ProMed: Shapley Information Gain Guided Reinforcement Learning for Proactive Medical LLMs

Hongxin Ding[1,2,3*], **Baixiang Zhang**[1,2,3*], **Yue Fang**[1,2,3*], **Weibin Liao**[1,2,3*], **Xinke Jiang**[1,2,3], **Jinyang Zhang**[1,2,3], **Yinghao Zhu**[5], **Zheng Li**[2], **Liantao Ma**[1,3,4], **Junfeng Zhao**[2,3†], **Yasha Wang**[1,3,4†]

[1]National Engineering Research Center for Software Engineering, Peking University, China
[2]School of Computer Science, Peking University, Beijing, China
[3]Key Laboratory of High Confidence Software Technologies, Ministry of Education
[4]Peking University Information Technology Institute, Tianjin Binhai, China
[5]School of Computing and Data Science, The University of Hong Kong

Abstract

Interactive medical questioning is essential in clinical consultations, where physicians must actively gather necessary patient information. Yet existing medical Large Language Models (LLMs) predominantly follow a reactive paradigm, risking diagnostic errors by answering before seeking sufficient details. To bridge this gap, we propose **ProMed**, a reinforcement learning framework that transitions LLMs toward a proactive paradigm, enabling them to ask clinically valuable questions before decision-making. Central to ProMed is the Shapley Information Gain (SIG) reward, which quantifies a question's clinical utility as the amount of newly acquired information, while considering its contextual importance via Shapley values. We integrate SIG into a two-stage training pipeline: (1) SIG-Guided Model Initialization uses Monte Carlo Tree Search to construct high-reward interaction trajectories for supervision, and (2) SIG-Augmented Policy Optimization, with a novel SIG-guided Reward Distribution Mechanism that prioritizes informative questions for fine-grained optimization. Experiments on partial-information medical benchmarks show that ProMed significantly outperforms state-of-the-art methods by 6.29% on average and delivers a 54.45% gain over the reactive paradigm, and generalizes robustly to out-of-domain cases. Our codes are available at https://github.com/hxxding/ProMed.

1 Introduction

Clinical diagnosis fundamentally relies on interactive medical consultation. Patients typically initiate a clinical encounter with vague or incomplete chief complaints, requiring clinicians to proactively elicit relevant historical and symptomatic information through targeted questioning before an accurate diagnosis can be established. Recently, leveraging large language models (LLMs) (Singhal et al., 2025; Wu et al., 2024; Zhang et al., 2023) for clinical decision-making has emerged as an active research direction. Benefiting from large-scale pretraining on medical knowledge, LLMs have demonstrated promising performance on static clinical tasks, such as medical examinations (Ding et al., 2024; Jiang et al., 2025) and disease diagnosis (McDuff et al., 2025; Xu et al., 2025b). However, when deployed in interactive clinical settings, existing LLMs remain constrained by a **reactive paradigm** that generates predictions solely based on an initial query. In such cases, partial or biased information in the initial patient input can induce cognitive bias in the model's reasoning process, leading to erroneous clinical decisions, potential misdiagnosis, and compromised patient safety (as illustrated in Figure 1). Therefore, *enabling LLMs to transition from a **reactive paradigm** to a **proactive paradigm**, where models can systematically acquire clinically informative evidence through interaction and inquiry, has become a critical research problem for unbiased and reliable clinical decision-making.*

Recent efforts on **proactive paradigm** for interactive medical LLMs primarily rely on prompt engineering or supervised fine-tuning (SFT). Despite their demonstrated effectiveness, prompt-based approaches (Li et al., 2024; Hu et al., 2024; Liu et al., 2025a; Wang et al., 2025b; Zhu and Wu, 2025) merely induce questioning behaviors through carefully designed prompts. Through this approach, LLMs passively follow question-asking instructions, rather than genuinely transitioning to an active reasoning paradigm, resulting in limited performance gains. In contrast, SFT-based methods (Liu et al., 2025b; Liao et al., 2023) attempt to simu-

*These authors contribute equally.
†Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32481–32515
July 2-7, 2026

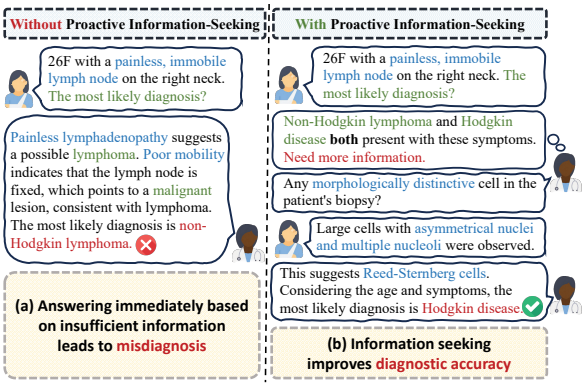


Figure 1: Clinical consultation example: relying on partial information causes misdiagnosis (a), actively seeking information enables accurate diagnosis (b).

late interactions by training on static multi-turn dialogues. However, such approaches lack robustness and adaptability to diverse and unpredictable patient scenarios encountered in real-world clinical practice. More critically, empirical studies (Liao et al., 2025a; Hong et al., 2024; Liao et al., 2025b) have shown that while SFT relying solely on positive feedback encourages models to ask clinically relevant questions, it also inadvertently amplifies incorrect or misleading inquiries. Therefore, *enabling a genuine* ***paradigm shift****, one that* ***promotes the generation of clinically informative questions*** *while* ***restraining erroneous or misleading inquiries****, remains a fundamental problem in interactive medical consultation.*

Reinforcement learning (RL) offers a promising solution for achieving this goal, as its **reward mechanism** enables explicit encouragement or suppression of model behaviors. Within this context, we identify **two key challenges** in applying RL to proactive interactive medical consultation.

- **Reward Modeling.** Quantifying the clinical value of a question has long been a central problem in interactive medical diagnosis. Existing studies typically rely on heuristic strategies, such as LLM–based scoring (Wang et al., 2025b) or leave-one-out evaluations that assess a question's impact on model confidence (Hu et al., 2024; Lee et al., 2025; Mazzaccara et al., 2024; Zhu and Wu, 2025). However, these approaches overlook a critical property of medical reasoning: its inherently **compositional nature**. Accurate diagnosis often depends on joint considerations of multiple clinical facts, and the utility of a single question may only emerge when combined with others. Consequently, isolated evaluation of individual questions can be fundamentally insufficient.
- **Reward Attribution.** Assigning terminal rewards to individual tokens along a trajectory is a long-standing challenge in RL. Naïve strategies such as uniformly distributing rewards across all tokens in a trajectory (e.g., GRPO (Shao et al., 2024)), **fail to distinguish clinically salient questions** from irrelevant ones, assigning them equal credits. Moreover, token-level averaging implicitly biases the model toward generating longer questions, as longer sequences receive more total rewards, thereby encouraging verbosity rather than clinical informativeness.

To address these challenges, we propose an RL framework for training **Pro**active **Med**ical LLMs (**ProMed**). **For Reward Modeling**, we introduce the novel *Shapley Information Gain (SIG)* reward mechanism. SIG utilizes Shapley values (Winter, 2002) from cooperative game theory to measure the importance of medical information while considering its interactions, yielding a context-aware information gain to precisely quantify questions' clinical utility. For **Reward Attribution**, we closely integrate SIG into RL through a two-stage design. *Stage 1: SIG-Guided Model Initialization* employs Monte Carlo Tree Search (MCTS) with SIG rewards to systematically explore optimal doctor-patient interaction trajectories for supervised warm-up, improving stability and convergence for weak initial policies (Wang et al., 2025a; Xu et al., 2025a) and alleviating the scarcity of high-quality medical interaction data. *Stage 2: SIG-Augmented Policy Optimization* incorporates SIG into GRPO with a novel *SIG-Guided Reward Distribution Mechanism*. Unlike standard GRPO that assigns uniform rewards to all tokens, our strategy allocates rewards proportionally to questions' utility, enabling more targeted, fine-grained policy optimization that reinforces the LLM's proactive ability.

Our contributions are as follows:

- **Insightfully**, we pioneer an RL framework ProMed to shift medical LLMs from reactive responders to proactive information seekers.
- **Technically**, we develop the SIG reward that leverages Shapley to model medical information interactions and quantify question utility, with a tailored reward distribution mechanism for fine-grained optimization.
- **Experimentally**, extensive evaluations demonstrate that ProMed significantly outperforms existing methods and exhibits robust generalization

to out-of-distribution (OOD) benchmarks.

- **Practically**, we construct two public benchmarks targeting interactive medical questioning with standardized splits to facilitate future research.

