming GPT-5.4 (45.30%), Qwen3-32B (42.24%), Med-Gemma-27B (40.88%), and DiagAgent-14B (42.10%). It also achieves the lowest average AEC of 23.52 USD. On the Private set, CoSteP-4B is less accurate than DiagAgent-14B (30.06% vs. 34.24%), but reduces AEC from 263.17 to 16.58 USD.

(2) Counterfactual step pricing provides gains beyond ordinary cost-aware reward shaping. Outcome-only GRPO achieves 42.80% average Acc with an AEC of 39.73 USD. Directly optimizing GRPO with the same cost-aware utility $U(\tau)$ improves these to 43.46% and 36.05 USD. The non-counterfactual variant uses the same step-wise utility estimator without subtracting the candidate-set counterfactual mean, which reaches 44.23% and 32.84 USD. CoSteP achieves 47.42% Acc and 23.52 USD AEC, showing that its gains cannot be explained solely by additional cost information or non-counterfactual process shaping.

(3) Explicit policy training outperforms the evaluated multi-agent baselines. MedAgent, MDAgent, and ToR achieve average Acc of 37.93%, 38.27%, and 39.75%, with AEC between 50.97 and 61.97 USD. CoSteP-4B achieves higher accuracy and lower cost, supporting explicit diagnostic-policy optimization for sequential examination ordering.

(4) Static medical capability does not directly transfer to active examination ordering. Huatuo-o1-8B reaches 19.07% average Acc, well below Qwen3-8B (37.74%), while Med-Gemma-27B (40.88%) is similar to Qwen3-32B (42.24%). This suggests that medical-domain pre-training alone does not ensure effective sequential decision-making under examination and cost constraints.

4.3 Ablation Studies

Table 2 evaluates the contribution of each component in CoSteP. Removing the individual utility penalties has a smaller effect on accuracy but substantially changes the examination profile. Without λ_{test}, AEN increases from 2.10 to 4.10, showing its role in limiting

Table 2: Ablation studies on the components of CoSteP with Qwen3-4B as the backbone on the MIMIC-IV test set.

Method	Acc. (%)	AEN	AEC (USD)
CoSteP (Ours)	**53.07**	**2.10**	**36.65**
w/o λ_{test}	51.85	4.10	78.40
w/o λ_{cost}	52.10	2.20	79.20
w/o λ_{na}	50.62	4.85	62.50

redundant examinations. Without λ_{cost}, AEC rises from 36.65 to 79.20 USD while AEN changes only slightly, indicating that this term specifically discourages expensive examinations. Removing λ_{na} increases AEN to 4.85 and AEC to 62.50 USD, demonstrating its importance in suppressing unavailable examination requests and the additional retries they induce.

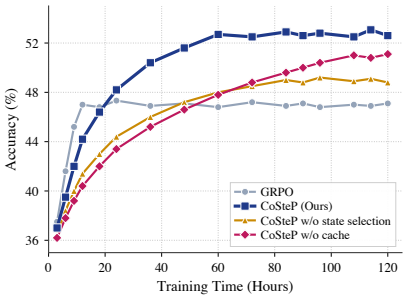


Figure 3: Training efficiency comparison on Qwen3-4B.

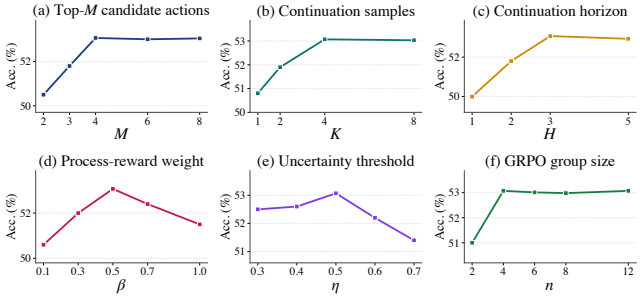


Figure 4: Hyperparameter sensitivity on Qwen3-4B. Each panel sweeps one hyperparameter while fixing the others at their defaults, and reports accuracy (%).

4.4 Training Efficiency Analysis

Figure 3 compares GRPO, CoSteP without state selection, CoSteP without rollout caching, and the full CoSteP under the same Qwen3-4B backbone and hardware budget. GRPO converges fastest but plateaus at the lowest accuracy, highlighting the limitation of outcome-only supervision. Removing the rollout cache requires counterfactual continuations to be freshly generated from the current policy, reducing reliance on potentially stale cached trajectories but increasing training time. Full CoSteP reuses compatible cached continuations, converges faster, and achieves higher accuracy. This may benefit from more diverse and stable samples for utility estimation, while reduced rollout cost enables more efficient optimization. Removing state selection also adds overhead by evaluating routine states. Together, rollout reuse and selective state evaluation make counterfactual process supervision computationally practical.

4.5 Expert Evaluation of Intermediate Decisions

We conduct an expert evaluation in which three senior physicians independently rate 300 randomly sampled intermediate decision states (100 per test set) along five dimensions. For each state, CoSteP and four baselines (GPT-5.4, Qwen3-32B, Med-Gemma-27B, and MDAgent) are evaluated under a fully paired setting. As shown in Figure 5, CoSteP achieves the highest score on all five dimensions. Its largest advantage appears in Diagnostic Gain (4.7 vs. 4.0 for the next-best baseline, Holm-adjusted $p < 0.001$), indicating that CoSteP selects more informative intermediate actions. It also leads in Cost-effectiveness (4.5 vs. 4.0, Holm-adjusted $p < 0.001$), consistent with the explicit cost penalties in its utility function. Overall, the expert ratings show that CoSteP's improvements in final accuracy are accompanied by higher-quality stepwise diagnostic decisions.

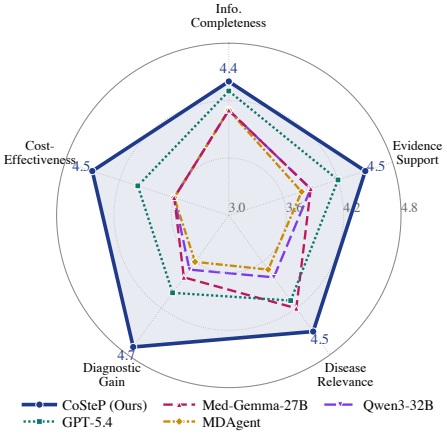


Figure 5: Expert evaluation of intermediate decisions across five dimensions.

4.6 Hyperparameter Sensitivity

Figure 4 sweeps the six hyperparameters with all others held at their defaults. The hyperparameters control distinct aspects of CoSteP: M is the number of top-ranked candidate actions retained at each high-uncertainty state; K is the number of continuation samples drawn per candidate action; H is

the continuation horizon, i.e., the maximum number of decision steps the model rolls out per branch; β is the global mixing weight of the process reward into the terminal reward; η is the uncertainty threshold that determines which states are selected for counterfactual evaluation; and n is the GRPO group size used for advantage normalization. As shown in Figure 4, accuracy peaks at M=4, K=4, H=3, β=0.5, η=0.5, and n=4, and we adopt these values as defaults.

4.7 Dynamic vs. Static Diagnosis

Table 3 compares CoSteP with a static baseline that receives all examination results at once. CoSteP must actively decide which examinations to issue rather than obtain all evidence at once. Despite this disadvantage, CoSteP improves accuracy from 50.23% to 53.07% while reducing AEC from 395.40 to 36.65 USD, corresponding to a 90.7% reduction. This confirms that CoSteP retains diagnostic competence under sequential, cost-aware constraints and achieves a markedly better accuracy-cost trade-off.

Table 3: Comparison of CoSteP and static diagnosis on the MIMIC-IV test set using the same Qwen3-4B backbone.

Method	Acc. (%)	AEC (USD)
Static Diagnosis	50.23	395.40
CoSteP (Ours)	**53.07**	**36.65** ▼ (−90.7%)

4.8 Case Study

Figure 6 presents an 80-year-old woman with slurred speech and memory impairment after a fall, atrial fibrillation, and an elevated INR of 5.2. Qwen3-4B immediately predicts ischemic stroke and produces an incorrect diagnosis. GRPO eventually identifies the subdural hematoma, but first requests a complete blood count and troponin, both of which provide limited discriminatory evidence, before ordering a head CT. In contrast, CoSteP recognizes that the fall and over-anticoagulation increase the likelihood of intracranial hemorrhage and directly requests a head CT, reaching the correct diagnosis with a single examination. Its counterfactual evaluation assigns the highest utility to head CT (+0.87), above brain MRI (+0.62), which is also correct but more expensive, while ECG and immediate stroke diagnosis receive lower utilities because they are non-discriminative or incorrect. This case illustrates how CoSteP favors diagnostically informative and cost-effective actions over redundant or misleading alternatives.

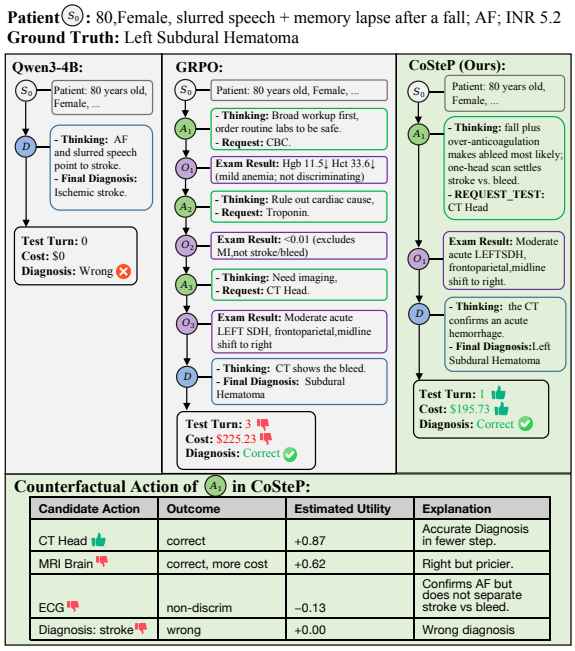


Figure 6: A representative case comparing Qwen3-4B, outcome-only GRPO, and CoSteP. CoSteP selects the most informative examination by evaluating its utility relative to counterfactual alternatives.

5 Conclusion

In this paper, we propose CoSteP, a reinforcement learning framework for sequential medical diagnosis. CoSteP identifies high-uncertainty decision states and evaluates each selected action relative to policy-generated counterfactual alternatives through downstream rollouts. A rollout cache further reduces the computational cost by reusing trajectories generated during training. Experiments on one in-domain and two out-of-domain benchmarks show that CoSteP achieves the highest average diagnostic accuracy and the lowest average examination cost. By enriching sparse outcome feedback with action-sensitive counterfactual signals from selected decision states, CoSteP provides a computationally practical form of step-informed trajectory supervision that alleviates credit dilution in sequential medical diagnosis.