2 Preliminaries

We formulate the **Interactive Medical Questioning** task, which mirrors realistic clinical consultations where patients provide incomplete information during initial inquiries. Each patient case in the dataset $\mathcal{D} = \{\mathcal{X}_i\}_{i=1}^{N}$ is defined as $\mathcal{X} = \{Q, \mathcal{F}, A^*\}$ and consists of:

- $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$: complete set of atomic facts fully describing the patient's clinical condition, where each f_i is a minimal, self-contained information unit (e.g., symptom or lab result);
- Q: atomic clinical inquiry without information;
- A^*: ground-truth answer based on the full set $\mathcal{F}$.

We model the LLM as an **interactive agent** that proactively acquires information through multi-turn questioning. The interaction starts with a **partial information question** $Q_p = (F_p, Q)$, where $F_p \subset \mathcal{F}$ represents limited patient information (e.g., a chief complaint). At each turn t, the model updates its internal belief state s_{t-1} about the patient based on the current dialogue history $\mathcal{H}_{t-1} = (q_1, r_1), \ldots, (q_{t-1}, r_{t-1})$. It decides whether information is sufficient and takes action a_t: either asking a **follow-up question** q_t and receiving the patient response r_t, or terminating the interaction by outputting an answer A'.

3 Methodology

3.1 Overview

As illustrated in Figure 2, ProMed comprises three modules: **Shapley Information Gain Reward** quantifies the clinical utility of questions to guide training. **SIG-Guided Model Initialization** uses MCTS with SIG to explore high-quality interaction trajectories for SFT. **SIG-Augmented Policy Optimization** integrates SIG into GRPO with a reward distribution for fine-grained optimization.

3.2 Shapley Information Gain Reward

To guide LLMs' question-asking with accurate and clinical-aware rewards, we propose *Shapley Information Gain* (SIG), which quantifies question utility by measuring newly acquired information amount, while accounting for fact importance and interactions via cooperative game theory.

Atomic Fact Foundation. We measure *incremental information* elicited by each question as the number of newly acquired atomic facts. Specifically, we leverage the pre-constructed ground-truth atomic fact set $\mathcal{F} = f_1, f_2, \ldots, f_n$ defined in Section 2. This enables explicit tracking of acquired facts and per-question information gain.

State Approximation via Dynamic Understanding Generation. At dialogue turn t, after the model poses a follow-up question q_t and receives the patient response r_t, the dialogue history is updated to: $\mathcal{H}_t = \{(q_1, r_1), (q_2, r_2), \ldots, (q_t, r_t)\}$. We approximate the model's internal belief state s_t using a *doctor understanding prompt* (Appendix K) that instructs the model to articulate its understanding of patient condition U_t based on Q_p and $\mathcal{H}_t$. U_t captures the model's current grasp of patient information and serves as a proxy for s_t.

Fact-Level Information Gain. To quantify the informational value of question q_t, we measure the incremental change in model understanding following its response r_t. Using a high-capacity LLM-powered fact-checker to determine fact entailment ($f_i \in \mathcal{F}$) in U_t, raw Information Gain (IG) is defined as the increase in fact coverage between current and previous model understanding states:

$$IG(q_t) = \frac{1}{|\mathcal{F}|} \sum_{f_i \in \mathcal{F}} [\mathbf{1}(f_i \subseteq U_t) - \mathbf{1}(f_i \subseteq U_{t-1})] \quad (1)$$

where $\mathbf{1}(\cdot)$ is the indicator function based on the fact-checker's judgments. The IG score captures the number of newly acquired facts elicited by q_t, but assumes equal and independent contributions across all facts.

Atomic Fact Shapley Calculation. In clinical practice, information differs in diagnostic value and exhibits nontrivial interactions. For instance, a chest CT scan is typically more informative for pneumonia than a reported fever(Balafar et al., 2024). The straightforward recall-based IG fails to capture such distinctions. Traditional "leave-one-out" evaluations (Hu et al., 2024; Zhu and Wu, 2025) measure information importance by their inflicted change in model uncertainty, which treat facts in isolation and overlook their complex dependencies and interactions. For instance, diagnosing acute appendicitis relies on combined evidence: elevated white blood cell count, right lower abdominal tenderness, and Blumberg's sign (Snyder et al., 2018). Omitting one symptom may not significantly affect the model's prediction if others are absent, thereby underestimating its true clinical

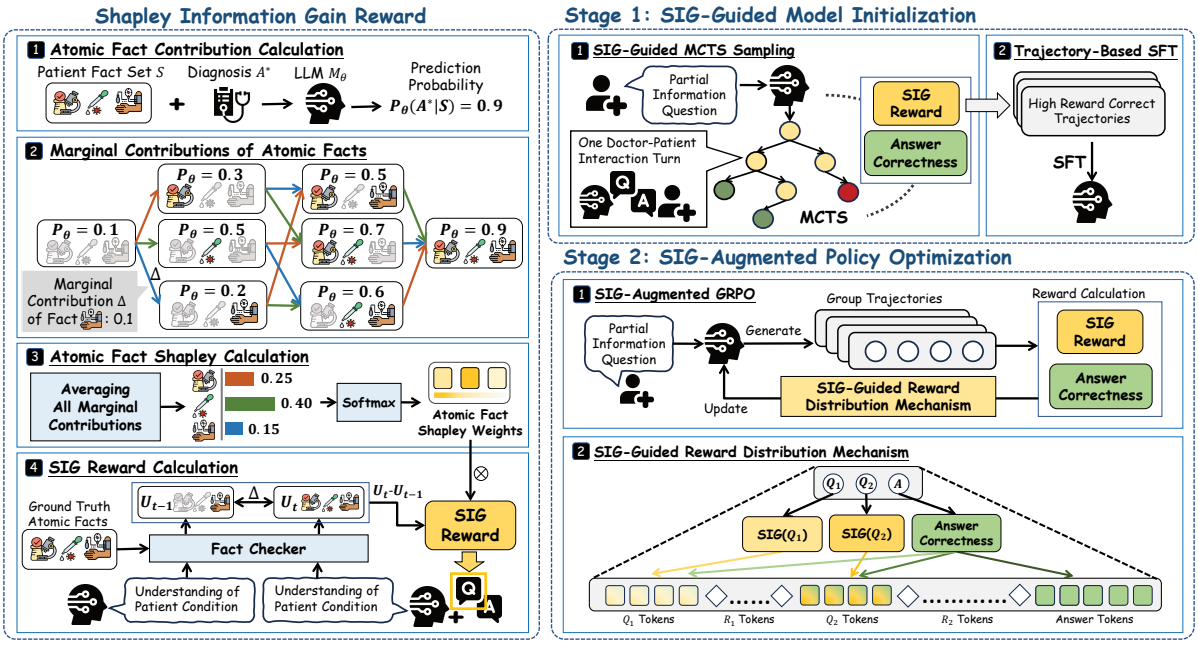


Figure 2: **ProMed framework. Shapley Information Gain Reward** calculates rewards for questions. **Stage 1** generates high-reward trajectories via MCTS for SFT. **Stage 2** distributes rewards and optimizes the policy via RL.

value. To capture both **(1) varying clinical importance** and **(2)complex interactions** of facts, we adopt the **Shapley value** (Winter, 2002) from cooperative game theory to more precisely and robustly attribute information importance.

Formally, given LLM M_θ parameterized by θ, atomic question Q and the desired answer A^*, the value of an atomic fact subset $S \subseteq \mathcal{F}$ is defined as the log-probability of predicting A^* based on Q and S:

$$v(S) = \log P_\theta(A^* \mid Q, S) \tag{2}$$

The value function quantifies the facts' contribution to the model's correct prediction, thus reflecting their utility. The Shapley value $\phi(f_i)$ of fact f_i is the expected marginal gain in $v(S)$ when f_i is added to all subsets:

$$\phi(f_i) = \sum_{S \subseteq \mathcal{F} \setminus \{f_i\}} \frac{1}{|\mathcal{F}| \binom{|\mathcal{F}|-1}{|S|}} [v(S \cup \{f_i\}) - v(S)] \tag{3}$$

This calculation captures individual fact importance via marginal contributions, and their complex interactions by considering all possible fact combinations, thus addressing the two aforementioned clinical factors.

Since enumerating $2^{|\mathcal{F}|}$ subsets is computationally infeasible, we employ a *Monte Carlo approximation* with online averaging. At iteration k, we sample a random permutation π_k of $\mathcal{F}$, compute the marginal contribution of each fact then update its Shapley estimation:

$$\phi^{(k)}(f_i) = \frac{k-1}{k} \phi^{(k-1)}(f_i) + \frac{1}{k} [v(S_i^{\pi_k} \cup f_i) - v(S_i^{\pi_k})]. \tag{4}$$

where $S_i^{\pi_k}$ denotes the set of facts preceding f_i in π_k. Crucially, we batch all evaluations within a single permutation into one model pass, leveraging parallel inference to significantly accelerate computation. The process terminates when estimates converge within a tolerance ϵ, enabling a controllable trade-off between efficiency and accuracy. Pseudo-codes and complexity analysis are provided in Appendix D.