AI USE STATEMENT

In this work, we used generative AI tools to assist with brainstorming, searching for relevant information, refining aspects of the research methodology and experimental design, and developing and editing software code. Generative AI tools were also used to improve the readability and linguistic clarity of the manuscript. All AI-assisted outputs were reviewed and assessed by the authors before being incorporated into the research or manuscript. Research decisions, experimental settings, implementation choices, result analysis, and scientific claims were determined and verified by the authors. We take full responsibility for the final content of this work, including any text, code, or research artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work investigates reinforcement learning for sequential medical diagnosis using de-identified clinical data. The MIMIC-IV and ClinicalBench datasets used in our experiments contain de-identified patient records, and the private hospital dataset used for out-of-domain evaluation consists of retrospective and fully de-identified inpatient cases. No personally identifiable patient information is used in the experiments reported in this work. No direct patient interaction or intervention was conducted. Access to and use of the private hospital dataset were authorized by the collaborating hospital and conducted in accordance with its institutional data governance and privacy requirements. The proposed method is evaluated retrospectively and is intended for research purposes only. Although CoSteP is designed to improve diagnostic accuracy and examination efficiency, model-generated diagnostic decisions may still be incorrect, incomplete, or inappropriate in real clinical settings. The system should therefore not be interpreted as a substitute for professional medical judgment or deployed for autonomous clinical decision-making without prospective validation, appropriate regulatory review, and human oversight.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. Section 3 provides a complete description of the proposed CoSteP framework, including the sequential diagnosis formulation, decision-critical state identification, counterfactual step pricing, and reward integration. Appendix A.1 describes the datasets, data processing and unified data format, as well as the construction of examination costs. Appendix A.2 provides the complete prompt templates used for sequential diagnosis, examination matching, and diagnosis evaluation. Appendix A.3 reports the model configuration, training and rollout settings, hyperparameters, computational setup, and evaluation protocol. Appendix A.4 further details the physician evaluation procedure and statistical analysis. The main experimental results are reported over three independent runs with different random seeds. In addition, the implementation of CoSteP is provided in the Supplementary Materials to facilitate reproduction of the reported experiments.

REFERENCES

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Medagentsim: Self-evolving multi-agent simulations for realistic clinical interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 362–372. Springer, 2025.

Ruike Cao, Shaojie Bai, Fugen Yao, Liang Dong, Jian Xu, and Li Xiao. Atpo: Adaptive tree policy optimization for multi-turn medical dialogue, 2026. URL https://arxiv.org/abs/2603.02216.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024. URL https://arxiv.org/abs/2412.18925.

Xi Chen, Hanyu Zhou, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Ziyuan Qin, Hairui Li, Xue Zhang, Yingman Guo, et al. Grounding large language models in clinical diagnostics. *Nature Communications*, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Yufei He, Juncheng Liu, Zhiyuan Hu, Yulin Chen, Yue Liu, Yuan Sui, Yibo Li, Nuo Chen, Jun Hu, Bryan Hooi, et al. Evoclinician: A self-evolving agent for multi-turn medical diagnosis via test-time evolutionary learning. *arXiv preprint arXiv:2601.22964*, 2026.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2607–2615, 2024.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*, 2024.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

Anisio Lacerda, Gisele Pappa, Adriano César Machado Pereira, Wagner Meira Jr, and Alexandre Guimaraes de Almeida Barros. Evaluation of medical large language models: taxonomy, review, and directions. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 10528–10536, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

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

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, et al. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*, 2025.

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

Qi Peng, Yi Cai, Jiankun Liu, Quan Zou, Xing Chen, Zheng Zhong, Zefeng Wang, Jiayuan Xie, and Qing Li. Integration of multi-source medical data for medical diagnosis question answering. *IEEE Transactions on Medical Imaging*, 44(3):1373–1385, 2024.

Qi Peng, Jialin Cui, Jiayuan Xie, Yi Cai, and Qing Li. Tree-of-reasoning: Towards complex medical diagnosis via multi-agent reasoning with evidence tree. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1744–1753, 2025.

Qi Peng, Jiatong Li, Sirui Huang, Yiyang Jiang, Kaisong Gong, Ronger Ding, Shijie Ye, Changmeng Zheng, Yi Cai, Xiaobo Yang, Jin Huang, Xiao-Yong Wei, and Qing Li. Aligning clinical needs and ai capabilities: A survey on llms for medical reasoning, 2026. URL https://arxiv.org/abs/2607.07761.

Pengcheng Qiu, Chaoyi Wu, Junwei Liu, Qiaoyu Zheng, Yusheng Liao, Haowen Wang, Yun Yue, Qianrui Fan, Shuai Zhen, Jian Wang, et al. Evolving diagnostic agents in a virtual clinical environment. *arXiv preprint arXiv:2510.24654*, 2025.

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

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-m1: Pushing the medical capability of large language models. *arXiv preprint arXiv:2502.12671*, 2025a.

Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization. *arXiv preprint arXiv:2505.15107*, 2025b.

Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2982–2990, 2022.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, and Jiayi Wang. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, et al. Process vs. outcome reward: Which is better for agentic rag reinforcement learning. *Advances in Neural Information Processing Systems*, 38: 58701–58729, 2026.

Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 419–428, 2024.

A Appendix

A.1 Dataset Description

MIMIC-IV. MIMIC-IV is a large-scale, publicly accessible critical-care database that contains de-identified electronic health records collected at a tertiary academic medical center, primarily covering intensive-care and emergency-department admissions. We use it as our in-domain corpus and follow the data-construction pipeline of DiagGym (Qiu et al., 2025): each discharge summary is parsed into a structured patient state, namely the chief complaint, history of present illness, past medical history, and physical examination, together with the chronologically ordered laboratory and imaging examinations, their results, and the final discharge diagnosis. The resulting cases predominantly cover critical-care and emergency-medicine conditions, including cardiovascular events, sepsis and other infectious processes, gastrointestinal and hepatobiliary disorders, acute renal injury, respiratory failure, and neurological emergencies. We use $23{,}377$ cases for training and hold out 750 further cases as the in-domain test set.

ClinicalBench. ClinicalBench (Yan et al., 2026) is an end-to-end multi-departmental clinical diagnostic benchmark constructed from real-world electronic medical records donated by Chinese tertiary hospitals, with strict de-identification and quality control. It contains $1{,}500$ patient cases spanning 24 clinical departments, ranging from internal medicine, cardiology, neurology, gastroenterology, and oncology to surgery, obstetrics and gynecology, pediatrics, and otolaryngology, and covers roughly 150 distinct diseases. Each case preserves the complete diagnostic trajectory: preliminary reception, history taking, ordered examinations and their reports, differential reasoning,

and final diagnosis. Because its source distribution differs substantially from MIMIC-IV, we use the entire 1,500-case set as a held-out out-of-domain test set; no part of it is used during training.

Private Dataset. The Private Dataset is curated in collaboration with a partner tertiary hospital and consists of 952 fully de-identified inpatient cases. Every case spans the complete admission-to-discharge trajectory and is annotated with chief complaint, history of present illness, past medical history, physical examination, laboratory and imaging examinations together with their reports, and the discharge diagnosis. The disease distribution exhibits a long-tailed shape and spans a broad spectrum of common inpatient conditions across multiple organ systems: hepatobiliary disease (e.g., hepatic cyst, hepatic hemangioma), cardiovascular disease (e.g., hypertension at multiple risk grades, aortic sclerosis), endocrine and metabolic disorders (e.g., electrolyte disturbances, nodular goiter), genitourinary and renal conditions (e.g., bilateral/unilateral renal cysts, renal calculi), gastrointestinal disease (e.g., chronic gastritis with or without erosion, intestinal obstruction), and respiratory disease (e.g., obstructive pneumonia, pulmonary nodules). The corpus thus complements MIMIC-IV with hospital admissions whose acuity profile and disease mix differ from those of intensive- and emergency-care units. We use the entire dataset as a second held-out out-of-domain test set.

Unified examination cost. To enable cost-aware learning and evaluation, we attach a unified U.S. dollar (USD) cost to every examination item across the three datasets, with values drawn from publicly available U.S. medical fee schedules. Laboratory tests are priced using the CMS Clinical Laboratory Fee Schedule (CLFS)[1]; imaging, electrocardiography, electroencephalography, echocardiography, and other non-laboratory diagnostic procedures are priced using the CMS Physician Fee Schedule (PFS)[2]; facility-heavy outpatient procedures are priced using the CMS Outpatient Prospective Payment System Addendum B[3]. To avoid double-charging when multiple sub-items belong to the same billing event, we further introduce a panel-aware billing-group map that consolidates fine-grained keys into parent groups: for example, `BLOOD Na`, `BLOOD K`, `BLOOD Cl`, `BLOOD Creat`, `BLOOD Glucose`, and `BLOOD UreaN` share one `BMP_PANEL`; `WBC`, `Hgb`, `Hct`, and `Platelet` share one `CBC_PANEL`; and urine micro-tests share one `UA_PANEL`. At runtime, ordering any sub-item charges its parent billing group exactly once per case.

Unified data format. We process all three datasets into a single record format so that the agent interacts with them through identical interfaces. Each record contains a case summary (chief complaint, present illness, past medical history, and personal, family, and allergy histories), a key-pertinent-results dictionary (physical examinations together with all laboratory and imaging findings), the final diagnosis, and the per-case examination cost map in USD. Table 4 shows one sample record drawn from MIMIC-IV.

Table 4: A case in the MIMIC-IV dataset.

"note_id": "xxxxxxxx-xx-xx",
"case_summary": "- Patient Information: ___ year old female\n- Chief Complaint: Abdominal pain\n- History of Present Illness: The patient reports gradual onset right lower quadrant (RLQ) pain with radiation to the back, accompanied by nausea and a small amount of diarrhea. She denies hematuria. The pain is described as sharp, moderate in severity, and sudden in onset. She states it feels similar to prior episodes of renal colic.\n- Past Medical History: Knee surgery and one episode of nephrolithiasis\n- Personal History: None\n- Family History: Non-contributory\n- Allergy History: No Known Allergies / Adverse Drug Reactions",
"key_pertinent_results_dict": {
"Vital Signs (Physical Examination)": "Temp: 99.2, HR: 80, BP: 122/81, Resp: 14, O(2)Sat: 100",
"Constitutional (Physical Examination)": "Comfortable",
"HEENT (Physical Examination)": "Normocephalic, atraumatic, Pupils equal, round and reactive to light, Extraocular muscles intact, Oropharynx within normal limits",

Continued on next page

[1] https://www.cms.gov/medicare/payment/fee-schedules/clinical-laboratory-fee-schedule-clfs

[2] https://www.cms.gov/medicare/payment/fee-schedules/physician

[3] https://www.cms.gov/medicare/payment/prospective-payment-systems/hospital-outpatient-pps/quarterly-addenda-updates