Shapley Information Gain Reward Calculation. Once the Shapley values $\{\phi(f_1), \phi(f_2), \ldots, \phi(f_n)\}$ are obtained, we compute softmax-normalized weights:

$$\tilde{\phi}_i = \frac{\exp(\phi(f_i))}{\sum_{j=1}^{n} \exp(\phi(f_j))}, \tag{5}$$

Shapley Information Gain (SIG) for question q_t is:

$$\mathrm{SIG}(q_t) = \sum_{f_i \in \mathcal{F}} \tilde{\phi}_i \left[\mathbf{1}(f_i \subseteq U_t) - \mathbf{1}(f_i \subseteq U_{t-1}) \right], \tag{6}$$

This formulation captures the importance-weighted information gain induced by a question, encouraging the model to prioritize acquiring information that is both novel and clinically impactful. It can be used to guide SFT data collection and drive policy optimization via reinforcement learning.

3.3 SIG-Guided Model Initialization

This stage initializes the LLM's information-seeking policy on high-quality interactions from *SIG-Guided MCTS* via *Trajectory-based SFT*.

SIG-Guided MCTS Sampling. To construct optimal interaction trajectories, we apply MCTS (Coulom, 2006) guided by SIG, simulating a dialogue tree rooted at the initial partial information question Q_p. Each intermediate node $n_t = (q_t, r_t)$ represents a follow-up question q_t and its corresponding response r_t. Each leaf node represents a final answer A'. Node expansion is governed by a system prompt that instructs the model to ask or answer based on current information sufficiency. The SIG reward (Eq 6) guides exploration for better paths by quantifying the clinical value of each questioning node. The MCTS proceeds through the following steps (see Appendix E for details, pseudo-codes and complexity analysis):

- **Selection.** Select a path via Upper Confidence Bound for Trees (UCT) (Kocsis and Szepesvári, 2006).
- **Expansion.** From node n_{t-1}, the model either generates a follow-up question q_t and receives r_t to expand node n_t, or predicts final answer A'.
- **Simulation.** Interaction continues until termination or a depth limit, with accumulated $SIG(q_t)$ rewards and a final correctness reward for A'.
- **Backpropagation.** The total reward of the trajectory is propagated to update all visited nodes.

The overall reward for a complete interaction trajectory $\tau = \{Q_p, (q_1, r_1), \ldots, (q_T, r_T), A'\}$ is:

$$R(\tau) = \alpha \cdot \mathbf{1}(A' = A^*) + \beta \sum_{t=1}^{T} \mathrm{SIG}(q_t), \tag{7}$$

where α and β are coefficients controlling outcome and process reward. The search process is conducted on patient cases in the training data. We retain the answer-correct trajectory with the highest $R(\tau)$ for each patient case.

Trajectory-Based SFT. We fine-tune the LLM on selected high-reward trajectories to imitate the optimal information-seeking behavior, learning when to ask, what to ask, and when to answer. We supervise only the model-generated tokens, i.e., the follow-up questions $\{q_1, \ldots, q_T\}$ and the final answer A', while masking out patient responses and prompts during loss computation and gradient propagation. The loss function is:

$$\begin{aligned} \mathcal{L}_{\text{SFT}} &= \mathbb{E}_{\tau \sim \mathcal{D}_{\text{SFT}}} \left[\sum_{t=1}^{T} \mathcal{L}^{(t)}_{\text{question}} + \mathcal{L}_{\text{answer}} \right], \\ \mathcal{L}^{(t)}_{\text{question}} &= -\log P_\theta(q_t \mid Q_p, \{(q_i, r_i)\}_{i=1}^{t-1}) \\ \mathcal{L}_{\text{answer}} &= -\log P_\theta(A' \mid Q_p, \{(q_i, r_i)\}_{i=1}^{T}) \end{aligned} \tag{8}$$

3.4 SIG-Augmented Policy Optimization

This stage further enhances the model's proactive information-seeking ability via RL by extending GRPO with SIG reward. A novel *SIG-Guided Reward Distribution Mechanism* decomposes trajectory-level rewards into action-level signals, prioritizing clinically valuable questions for targeted, fine-grained optimization.

SIG-Guided Reward Distribution Mechanism. Following GRPO, for each partial information question Q_p, a group of trajectories $\mathcal{G} = \{\tau_1, \ldots, \tau_K\}$ is sampled from the current policy $\pi_{\theta_{old}}$. The trajectory-level reward $R(\tau_i)$ is computed via Eq 7, capturing both the outcome correctness and cumulative information gain.

In standard GRPO, the trajectory-level reward is used to derive the group-relative advantage by comparing the performance of trajectories within the group. This advantage is uniformly assigned to all model-generated tokens, assuming that each token, whether part of a question or the final answer, contributes equally to the outcome. While this approach captures the overall quality of the trajectory, it overlooks its internal heterogeneity: some questions may elicit more clinical information while others may be redundant or irrelevant. Consequently, such equal feedback fails to prioritize questions with higher clinical values.

To address this, we introduce the SIG-Guided Reward Distribution Mechanism, decomposing the trajectory-level reward $R(\tau)$ into action-specific rewards for each question and the final answer:

- Each follow-up question q_t receives:

$$\begin{aligned} R(q_t) &= \beta \cdot \mathrm{SIG}(q_t) + \lambda_q \cdot w_t \cdot \mathbf{1}(A' = A^*), \\ \text{where} \quad w_t &= \frac{\mathrm{SIG}(q_t)}{\sum_{j=1}^{T} \mathrm{SIG}(q_j)}. \end{aligned} \tag{9}$$

- The final answer A' receives:

$$R(A') = \lambda_a \cdot \mathbf{1}(A' = A^*). \tag{10}$$

Here, $\lambda_q + \lambda_a = \alpha$ ensures the total correctness reward is preserved and fully distributed across actions. The normalized SIG weight w_t reflects the relative contribution of each question to the final

answer. This decomposition guarantees that action-level rewards add up to the trajectory reward:

$$\sum_{t=1}^{T} R(q_t) + R(A') = R(\tau) \tag{11}$$

To provide token-level feedback, action rewards are further propagated to individual tokens. Let $\{x_1, x_2, \ldots, x_N\}$ denote the token sequence of trajectory τ. Each token x_i inherits the reward of the action that it belongs to:

$$r(x_i) = \begin{cases} R(q_i), & \text{if } x_i \in q_t \\ R(A'), & \text{if } x_i \in A' \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

This assignment ensures that actions providing more clinical utility receive proportionally higher rewards, encouraging informative questions that contribute meaningfully to the correct outcome.

Final Optimization Objective. Next, we normalize token-level rewards across the group $\mathcal{G}$ to obtain group-relative token-level advantages. Let $\mathcal{R}_\mathcal{G} = \{r(x_i) \mid x_i \in \tau_k, \tau_k \in \mathcal{G}\}$ be the set of all token rewards across the group. The advantage $\hat{A}(x)$ for each token x is:

$$\hat{A}(x_i) = \frac{r(x_i) - \text{mean}(\mathcal{R}_\mathcal{G})}{\text{std}(\mathcal{R}_\mathcal{G})}, \tag{13}$$

Finally, we apply token-level advantages to the optimization objective. Let $\hat{A}_{k,i}$ denote the advantage of token x_i in trajectory τ_k. The objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_{Q_p \sim \mathcal{D},\ \{\tau_k\}_{k=1}^{K} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\tau_k|} \sum_{i=1}^{|\tau_k|} \min\left(r_{k,i}\hat{A}_{k,i},\ \text{clip}(r_{k,i}, 1 \pm \epsilon)\hat{A}_{k,i}\right) \right] \tag{14}$$

where the importance ratio is defined as:

$$r_{k,i} = \frac{\pi_\theta(\tau_{k,i} \mid Q_p, \tau_{k,<i})}{\pi_{\theta_{\text{old}}}(\tau_{k,i} \mid Q_p, \tau_{k,<i})}. \tag{15}$$

Here, π_θ is the policy model, and $\tau_{k,<i}$ the decoding context preceding token x_i. This objective assigns fine-grained credit within trajectories, providing differentiated gradients at the token level.

Model-Aware Dynamic Rewarding. Rather than static precomputation, the atomic fact Shapley values within our SIG reward are dynamically computed during training. By recalculating these values based on the model's evolving prediction probabilities (Eq 3) at each update step, the SIG reward remains strictly model-aware. This ensures the reward accurately captures the value of information relative to the model's current state.