Table 4 – Continued from previous page

"Chest (Physical Examination)": "Clear to auscultation",
"Cardiovascular (Physical Examination)": "Regular Rate and Rhythm, Normal first and second heart sounds",
"Abdominal (Physical Examination)": "Soft, Nondistended, TTP mcburney's point",
"GU/Flank (Physical Examination)": "No costovertebral angle tenderness",
"Extr/Back (Physical Examination)": "No cyanosis, clubbing or edema",
"Skin (Physical Examination)": "No rash, Warm and dry",
"Neuro (Physical Examination)": "Speech fluent",
"Psych (Physical Examination)": "Normal mood, Normal mentation",
"Additional Findings (Physical Examination)": "No petechiae",
"URINE CULTURE": "Organism: ESCHERICHIA COLI; Isolate Number: 1.0; Antibiotic: AMPICILLIN (Dilution Text: 8, Comparison: =, Value: 8.0); Interpretation: S; Comments: Culture workup discontinued. Further incubation showed contamination with mixed skin/genital flora. Clinical significance of isolate(s) uncertain. Interpret with caution.",
"Bacteria": "Comments: FEW.",
"Urine Analysis": "Bilirubin: Units: mg/dL; Comments: NEG.; Epithelial Cells: Value: 17; Numeric Value: 17.0; Units: #/hpf; Glucose: Units: mg/dL; Comments: NEG.; Ketone: Units: mg/dL; Comments: TR.; Leukocytes: Comments: SM.; Nitrite: Comments: NEG.; pH: Value: 6.5; Numeric Value: 6.5; Units: units; Reference Range: 5.0–8.0; Protein: Value: 30; Numeric Value: 30.0; Units: mg/dL; RBC: Value: 1; Numeric Value: 1.0; Units: #/hpf; Reference Range: 0.0–2.0; Specific Gravity: Value: 1.024; Numeric Value: 1.024; Reference Range: 1.001–1.035; Urine Appearance: Comments: Hazy.; Urine Color: Comments: Yellow.; Urine Mucous: Comments: MANY.; Urobilinogen: Value: 2; Numeric Value: 2.0; Units: mg/dL; Reference Range: 0.2–1.0; Flag: abnormal; WBC: Value: 5; Numeric Value: 5.0; Units: #/hpf; Reference Range: 0.0–5.0",
"Blood": "Comments: NEG.",
"Yeast": "Value: NONE",
"Urine Pregnancy Test": "HCG, Urine, Qualitative: Units: +/-; Comments: NEGATIVE. FOR QUANTITATION OF POSITIVES, SEND SERUM FOR HCG.",
"Length of Urine Collection": "Comments: RANDOM.",
"Anion Gap": "Value: 18; Numeric Value: 18.0; Units: mEq/L; Reference Range: 8.0–20.0",
"Comprehensive Metabolic Panel": "Bicarbonate: Value: 24; Numeric Value: 24.0; Units: mEq/L; Reference Range: 22.0–32.0",
"Kidney Function Tests": "Chloride: Value: 101; Numeric Value: 101.0; Units: mEq/L; Reference Range: 96.0–108.0; Creatinine: Value: 0.9; Numeric Value: 0.9; Units: mg/dL; Reference Range: 0.4–1.1; Glucose: Numeric Value: 149.0; Units: mg/dL; Reference Range: 70.0–100.0; Flag: abnormal; Comments: IF FASTING, 70–100 NORMAL, >125 PROVISIONAL DIABETES.; Potassium: Numeric Value: 4.8; Units: mEq/L; Reference Range: 3.3–5.1; Comments: HEMOLYSIS FALSELY ELEVATES K.; Sodium: Value: 138; Numeric Value: 138.0; Units: mEq/L; Reference Range: 133.0–145.0; Urea Nitrogen: Value: 17; Numeric Value: 17.0; Units: mg/dL; Reference Range: 6.0–20.0",
"Estimated GFR": "Estimated GFR (MDRD equation): Comments: Using this patient's age, gender, and serum creatinine value of 0.9. Estimated GFR = 67 if non African-American (mL/min/1.73 m^2). Estimated GFR $>$ 75 if African-American (mL/min/1.73 m^2). For comparison, mean GFR for age group 40–49 is 99 (mL/min/1.73 m^2). GFR $<$ 60 = Chronic Kidney Disease, GFR $<$ 15 = Kidney Failure.",
"Light Green Top Hold": "Comments: HOLD.",
"Liver Function Test": "INR(PT): Value: 0.9; Numeric Value: 0.9; Reference Range: 0.9–1.1; PT: Numeric Value: 9.7; Units: sec; Reference Range: 9.4–12.5; Comments: VERIFIED.",
"Coagulation Profile": "PTT: Numeric Value: 23.2; Units: sec; Reference Range: 25.0–36.5; Flag: abnormal; Comments: VERIFIED.",

Continued on next page

Table 4 – Continued from previous page

“Complete Blood Count”: “Basophils: Value: 0.4; Numeric Value: 0.4; Units: %; Reference Range: 0.0–2.0; Eosinophils: Value: 0.6; Numeric Value: 0.6; Units: %; Reference Range: 0.0–4.0; Hematocrit: Value: 46.7; Numeric Value: 46.7; Units: %; Reference Range: 36.0–48.0; Hemoglobin: Value: 16.1; Numeric Value: 16.1; Units: g/dL; Reference Range: 12.0–16.0; Flag: abnormal; Lymphocytes: Value: 6.5; Numeric Value: 6.5; Units: %; Reference Range: 18.0–42.0; Flag: abnormal; MCH: Value: 32.3; Numeric Value: 32.3; Units: pg; Reference Range: 27.0–32.0; Flag: abnormal; MCHC: Value: 34.6; Numeric Value: 34.6; Units: %; Reference Range: 31.0–35.0; Monocytes: Value: 4.5; Numeric Value: 4.5; Units: %; Reference Range: 2.0–11.0; Neutrophils: Value: 88.0; Numeric Value: 88.0; Units: %; Reference Range: 50.0–70.0; Flag: abnormal; Platelet Count: Value: 273; Numeric Value: 273.0; Units: K/uL; Reference Range: 150.0–440.0; RDW: Value: 12.9; Numeric Value: 12.9; Units: %; Reference Range: 10.5–15.5; Red Blood Cells: Value: 5.00; Numeric Value: 5.0; Units: m/uL; Reference Range: 4.2–5.4; White Blood Cells: Value: 18.5; Numeric Value: 18.5; Units: K/uL; Reference Range: 4.0–11.0; Flag: abnormal”,

“Mean Corpuscular Volume”: “MCV: Value: 93; Numeric Value: 93.0; Units: fL; Reference Range: 82.0–98.0”,

“Green Top Hold, plasma”: “Green Top Hold (plasma): Value: HOLD. DISCARD GREATER THAN 4 HOURS OLD.”,

“Lactate”: “Value: 2.2; Numeric Value: 2.2; Units: mmol/L; Reference Range: 0.5–2.0; Flag: abnormal”,

“CT of abdomen and pelvis”: “Preliminary Report 1. Acute appendicitis.”