4 Experiments

4.1 Experimental Setup

Datasets. Experiments are conducted on two datasets derived from public multiple-choice medical benchmarks: **MedQA** (Jin et al., 2021) and **CMB** (Wang et al., 2024). Each original question is decomposed into atomic facts $\mathcal{F}$ and an atomic question Q that excludes factual information. The model input is a partial information question Q_p, consisting of Q and a subset of $\mathcal{F}$: 50% facts for CMB and the chief complaint for MedQA (following MEDIQ (Li et al., 2024)). Dataset construction and statistics can be found in Appendix G.

Baselines. We compare with diverse baselines:

- **Prompt-based. Direct** generates answers without interaction. **Vanilla** uses a system prompt to encourage questioning. **COT** adds "Let's think step by step" to promote reasoning. **MCTS-BT** uses MCTS with self-evaluated reward for inference-time scaling and selects the best trajectory, while **MCTS-MV** adopts majority voting over sampled trajectories. **MEDIQ** (Li et al., 2024) implements an abstention module. **UoT** (Hu et al., 2024) selects questions by maximizing expected entropy reduction.
- **SFT-based. DialogT** (Liu et al., 2025b) reformulates QAs as dialogues for fine-tuning. **SFT-GT** uses gold answers for supervision. **SFT** samples correct trajectories without SIG-guided MCTS.
- **SFT+RL.** RL is conducted upon ProMed (S#1) initialization for fair comparisons. **DPO** contrasts correct and incorrect trajectories sampled from the model. **GRPO** uses trajectory-level correctness reward.

Evaluation Metrics. Exact Match (EM) accuracy is reported, where options (e.g., "A", "CD") extracted from the model's final answer must be identical to the ground-truth set.

$$\textit{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Questions}} \times 100$$

Implementations. Experiments are conducted on *Instruct* LLMs: *LLaMA3.1-8B*, *LLaMA3.2-3B* (Dubey et al., 2024), and *Qwen3-1.7B* (Yang et al., 2025), with patient simulator by *Qwen2.5-72B* and SFT data sampled from DEEPSEEK-R1 (Guo et al., 2025). Training-based methods are trained on benchmark-specific training sets. Prompt-based methods are directly tested. Implementation details are in Appendix H.

Type	*LLM Turbo* Method	*Qwen3-1.7B* MedQA ↑	*Qwen3-1.7B* CMB-Exam ↑	*LLaMA3.2-3B* MedQA ↑	*LLaMA3.2-3B* CMB-Exam ↑	*LLama3.1-8B* MedQA ↑	*LLama3.1-8B* CMB-Exam ↑
Prompt	Direct	36.34±1.34	19.34±0.91	27.48±1.23	43.10±1.10	48.32±1.41	44.10±1.14
	Vanilla	31.05±1.30	29.01±1.02	30.81±1.26	33.45±1.07	40.11±1.32	44.49±1.10
	CoT	29.54±1.23	31.04±1.05	35.61±1.33	34.79±1.13	43.12±1.41	44.75±1.13
	MCTS-BT	29.77±1.29	30.28±1.05	30.24±1.32	21.50±0.95	34.41±1.35	42.95±1.15
	MCTS-MV	33.39±1.33	37.93±1.06	30.64±1.26	30.03±1.02	42.50±1.43	50.44±1.18
	MEDIQ	37.13±1.44	31.13±1.04	35.73±1.50	15.97±0.77	44.78±1.95	33.16±1.40
	UoT	36.94±1.30	43.79±0.99	36.71±1.18	40.57±1.01	36.68±1.32	44.04±1.02
SFT	DialogT	28.54±1.23	32.43±1.07	28.83±1.35	34.07±1.08	33.42±1.36	38.37±1.10
	SFT-GT	33.49±1.25	43.65±1.13	42.76±1.38	42.37±1.11	48.74±1.40	49.93±1.13
	SFT	36.85±1.36	44.30±1.10	41.83±1.35	43.60±1.13	49.39±1.43	47.15±1.12
	ProMed(S#1)	37.61±1.37	45.69±1.16	43.69±1.37	45.07±1.17	52.63±1.44	51.78±1.15
SFT+RL	ProMed(S#1)+DPO	38.06±1.37	42.42±1.09	43.44±1.46	42.87±1.14	52.80±1.40	47.87±1.14
	ProMed(S#1)+GRPO	37.62±1.39	46.61±1.22	46.32±1.37	44.76±1.31	54.60±1.37	51.43±1.13
	ProMed(S#1+2)	**39.93**±1.43	**51.98**±1.17	**47.38**±1.54	**46.25**±1.13	**55.60**±1.30	**59.33**±1.11
**Performance Gain (%)*		*+4.91*	*+11.52*	*+2.29*	*+2.62*	*+1.83*	*+14.58*

Table 1: Performance comparison (%) on **MedQA** and **CMB-Exam**. **Bold** indicates the best performance, underline the second-best. Performance gains are computed as the relative improvements over the second-best performances.

Ablation	MedQA ↑	CMB (OOD) ↑
Vanilla	40.11	44.49
w/o Stage 1	35.59	43.78
w/o Stage 2	53.26	41.44
w/o SIG	54.42	40.83
w/o Shapley	54.55	42.73
w/o Distribution	54.60	45.00
ProMed	**55.60**	**45.48**

Table 2: Ablation studies of ProMed.

4.2 Main Results

Experiment results are shown in Table 1. We summarize the key findings below:

Training for proactive questioning is essential. Direct answering without acquiring information yields poor accuracy (e.g., 19.34% on CMB, *Qwen3-1.7B*), demonstrating the necessity to address the reactive paradigm of LLMs under information insufficiency. Prompt-based methods, including advanced frameworks like UoT and MEDIQ, fail to consistently outperform direct answering (e.g., UoT on LLaMA3.2-3B MedQA). These results highlight the limitations of prompting and the importance of targeted training.

ProMed significantly outperforms existing methods and enhances LLMs' proactive information-seeking ability. ProMed consistently achieves the highest accuracy, surpassing baselines by an average relative improvement of **6.29%** over second-best results, and a striking **54.45%** gain over direct answering. This demonstrates that ProMed effectively shifts LLMs from passively reacting to proactively acquiring information, which supports its potential for practical clinical consultations.

ProMed(S#1) provides high-quality supervision via SIG-guided MCTS. Among SFT methods, ProMed(S#1) leverages SIG-guided MCTS to sample clinically valuable interactions and achieves the best performance. In contrast, DialogT constructs general multi-turn dialogues that fail to target information gaps. Standard SFT, which retains answer-correct samples without assessing question quality, also underperforms. To ensure the performance gain stems from improved questioning rather than answer memorization, we also fine-tune the model using ground-truth answers (SFT-GT) instead of sampled trajectories. Despite leveraging more data, SFT-GT still underperforms ProMed, confirming that ProMed(S#1) offers higher-quality supervision and a stronger initialization for proactive ability.

ProMed(S#2) further boosts proactive ability. Starting from the same ProMed(S#1) initialization, our SIG-Augmented Policy Optimization consistently outperforms other RL methods. While DPO and GRPO occasionally fail to improve the SFT-model, Stage 2 offers stable and significant gains, underscoring the benefit of our tailored SIG reward and reward distribution in optimizing the model's information-seeking strategy.

Supplementary experiments, including hyperparameter and Shapley analyses are in Appendix C. Runtime analysis is provided in Appendix F.

4.3 Ablation Studies

We systematically ablate key components of ProMed to validate their effectiveness. Experiments are conducted on MedQA-trained LLaMA3.1-8B, and evaluated on both in-domain and OOD settings (Table 2). The OOD setting

corresponds to a train-test distribution shift, where models trained on MedQA are evaluated on CMB-Exam, referred to as CMB (OOD).

Both stages are essential and complementary. Removing either SIG-guided Model Initialization or SIG-Augmented Policy Optimization leads to substantial performance drops. Notably, removing Stage 1 causes the largest in-domain drop, highlighting that a good initialization is crucial for avoiding poor RL convergence. Removing Stage 2 severely hurts OOD performance, confirming the importance of SIG-based policy optimization for cross-distribution generalization.

Each component in the reward design contributes to model performance. Removing the process reward SIG reduces both in-domain performance and OOD generalization. Removing the Shapley weighting or the reward distribution mechanism also leads to consistent drops, confirming the importance of modeling question utility and allocating fine-grained reward accordingly.