},

“final_diagnosis”: “Acute appendicitis. The diagnosis is supported by the patient’s clinical presentation of RLQ pain, nausea, and diarrhea, along with laboratory findings of leukocytosis and neutrophilia, elevated lactate, and CT imaging confirming a dilated, fluid-filled appendix.”,

“diagnosis_results”: “Acute appendicitis”,

“exam_cost_map (USD)”: {

“Vital Signs (Physical Examination)”: 0.00,

“Constitutional (Physical Examination)”: 0.00,

“HEENT (Physical Examination)”: 0.00,

“Chest (Physical Examination)”: 0.00,

“Cardiovascular (Physical Examination)”: 0.00,

“Abdominal (Physical Examination)”: 0.00,

“GU/Flank (Physical Examination)”: 0.00,

“Extr/Back (Physical Examination)”: 0.00,

“Skin (Physical Examination)”: 0.00,

“Neuro (Physical Examination)”: 0.00,

“Psych (Physical Examination)”: 0.00,

“Additional Findings (Physical Examination)”: 0.00,

“URINE CULTURE”: 8.50,

“Bacteria”: 5.00,

“Urine Analysis”: 5.00,

“Blood”: 5.00,

“Yeast”: 5.00,

“Urine Pregnancy Test”: 5.00,

“Length of Urine Collection”: 5.00,

“Anion Gap”: 8.46,

“Comprehensive Metabolic Panel”: 10.56,

“Kidney Function Tests”: 8.68,

“Estimated GFR”: 8.68,

“Light Green Top Hold”: 5.00,

“Liver Function Test”: 5.00,

“Coagulation Profile”: 18.00,

“Complete Blood Count”: 6.47,

“Mean Corpuscular Volume”: 5.00,

“Green Top Hold, plasma”: 5.00,

“Lactate”: 11.57,

Continued on next page

Table 4 – Continued from previous page

"CT of abdomen and pelvis": 29.73
}

A.2 PROMPT TEMPLATES

This section shows the prompts used in our framework, each driving a distinct module of the sequential-diagnosis pipeline. The *Sequential Diagnosis Agent Prompt* is the policy prompt: it instructs the actor to maintain an explicit short list of differential diagnoses at every step, choose a single examination that best discriminates among them, and commit to a final diagnosis once the evidence is sufficient. The *Exam Key Matching Prompt* is a lightweight terminology-alignment module: it maps a free-form examination request from the agent to one of the available examination keys in the case record, so that the environment can return the matching result without leaking values. The *Diagnosis Evaluation Prompt* is the outcome judge: given a model prediction and the ground-truth diagnosis, it returns a single binary verdict "Correct"/"Wrong" that defines the terminal accuracy reward.

Sequential Diagnosis Agent Prompt

You are a medical AI assistant performing sequential diagnosis. You are given a patient's initial clinical information and must iteratively request medical examinations/tests to gather more information, then provide a final diagnosis when you have enough evidence.

Core Strategy

- Start broad: prioritize the examination with the greatest expected diagnostic value relative to its cost. Use foundational tests only when they meaningfully narrow the current differential.
- Be efficient: each test should meaningfully change your differential or confirm/exclude a leading diagnosis.
- Be specific: when diagnosing, name the precise condition (e.g., "Community-acquired pneumonia" not just "Lung infection").

Rules

1. At each step, explicitly list your top 2–3 differential diagnoses before deciding the next action.
2. Choose the test that best discriminates between your leading differentials.
3. Request only ONE specific test per turn.
4. Once key discriminating evidence is obtained, commit to a final diagnosis — do not over-investigate.
5. Use standard medical terminology for test names.

Available Examinations (representative examples)

Laboratory: Complete Blood Count, Comprehensive Metabolic Panel, Liver Function Test, Kidney Function Tests, Urine Analysis, Coagulation Profile, Troponin, Lactate, Blood Culture, Urine Culture, Arterial Blood Gas, Lipase, D-dimer, Thyroid Function Tests, Blood Glucose, HbA1c, Procalcitonin, BNP/NT-proBNP, Iron Studies, Vitamin B12/Folate.

Imaging & Procedures: Chest X-ray, CT Abdomen and Pelvis, CT Head, CT Chest, Abdominal Ultrasound, Echocardiography, ECG, MRI Brain, Lumbar Puncture, Endoscopy.

Common Diagnoses in This Setting

Pneumonia, Urinary tract infection, Acute appendicitis, Cellulitis, Acute pancreatitis, Acute cholecystitis, Heart failure, Subarachnoid hemorrhage, Acute kidney injury, Upper GI bleed, Sepsis, Pulmonary embolism, Diabetic ketoacidosis, Acute myocardial infarction, Stroke, Bowel obstruction, Deep vein thrombosis.