5 Analysis

5.1 Out-of-Domain Generalization

To verify that ProMed enhances intrinsic proactivity rather than task-specific overfitting, we conduct comprehensive OOD evaluations on CMB-trained LLaMA3.1-8B across *tasks* and *domains*:

- **Multiple-Choice Evaluation.** We evaluate zero-shot performance on the unseen MedQA.
- **Open-Ended Medical QA.** We examine whether the learned proactive behavior generalizes to open-ended tasks. **(1) MedQA without options.** We remove answer options and convert MedQA into a free-form generation task. **(2) Clinical Reasoning** CMB-Clin (Wang et al., 2024) consists of EHR-based complex clinical analysis (e.g., differential diagnosis, treatment planning), measured by BLEU-4 and ROUGE against reference expert answers.
- **General-Domain QA.** We test whether ProMed generalizes beyond healthcare on general-domain Abg-CoQA (Guo et al., 2021), assessing action-level accuracy in distinguishing between ambiguous (requiring clarification) and non-ambiguous queries.
- **Heterogeneous Clinical Data.** We include OOD tests on MIMIC-IV in Appendix C.1.

Results. As shown in Table 3, ProMed consistently outperforms all baselines across OOD medical tasks (57.50% MCQ, 25.37% Open-Ended for MedQA, 36.18 ROUGE for CMB-Clin), demonstrating robust reasoning that transfers to complex realistic clinical scenarios. On general-domain Abg-CoQA, ProMed effectively handles ambiguity, achieving the highest Macro-Acc. Its occasional over-questioning in non-ambiguous cases reflects a "cautious-proactive" strategy, prioritizing information gathering over premature answering, a desirable property for high-stakes domains. Overall, these results confirm that ProMed effectively shifts the model paradigm from reactive to proactive.

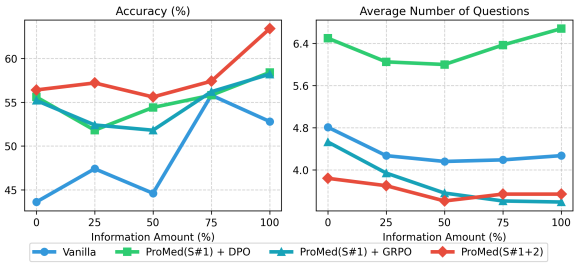


Figure 3: Model performance on CMB and question efficiency under varying initial information amounts.

5.2 Varying Initial Information Amount

To investigate models' proactive behavior across different levels of information scarcity, we vary the proportion of initial facts from 0% to 100% on 500 sampled CMB test instances and evaluate CMB-tuned LLaMA3.1-8B (Figure 3).

ProMed consistently achieves superior accuracy and questioning efficiency across all information levels, demonstrating robust generalization regardless of how much prior context is provided. In contrast, the untrained Vanilla struggles under low-information settings (0%–50%), indicating its limited questioning ability to compensate for missing evidence, while baselines initialized with ProMed Stage 1 SFT deliver better results. DPO relies on relatively more questions while ProMed Stage 2 exhibits targeted and information-efficient behavior, reaching higher accuracy with fewer questions. These results highlight the robustness and flexibility of ProMed, which successfully trains LLMs to dynamically adjust questioning strategies to gather essential information.

5.3 Robustness to Patient Simulator Variations

To assess potential bias from LLM-simulated patients and ensure robustness to simulator choice, we evaluate MedQA-trained LLaMA3.1-8B under

Method	MedQA (Acc. ↑)		CMB-Clin (Gen. ↑)		Abg-CoQA (Action Level Acc. ↑)		
	MCQ	Open-Ended	BLEU-4	ROUGE-L	Ambiguous	Non-Ambiguous	Macro-Acc
Vanilla	40.11	20.35	57.25	19.41	60.16	**48.11**	54.15
ProMed(S#1)+DPO	49.88	21.29	57.79	20.45	**93.44**	9.11	51.28
ProMed(S#1)+GRPO	51.29	21.92	50.34	31.00	77.24	30.93	54.08
ProMed(S#1+2)	**57.50**	**25.37**	**60.90**	**36.18**	79.51	34.30	**56.91**

Table 3: Results of OOD evaluations. CMB-trained LLaMA-8B is evaluated in cross-task, cross-domain settings. **Bold** indicates the best performance, and underline indicates the second-best performance.

Patient Simulator Setting	MedQA	CMB (OOD)
Backbone Replacement		
QWEN2.5-72B (default)	55.60	**45.48**
QWEN3-32B	56.23	44.35
DEEPSEEK-V3.2	56.77	44.83
GPT-5	57.23	44.14
Enhanced Simulation Mechanisms (Qwen3-32B)		
+ Fact-checker	**59.07**	45.18
+ MBTI persona	58.17	44.32

Table 4: Sensitivity analysis on patient simulators.

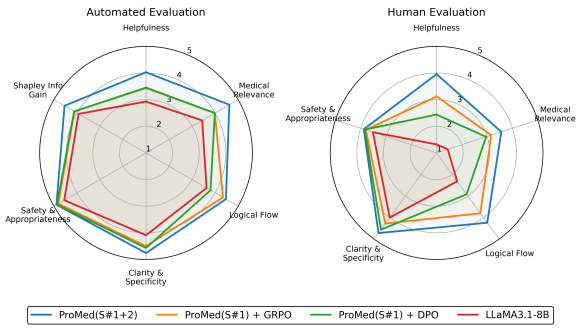


Figure 4: Question quality by both automated evaluations (left) and human clinicians (right).

diverse simulator configurations, including:

- varying the simulator's LLM backbone;
- introducing enhanced mechanisms, including a fact-checker to mitigate hallucinations and MBTI personas for diverse role-playing behaviors.

Detailed configurations and simulator quality analyses are reported in Appendix B. As shown in Table 4, performance remains remarkably stable across all simulator variants on both in-domain MedQA and OOD CMB. Adding a Fact-checker slightly improves test time performance, suggesting that higher-quality feedback during interaction further benefits the model. The results confirm that ProMed's effectiveness does not depend on a specific patient simulator design during training.

5.4 Intermediate Question Quality

Beyond final answer accuracy, we analyze the quality of model-generated questions using both LLM-based evaluators and human clinicians. Each trajectory is evaluated along five clinically motivated dimensions: **Helpfulness**, **Medical Relevance**, **Logical Flow**, **Clarity and Specificity**, and **Safety and Appropriateness** (detailed evaluation protocols in Appendix I). We additionally report the accumulated SIG reward over each trajectory as a quantitative measure of question values.

As shown in Figure 4, untrained models produce low-utility questions, especially in Helpfulness and Medical Relevance, reflecting generic, weakly-grounded questions. In contrast, ProMed-initialized and RL-optimized models show improved question quality across all dimensions. Notably, ProMed(S#1+2) achieves the best overall performance, indicating a stronger ability to identify and query missing critical information. Improvements in Logical Flow suggest that the model better conditions its questions on prior evidence, thereby reducing redundancy and enhancing coherence. Although clinicians apply stricter standards, the trends remain consistent. Overall, ProMed substantially improves the *clinical value* and *strategic quality* of proactive questions, enabling more effective and reliable consultations.

Qualitative examples and analyses of representative interaction trajectories are in Appendix J.

6 Conclusions and Future Works

We present ProMed, a novel RL framework that shifts medical LLMs from reactive to proactive for interactive medical consultation. By introducing SIG reward, ProMed quantifies question utility while accounting for interactions among clinical information. Building on this, a two-stage SIG-augmented RL pipeline enables stable initialization and fine-grained optimization via a reward distribution mechanism. Experiments show that ProMed outperforms existing methods and generalizes well to OOD settings. Future work will extend this paradigm to long-term reasoning and broader interactive decision-making tasks.

Limitations

While ProMed framework offers a promising step toward proactive medical LLMs, enabling them to actively seek information in clinical settings, several limitations remain. First, training and evaluation are conducted in simulated dialogue environments, which may not fully capture the complexity and variability of real-world patient interactions. Second, although our experiments primarily focus on two multiple-choice clinical benchmarks covering diagnosis, medication, and test recommendation, we also validate the effectiveness of our approach on several open-ended medical questions and general-domain tasks. Nevertheless, further evaluation on more diverse, complex, and real-world clinical scenarios, such as free-form treatment planning or multi-patient longitudinal cases, is necessary. Third, due to computational resource constraints, we use models up to 8B parameters; while representative, these models may not fully reveal the performance ceiling achievable with larger-scale LLMs. Finally, our current implementation operates on text-based medical facts. Incorporating multimodal clinical data, such as time series, imaging, and data from multiple comprehensive datasets, remains an important direction toward building more general and broadly applicable proactive LLMs.