Response Format

If you need more information (request a test):

```
THINKING: [Top 2--3 differential diagnoses ranked by likelihood,
and why this test best discriminates between them]
ACTION: REQUEST_TEST
Test needed:  [one specific test/examination name]
Reason:  [why this test is needed]
```

If you have enough information (final diagnosis):

```
THINKING: [Brief summary of key findings supporting the diagnosis]
ACTION: FINAL_DIAGNOSIS
Diagnosis:  [your final diagnosis]
Reason:  [clinical reasoning supporting the diagnosis]
```

Example (Requesting imaging)
`THINKING`: Patient presents with fever, productive cough, and crackles on lung auscultation. Top differentials: (1) Community-acquired pneumonia, (2) Bronchitis, (3) Atypical pneumonia. Chest X-ray will best discriminate by detecting pulmonary infiltrates.
`ACTION: REQUEST_TEST`
`Test needed`: Chest X-ray
`Reason`: To confirm pulmonary infiltrates and assess extent of infection.

Figure 7: The system prompt that drives the diagnostic agent at every interaction step.

Exam Key Matching Prompt

You are a medical terminology matching assistant. Given a list of available examination names and a requested test, determine which examination name (if any) matches the request.
Rules
- Match using medical knowledge: abbreviations, synonyms, and alternative names all count (e.g., "CBC" = "Complete Blood Count", "ECG" = "Electrocardiogram", "CXR" = "Chest X-ray").
- Return EXACTLY the matching name from the available list — do not modify it.
- If no match exists, output "`NO_MATCH`".

Response Format
If matched: output ONLY the exact examination name from the list, nothing else.
If not matched: output ONLY "`NO_MATCH`", nothing else.
Examples
Available: `["Complete Blood Count", "Chest X-ray", "Basic Metabolic Panel"]`
Requested: `CBC`
Response: `Complete Blood Count`
Available: `["Electrocardiogram", "Troponin I", "Chest X-ray"]`
Requested: `ECG`
Response: `Electrocardiogram`
Available: `["Complete Blood Count", "Basic Metabolic Panel"]`
Requested: `CT Abdomen`
Response: `NO_MATCH`
User Template
`Available: {exam_keys}`
`Requested: {requested_test}`
`Response:`

Figure 8: The lightweight prompt that aligns the agent's free-form test request with one of the case's available examination keys.

Diagnosis Evaluation (Judge) Prompt

You are a medical diagnosis evaluation system. Your task is to judge whether a model-predicted diagnosis is clinically correct when compared to the ground truth diagnosis.
Evaluation Rules
1. The same disease may have multiple names or aliases (e.g., "Heart attack" = "Myocardial infarction" = "MI").
2. Abbreviations and full names are equivalent (e.g., "COPD" = "Chronic Obstructive Pulmonary Disease").
3. If the predicted diagnosis captures the core disease correctly but mentions additional complications or details, it is still CORRECT.
4. If the predicted diagnosis is a broader category that includes the ground truth (e.g., `pred`="Cardiac disease" vs. `gt`="Acute myocardial infarction"), it is WRONG — the prediction must be specific enough.
5. If the predicted diagnosis is a different disease entirely, it is WRONG.
6. Only evaluate the diagnosis itself — ignore differences in reasoning, treatment, or symptom descriptions.

Response Format
Output ONLY one word: "`Correct`" or "`Wrong`". Do not output any other content.

Examples
Example 1 — Correct (same disease, different wording):
`[pred_diag]`: Acute myocardial infarction
`[gt_diag]`: Heart attack
`Output: Correct`
Example 2 — Correct (prediction includes ground truth with extra detail):
`[pred_diag]`: Community-acquired pneumonia with possible sepsis
`[gt_diag]`: Pneumonia
`Output: Correct`
Example 3 — Wrong (different disease):
`[pred_diag]`: Gastroesophageal reflux disease
`[gt_diag]`: Peptic ulcer disease
`Output: Wrong`
User Template
`[pred_diag]`
{`pred_diag`}
`[gt_diag]`
{`gt_diag`}

Figure 9: The judge prompt that turns the predicted final diagnosis into the binary terminal reward used during training and evaluation.

A.3 Hyperparameters and Training Details

We detail in this section the hyperparameters used for CoSteP training and evaluation. We use Qwen3-4B-Instruct as the backbone for all training variants. To process examination actions, we first extract the examination action from the model response using regular-expression rules. The extracted action is then matched to the case-specific examination list through both regex matching and LLM-based semantic matching. The same extracted action token is used to compute action entropy, which guides the selection of high-uncertainty decision states. Any action that cannot be parsed is treated as invalid and incurs the unavailable-action penalty. Diagnosis correctness is judged by an LLM-as-a-Judge using GPT-5.4. We base our implementation on the verl framework and report a typical training configuration. All training runs are performed on 2 NVIDIA H20 GPUs.

Backbone and Rollout. The actor is initialized from Qwen3-4B-Instruct-2507 and trained with multi-turn rollouts capped at 8 assistant turns per trajectory, a maximum prompt length of 2048 tokens, and a maximum response length of 4096 tokens. Rollouts are served by SGLang with FP8 quantization, asynchronous mode, and a tensor-parallel size of 2. Diagnosis correctness is judged by a closed-source GPT-5.4 model accessed via API, using the prompt detailed in Figure 9; the same judge is used during both training and evaluation to ensure that the scoring criterion remains consistent throughout.

Optimization. We optimize the actor with GRPO: $n = 4$ rollouts per prompt, a global training batch of 16 prompts (yielding $16 \times 4 = 64$ trajectories per iteration), a PPO mini-batch of 16 and a per-GPU PPO micro-batch of 8, and a single PPO inner epoch. The actor uses learning rate 5×10^{-6}, bfloat16 parameters with FSDP and gradient checkpointing, and FlashAttention-2; KL regularization to the reference model is disabled (`use_kl_loss=False`, `use_kl_in_reward=False`). We run 3 epochs of training and save/evaluate every 200 optimizer steps.

CoSteP-specific Hyperparameters. The counterfactual rollout settings follow the notation in the main paper: top-M candidate actions $M = 4$, $K = 4$ continuations per branch, continuation horizon $H = 3$, and $n_s = 8$ action samples for uncertainty estimation, with the uncertainty threshold $\eta = 0.5$. To keep the different reward components on comparable numerical scales and improve optimization stability, we map the raw monetary cost of each examination to one of several standardized ordinal cost levels, ranging from 0 to 6, where a higher level indicates a more expensive examination. These normalized cost levels are used only for reward computation during training, while the original monetary costs are retained for reporting the Average Exam Cost in evaluation. The terminal-reward shaping coefficients are $\lambda_{\text{test}} = 0.05$, $\lambda_{\text{cost}} = 0.02$, and $\lambda_{\text{na}} = 0.10$. The process reward is mixed with the terminal reward using $\beta = 0.5$ and clipped to $[-c, c]$, where $c = 1.0$.

The continuation cache is maintained across training batches with a capacity of 8192 entries and is continuously updated as new trajectories are generated.

Evaluation Setup. At evaluation time we use SGLang with the same generation length limits as training (max prompt 2048, max response 4096). The judge prompt (Figure 9) returns a binary correctness label, which is averaged over the test set to obtain Accuracy; Average Number of Examinations (AEN) is obtained from the real trajectory and Average Exam Cost (AEC, USD) is computed using the unified USD cost map, respectively. All test sets (in-domain MIMIC-IV, ClinicalBench, Private) are evaluated with the same configuration; no test-time tuning is performed.

A.4 Expert Evaluation Protocol

This appendix details the expert evaluation reported in the main text. The study aims to assess the quality of *intermediate* diagnostic decisions, rather than only the final diagnosis.

Participants and Sampling. We invited three senior physicians with at least five years of clinical experience. From each of the three test sets (MIMIC-IV, ClinicalBench, Private) we randomly sampled 100 intermediate decision states, yielding 300 states in total. For each sampled state, all five compared methods (CoSteP, GPT-5.4, Qwen3-32B, Med-Gemma-27B, MDAgent) are evaluated on the *same* state: each method's reasoning and proposed next action are collected at that state and presented to the raters in randomized order with model identity blinded. Each rater therefore scores five method outputs per state, conditioned on identical patient information (chief complaint, history, prior findings), so that method comparisons within a state are fully paired. Downstream outcomes and the final diagnosis are withheld to avoid bias.

Scoring Dimensions. Each node is rated on a 1–5 Likert scale along five dimensions. The first two assess the *reasoning* quality; the last three assess the proposed *next action*.

- **Information Completeness.** Whether the model fully and accurately exploits the available case state, including symptoms, history, prior results, and key positive/negative findings.
  - 1: Clearly misses or misreads key information, breaking the basis of subsequent reasoning.
  - 2: Uses only part of the information; many important clinical findings are overlooked or misread.
  - 3: Captures the main information but still misses some important details.
  - 4: Captures most key information accurately, with only minor omissions.
  - 5: Comprehensively and accurately captures all key positives and negatives.
- **Evidence Support.** Whether each step of reasoning is grounded in the currently available evidence, with no unwarranted leaps or contradictions.
  - 1: Reasoning lacks evidential support or contradicts the current state.
  - 2: Contains many unsupported inferences, over-interpretations, or weakly grounded claims.
  - 3: Mostly acceptable but with some weakly supported judgments.
  - 4: Generally well supported by the current evidence, with only minor gaps.
  - 5: Every key inference is explicitly traceable to the current evidence.
- **Disease Relevance.** Whether the proposed next action is directed at the current target disease, suspected diseases, or critical differentials.
  - 1: Action is essentially unrelated to any plausible disease.
  - 2: Only indirectly related; weakly aligned with the most likely candidate diseases.
  - 3: Loosely related to some candidate but not well targeted.
  - 4: Reasonably aligned with the main suspected diseases or differentials.
  - 5: Tightly focused on the most likely disease or the most critical differential.
- **Diagnostic Gain.** Whether the action is expected to provide new and valuable diagnostic information — e.g., reducing uncertainty, discriminating between key differentials, or unblocking the next reasoning step.
  - 1: Provides almost no new diagnostic information.

2: Carries some informational value but contributes little to narrowing the differential.
3: Provides some information but is not the most informative option here.
4: Has clear diagnostic value and meaningfully helps separate candidates.
5: A pivotal step that markedly reduces diagnostic uncertainty.

- **Cost-effectiveness.** Whether the action is economically and clinically proportionate, considering monetary cost, time, invasiveness, risk, and redundancy with prior examinations.
  1: Clearly excessive: cost, invasiveness, or risk far outweighs the expected benefit.
  2: Has some necessity but is too costly, burdensome, or complex for the current stage.
  3: Acceptable but not the most economical or appropriate option.
  4: Reasonably priced and appropriately necessary for the current diagnostic stage.
  5: High diagnostic value with low burden and clear necessity.

Protocol and Statistics. The three raters scored every sampled state independently, with no communication during scoring. Per-state per-method scores are obtained by averaging the three raters' values, and per-method scores reported in the main paper are means over all 300 sampled states. Inter-rater agreement, measured by Krippendorff's α on the ordinal scale, was above 0.7 on every dimension, indicating substantial agreement. To test whether CoSteP outperforms each baseline, we perform a paired Wilcoxon signed-rank test on the per-state mean scores (Co vs. baseline) along each of the five dimensions, with Holm correction for multiple comparisons across dimensions.