Ethical considerations

This study develops a reinforcement learning framework guided by Shapley Information Gain to enhance the proactive ability of LLMs for interactive medical consultations. All experiments are conducted on public, de-identified datasets (MedQA, CMB, CMB-Clin, and Abg-CoQA) that do not contain personally identifiable information. Human evaluations are conducted under ethical approvals. While the trained model demonstrates improved proactive questioning under partial information, it is intended purely for research purposes and is not deployed in real-world clinical scenarios. The model is not designed to replace medical professionals, and any future application would require rigorous clinical validation, safety testing, and adherence to medical regulatory standards. We note that some datasets may reflect population-specific characteristics: for example, the Chinese datasets primarily represent East Asian populations, whereas the English datasets mainly reflect Western populations. Nevertheless, our proposed framework is model-agnostic and task-agnostic, and the underlying methodology is generalizable across populations and clinical contexts. We believe this work supports the responsible advancement of medical AI by addressing the risks of hallucination and unreliable responses that may arise in reactive medical LLMs when operating under incomplete patient information.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62576013, No.U23A20468), Prevention and Control of Emerging and Major Infectious Diseases-National Science and Technology Major Project (2025ZD01906000, 2025ZD01906004), the National Natural Science Foundation of China (62402017), Peking University Clinical Medicine Plus X Pilot Program-Key Technologies Project (2024YXXLHGG007), and "TengYun" Clinical Research Program (TY2025015).

Liantao Ma is supported by the Beijing Traditional Chinese Medicine Science and Technology Development Fund (BJZYZD-2025-13), the Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan (20250628).

References

Moloud Balafar, Mahboub Pouraghaei, Mahnaz Ranjkesh, Mahshid Dehghan, Ali Delkhorrami, and Samad Shams Vahdati. 2024. Comparison of the diagnostic value of ultrasound with chest ct scan in patients with unspecified pulmonary pneumonia in the emergency department. *Journal of Emergency Practice and Trauma*, 9(2):87–91.

Lynn Bickley and Peter G Szilagyi. 2012. *Bates' guide to physical examination and history-taking*. Lippincott Williams & Wilkins.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Rémi Coulom. 2006. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

Hongxin Ding, Yue Fang, Runchuan Zhu, Xinke Jiang, Jinyang Zhang, Yongxin Xu, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. 3ds: Decomposed difficulty

data selection's case study on llm medical domain adaptation. *arXiv preprint arXiv:2410.10901*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

CJ Eastmond and V Wright. 1979. The nail dystrophy of psoriatic arthritis. *Annals of the Rheumatic Diseases*, 38(3):226–228.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei W Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in llms. *Advances in Neural Information Processing Systems*, 37:24181–24215.

Saket Jha, Rajiv Ranjan Kumar, Aadhaar Dhooria, and Aman Sharma. 2019. Nail pitting: a key clinical sign of psoriatic arthritis. *Rheumatology*, 58(12):2250–2250.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025. HyKGE: A hypothesis knowledge graph enhanced RAG framework for accurate and reliable medical LLMs responses. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11856, Vienna, Austria. Association for Computational Linguistics.

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

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback.

Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer.

Suzanne Kurtz, Juliet Draper, and Jonathan Silverman. 2017. *Teaching and learning communication skills in medicine*. CRC press.

Suzanne M Kurtz and Jonathan D Silverman. 1996. The calgary—cambridge referenced observation guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. *Medical education*, 30(2):83–89.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Dong-Ho Lee, Hyundong Cho, Jonathan May, and Jay Pujara. 2025. What is a good question? utility estimation with llm-based simulations. *arXiv preprint arXiv:2502.17383*.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Weibin Liao, Xu Chu, and Yasha Wang. 2025a. Tpo: Aligning large language models with multi-branch & multi-step preference trees. In *The Thirteenth International Conference on Learning Representations*.

Weibin Liao, Xin Gao, Tianyu Jia, Rihong Qiu, Yifan Zhu, Yang Lin, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025b. Learnat: Learning nl2sql with ast-guided task decomposition for large language models. *arXiv preprint arXiv:2504.02327*.

Weibin Liao, Tianlong Wang, Yinghao Zhu, Yasha Wang, Junyi Gao, and Liantao Ma. Magical: Medical lay language generation via semantic invariance and layperson-tailored adaptation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2023. An automatic evaluation framework for multi-turn medical consultations capabilities of large language models. *arXiv preprint arXiv:2309.02077*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Zijie Liu, Xinyu Zhao, Jie Peng, Zhuangdi Zhu, Qingyu Chen, Xia Hu, and Tianlong Chen. 2025b. Dialogue is better than monologue: Instructing medical llms via strategical conversations. *arXiv preprint arXiv:2501.17860*.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. 2024. Learning to ask informative questions: Enhancing llms with preference optimization and expected information gain. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5064–5074.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jinsheng Shi, Yuyu Yuan, Ao Wang, and Meng Nie. 2025. Fine-tuning a personalized openbiollm using offline reinforcement learning. *Applied Sciences (2076-3417)*, 15(5).

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Matthew J Snyder, Marjorie Guthrie, and Stephen Cagle. 2018. Acute appendicitis: efficient diagnosis and management. *American family physician*, 98(1):25–33.

Richard S Sutton, Andrew G Barto, and 1 others. 1999. Reinforcement learning. *Journal of Cognitive Neuroscience*, 11(1):126–134.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, and 1 others. 2024. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205.

Ziyu Wang, Hao Li, Di Huang, Hye-Sung Kim, Chae-Won Shin, and Amir M Rahmani. 2025b. Healthq: Unveiling questioning capabilities of llm chains in healthcare conversations. *Smart Health*, page 100570.

Eyal Winter. 2002. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. 2025a. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*.

Yongxin Xu, Xinke Jiang, Xu Chu, Rihong Qiu, Yujie Feng, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2025b. Dearllm: Enhancing personalized healthcare via large language models-deduced feature correlations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 941–949.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, and 1 others. 2023. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Jiayuan Zhu and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. *arXiv preprint arXiv:2502.07143*.

A Related Work

LLMs for Interactive Medical Questioning. Efforts have aimed to equip LLMs for multi-turn medical consultations. MMD-eval (Liu et al., 2025a) simulates doctor-patient interactions to evaluate LLMs. MEDIQ (Li et al., 2024) designs an abstention module and UoT (Hu et al., 2024) selects best questions according to simulated entropy-reduction. Others (Liu et al., 2025b; Liao et al., 2023) fine-tune LLMs on constructed dialogues. However, these approaches rely on backbone capacity or static data, failing to truly enhance LLMs' proactive ability in dynamic dialogues. Question evaluation methods often rely on LLMs scored heuristics like usefulness or relevance (Wang et al., 2025b; Liao et al.) or leave-one-out estimations (Hu et al., 2024; Lee et al., 2025; Zhu and Wu, 2025; Mazzaccara et al., 2024), which fail to rigorously quantify questions' values in complex medical contexts. Overall, there remains a lack of training frameworks that accurately assess question and optimize for LLMs' dynamic proactive ability.

Reinforcement Learning for LLMs. Reinforcement learning (RL) (Sutton et al., 1999; Kaelbling et al., 1996) has proven effective for enhancing LLMs. RLHF (OpenAI, 2023; Kaufmann et al., 2024) aligns models via reward modeling and PPO (Schulman et al., 2017). DPO (Rafailov et al., 2023) bypasses explicit reward modeling by learning from preference pairs. GRPO (Shao et al., 2024) and DAPO (Yu et al., 2025) leverage group-wise advantages for optimization and enhance reasoning abilities. However, RL for interactive medical consultations with tailored rewards remains underexplored.

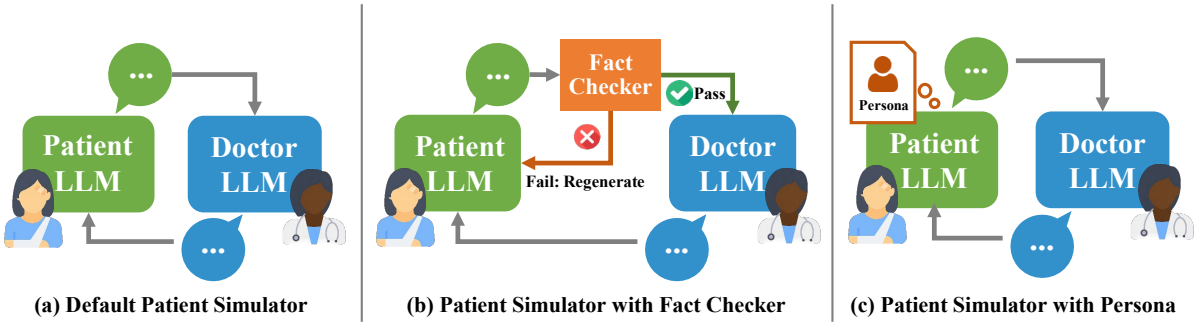
B Patient Simulator

To enable scalable training and evaluation for interactive medical consultation, we implement an LLM-based patient simulator. While such simulation reduces interaction costs, it may introduce biases stemming from the backbone model, response control mechanisms, or behavioral styles. To ensure that ProMed does not depend on a specific simulator instantiation, we systematically design and evaluate multiple patient simulator variants.

Simulator Designs. Specifically, we consider simulator designs along three dimensions: (1) the backbone LLM used to generate patient responses, which determines the linguistic capacity and medical knowledge of the simulator; (2) the interaction process, including whether question quality is explicitly validated; and (3) the behavioral variability introduced via persona modeling. Figure 5 provides an overview of the three representative simulator designs evaluated in this work.

- **Default Patient Simulator.** In the default setting, the patient simulator directly answers the doctor LLM's questions based on the complete fact set $\mathcal{F}$. Given a question, the simulator generates a response grounded in $\mathcal{F}$ whenever possible; otherwise, it replies "I don't know." Under this setting, we experiment with multiple backbone models, including *Qwen2.5-72B-Instruct*, *Qwen3-32B*, *GPT-5*, and *DeepSeek-V3.2*, to examine the impact of backbone choice.
- **Patient Simulator with Fact Checker.** To improve response reliability, we introduce a fact-checking module that explicitly evaluates each generated patient response along two dimensions: relevance, i.e., whether the response addresses the doctor's question, and factuality, i.e., whether the response is supported by the atomic fact set $\mathcal{F}$. If either criterion is violated, the response is rejected and the simulator is prompted to regenerate, with explicit feedback indicating the detected issue (e.g., irrelevance or factual inconsistency). This process is repeated until both criteria are satisfied. Such a design enforces strict quality control of the patient simulator.
- **Patient Simulator with Persona.** To model realistic variability in patient communication, we construct a persona-based simulator by conditioning the backbone LLM on randomized MBTI persona profiles. Each profile provides a structured description of the patient's personality traits, together with explicit guidance on corresponding communication styles, such as verbosity, tone, and level of certainty. The simulator is instructed to consistently adhere to the assigned persona when responding to the doctor's questions, resulting in diverse yet coherent interaction patterns (e.g., concise and cautious responses versus detailed and expressive ones), while still grounding answers in the atomic fact set $\mathcal{F}$.

Automatic Evaluation. We evaluate simulator response quality along two dimensions: **Relevance**, measuring whether the response meaningfully addresses the doctor's question, and **Factuality**, measuring whether the response is supported by the atomic facts. Both criteria are assessed using an LLM-based evaluator (based on DeepSeek-



(a) Default Patient Simulator (b) Patient Simulator with Fact Checker (c) Patient Simulator with Persona

Figure 5: Illustration of patient simulator designs. **(a) Default**: the simulator directly responds to questions. **(b) Fact checker**: a fact checker module triggers regeneration of responses to ensure factuality and relevance. **(c) Persona**: the simulator is conditioned on randomized MBTI persona traits to induce diverse communication behaviors.

Patient Simulator Setting	Relevance (%)	Factuality (%)
Backbone Replacement (Default Simulator)		
QWEN2.5-72B	**96.80**	92.40
QWEN3-32B	95.60	92.40
DEEPSEEK-V3.2	93.60	95.60
GPT-5	96.40	96.80
Enhanced Simulation Mechanisms		
QWEN3-32B + Fact-checker	94.40	**97.20**
QWEN3-32B + MBTI persona	95.20	94.40
Human Verification Result		
QWEN2.5-72B	96.00	92.00

Table 5: Quality evaluation of patient simulator configurations. Relevance measures whether the simulated patient responses address the doctor's questions, while factuality measures consistency with the provided facts.

V3.2 (DeepSeek-AI, 2025)) with binary (Yes/No) judgments. For each patient simulator configuration, we randomly sample 250 interactions from both the CMB and MedQA benchmarks, resulting in 500 evaluated interactions per simulator. Table 5 summarizes the relevance and factuality accuracy across different simulator configurations. Overall, LLM-based patient simulators demonstrate consistently high relevance and factuality, making them reliable participants in interactive medical consultations for model training and evaluation. Fact checker improves factuality at a small cost to relevance, while persona modeling introduces mild behavioral variability without degrading overall response quality. For simplicity, we implement default patient simulator using Qwen2.5-72B, which achieves a strong balance between response quality and computational cost.

Human Verification. To further validate the reliability of our chosen simulator, we manually annotate a small subset (50 samples) of interactions generated by the Qwen2.5-72B-based patient. The human evaluation closely matches the automatic assessment, confirming that the simulator provides coherent, relevant, and factually grounded responses suitable for interactive medical consultation experiments.

Model	Outcome Prediction (%)	Readmission Prediction (%)
Vanilla	56.90	41.60
ProMed(S#1)+GRPO	59.90	50.40
ProMed(S#1+2)	**69.50**	**76.20**

Table 6: Performance on MIMIC-IV prediction tasks.

C Supplementary Experiments

C.1 Experiments on MIMIC-IV

While our primary evaluation utilizes natural-language dialogue to simulate the most common mode of doctor–patient interaction, real-world clinical intelligence must handle heterogeneous data sources. To further validate this capability, we conduct additional experiments on the MIMIC-IV (Johnson et al., 2023) dataset, demonstrating that our model can proactively acquire and distinguish critical information from uncurated, noisy, and mixed structured–unstructured medical data.

Data. We randomly sample 1000 patient records from MIMIC-IV. Each record contains demographic information (e.g., age and gender) and multiple hospital admissions. Each admission includes both structured laboratory measurements (numerical values) and unstructured clinical notes.

Information Sufficiency Setting. To simulate partial observability: The doctor LLM initially observes patient demographics, 50% randomly sampled laboratory tests, and the first 500 characters of clinical notes. Patient Simulator has access to the

full patient record. Neither side has access to the ground-truth prediction target, preventing information leakage and the model trivially asking for the answer.

Tasks. We evaluate two clinically relevant prediction tasks:

- **Mortality outcome prediction**: Predict patient outcome (mortality) from longitudinal visit sequences.
- **Readmission prediction**: Predict whether a patient will have a subsequent admission based on a single visit.

Structured data are serialized into textual form with explicit column names (e.g., *Heart Rate: 90.0*) and concatenated with clinical notes as model input.

Models. LLaMA-3.1-8B-Instruct is the vanilla backbone model. ProMed(S#1+GRPO) is the GRPO baseline trained with answer-correctness rewards. ProMed(S#1+2) is trained with our framework. Training is conducted on the CMB dataset.

Results. As shown in table 6, ProMed significantly outperforms all baselines on both tasks. The untrained backbone model performs near or below random guessing, highlighting the difficulty of reasoning over noisy and heterogeneous clinical data. The GRPO-trained model shows modest improvement, reflecting slightly enhanced medical abilities; however, because its RL objective optimizes only final answer correctness, it fails to consistently acquire critical intermediate information in this OOD setting. In contrast, ProMed explicitly rewards high-value questions, which enables the model to actively acquire and extract task-relevant signals from mixed structured and unstructured data.

Case Study Analysis. Further qualitative analysis (see Appendix J for details) reveals:

- In the **outcome prediction** case: The GRPO model is misled by superficially positive signals (e.g., stable vitals). In contrast, ProMed retrieves high-impact factors such as respiratory failure, multimorbidity, and functional decline, leading to the correct prediction.
- In the **readmission prediction** case: The GRPO model focuses narrowly on ongoing medical issues, resulting in an incorrect prediction. In contrast, ProMed performs more comprehensive information retrieval across clinical and care-related dimensions (e.g., functional status, caregiver support, hospice care), enabling the correct decision.

Dataset	Model	Prompt	Accuracy (%)	Avg. #Questions
CMB	LLaMA3-8B	Original	**44.49**	4.13
CMB	LLaMA3-8B	SOP-guided	43.30	4.48
CMB	ProMed	Original	**59.33**	4.40
CMB	ProMed	SOP-guided	52.75	3.67
MedQA	LLaMA3-8B	Original	40.11	4.04
MedQA	LLaMA3-8B	SOP-guided	**47.10**	5.66
MedQA	ProMed	Original	**57.50**	3.70
MedQA	ProMed	SOP-guided	52.42	5.74

Table 7: Impact of SOP-guided questioning on accuracy and efficiency.

Summary. These findings suggest that our training strategy equips the model with the ability to handle noisy, partially observed, and heterogeneous medical data, which is essential for real-world clinical deployment.

C.2 Clinical SOP-Guided Questioning

Question ordering plays an important role in clinical reasoning, potentially affecting both diagnostic efficiency and accuracy. To systematically study this factor, we conduct experiments incorporating an explicit clinical questioning protocol into the model prompting.

Clinical Inquiry Protocol. We introduce a **Clinical Inquiry Protocol** into the system prompt, requiring the doctor model to follow the Calgary–Cambridge Framework (Kurtz and Silverman, 1996; Kurtz et al., 2017), a widely adopted standard for medical interviewing. The protocol enforces a structured questioning order: (1) symptom clarification, following the OPQRST scheme (Bickley and Szilagyi, 2012), (2) past medical history, (3) social and family history, and (4) review of systems.

Experimental Setup. We compare our original prompting and SOP-guided prompting across the backbone model and our trained ProMed model, evaluated on CMB (in-domain) and MedQA (OOD). We report final accuracy and average number of questions asked (as a proxy for efficiency).

Results. Model- and context-dependent effectiveness. The impact of SOP guidance varies across models and datasets. On the CMB dataset, enforcing the protocol degrades performance for both models. On MedQA, improvements are observed only for the untrained backbone model, while ProMed experiences a performance drop.

Efficiency trade-offs. SOP guidance generally increases the number of questions asked (in three out of four settings), indicating reduced efficiency.

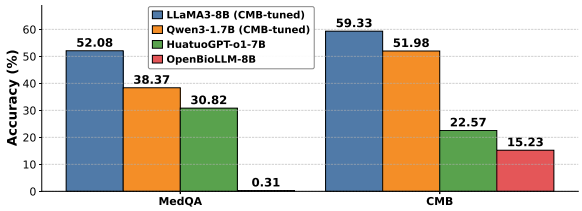


Figure 6: Performance comparison of ProMed with various SOTA medical LLMs.

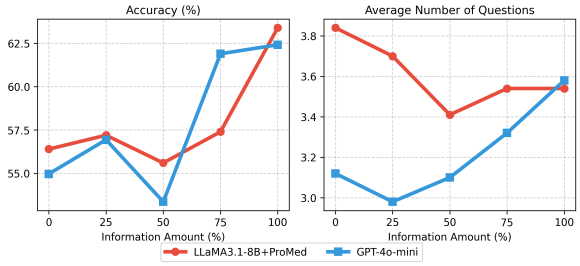


Figure 7: Comparison with GPT-4o-mini.

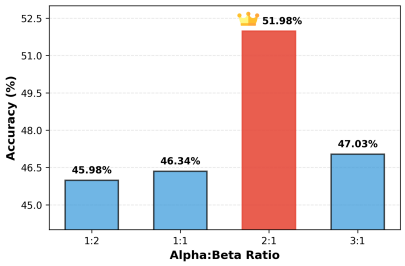


Figure 8: Sensitivity analysis of the reward weights α and β on the CMB dataset using Qwen3-1.7B.

Even when accuracy improves (e.g., LLaMA3-8B on MedQA), this gain comes at the cost of additional queries.

Interference with learned policies. For the ProMed model, which has already learned effective information-seeking strategies, imposing a rigid protocol consistently harms performance. This suggests a mismatch between fixed procedural constraints and the model's learned questioning policy.

Summary. These findings indicate that while structured questioning is clinically meaningful, a single fixed SOP does not consistently improve performance across tasks or domains. In practice, effective questioning strategies are often domain-specific and adaptive. Rigid protocols may therefore limit the model's ability to optimize information acquisition. In contrast, approaches that prioritize information gain provide a more flexible and robust foundation for clinical reasoning.

C.3 Comparison with Medical LLMs

To further validate the necessity of optimizing proactive information-seeking, we compare ProMed-optimized models (LLaMA3-8B and Qwen3-1.7B, trained on CMB) with open-source medical LLMs *HuatuoGPT-o1-7B* (Chen et al., 2024) and *OpenBioLLM-8B* (Shi et al., 2025), as shown in Figure 6. **ProMed enables strong interactive reasoning beyond medical training.** Despite being of comparable or smaller scales, ProMed models outperform existing medical LLMs on both benchmarks by large margins, demonstrating that medical pre-training and SFT alone does not endow LLMs with robust interactive diagnostic abilities. These findings confirm that targeted training is essential for enabling clinically valuable interaction.

C.4 Comparison with proprietary LLMs

To further evaluate the effectiveness of ProMed, we compare our fine-tuned LLaMA3.1-8B against the proprietary GPT-4o-mini, across varying levels of initial information availability. As shown in Figure 7, our model consistently matches or outperforms GPT-4o-mini in accuracy. This suggests that ProMed training successfully empowers an open-sourced model to reach, and even exceed, the reasoning capabilities of proprietary counterparts in specialized medical domains. The distinction also lies in the models' adaptive questioning behavior:

- **ProMed Efficiency:** Our model exhibits a highly adaptive inquiry strategy. It intelligently reduces the number of questions as the initial information becomes more sufficient (from 0% to 50%).
- **GPT-4o-mini Redundancy:** In contrast, GPT-4o-mini shows an inefficient trend; its average number of questions actually *increases* with more available information. This indicates a lack of awareness regarding information sufficiency, leading to redundant queries even when the evidence is already conclusive.

These results underscore that ProMed instills a "self-awareness" of information gaps. By precisely quantifying the utility of questions through SIG rewards, our model learns to prioritize high-gain inquiries and refrain from redundant dialogue, achieving a more professional and human-like clinical reasoning process than general-purpose proprietary models.

C.5 Hyperparameter Sensitivity Analysis

To investigate the impact of the reward components on model performance, we conduct a sensitivity analysis on the weighting coefficients α (answer outcome reward) and β (question process reward). We evaluate four different ratios: $\alpha : \beta \in \{1:1, 2:1, 3:1, 1:2\}$ by training Qwen3-1.7B on CMB. As illustrated in Figure 8, Our primary configuration ($\alpha : \beta = 2:1$) achieves the highest accuracy of 51.98%. When the outcome reward is further emphasized, performance decreases, indicating that overly prioritizing answer correctness can weaken the model's incentive to ask informative questions during the interaction process. On the other hand, increasing question process reward weight also degrades performance, implying that excessive focus on question quality without sufficient guidance from outcome supervision may lead to suboptimal diagnostic decisions. Overall, these results highlight the importance of a balanced reward design. A moderate emphasis on the outcome reward, while still preserving a strong signal for question quality, yields the best performance. This observation supports our design choice of $\alpha : \beta = 2:1$ and demonstrates that effective interactive medical reasoning requires both accurate final predictions and high-quality information-seeking behavior.

C.6 Information Shapley Analysis

To further validate the effectiveness of Shapley values in capturing clinically meaningful information, we design a noise-injection experiment. Specifically, we randomly sample 100 patient cases in MedQA and inject irrelevant facts, such as "the patient's zodiac sign" or "the patient's hair color" that are unrelated to clinical outcomes. We then compute the importance scores of all atomic facts using two methods: Leave-One-Out (LOO) and our proposed atomic fact Shapley. In LOO, the importance of each fact is measured individually by adding it to the input and observing its impact on the model's probability to predict the correct answer. For each method, we rank all facts by their estimated importance and calculate Recall@K, which measures the proportion of truly relevant medical facts appearing in the top-K positions.

As shown in Table 8, Shapley values consistently outperform LOO across all K. Notably, Shapley achieves a Recall@1 of 95.96% and Recall@3 of 90.24%, compared to LOO's 80.16% and 76.06%, respectively. These results demonstrate that Shapley values provide a finer-grained and more accurate reflection of clinical relevance than LOO, enabling more reliable identification of medically salient information. This superior sensitivity to fact-level importance supports the use of Shapley values as the foundation of our reward design in interactive medical questioning.

Method	Recall@1	Recall@3	Recall@10
Shapley	**0.9596**	**0.9024**	**0.9058**
LOO	0.8016	0.7606	0.8275

Table 8: Comparison of Recall@K between Shapley Value and Leave-One-Out (LOO) under noise injection setting. Higher values indicate better ability to identify clinically relevant facts.

Tolerance (ϵ)	Average Time (s)	Std Dev
0.10	0.34	0.25
0.01	1.03	0.82
0.001	3.69	2.91

Table 9: Impact of Convergence Tolerance (ϵ) on Computation Time.

D Monte Carlo Shapley

D.1 Algorithm Details

To mitigate the computational overhead of Monte Carlo Shapley calculation, we implement a batched inference strategy leveraging the vLLM (Kwon et al., 2023) framework. For each sampled permutation π_k, we construct a sequence of n prefixes, $\{S_1, \ldots, S_n\}$, representing the cumulative addition of facts. Rather than performing n independent sequential passes, we pack these prefixes into a single inference batch. Specifically, we utilize the prompt logprobs feature of the inference engine to extract the log-probabilities of the target answer A conditioned on each prefix. This batching approach reduces the number of model forward passes from $O(K \times n)$ to $O(K)$, where K is the number of MC iterations and n is the number of atomic facts.

In Algorithm 1, we provide the pseudo-codes of the batched Monte Carlo approximation algorithm for calculating the atomic fact Shapley in Section 3.2.

D.2 Complexity Analysis

We analyze the SIG reward's efficiency, focusing on how the computational cost scales with information volume (number of atomic facts), precision

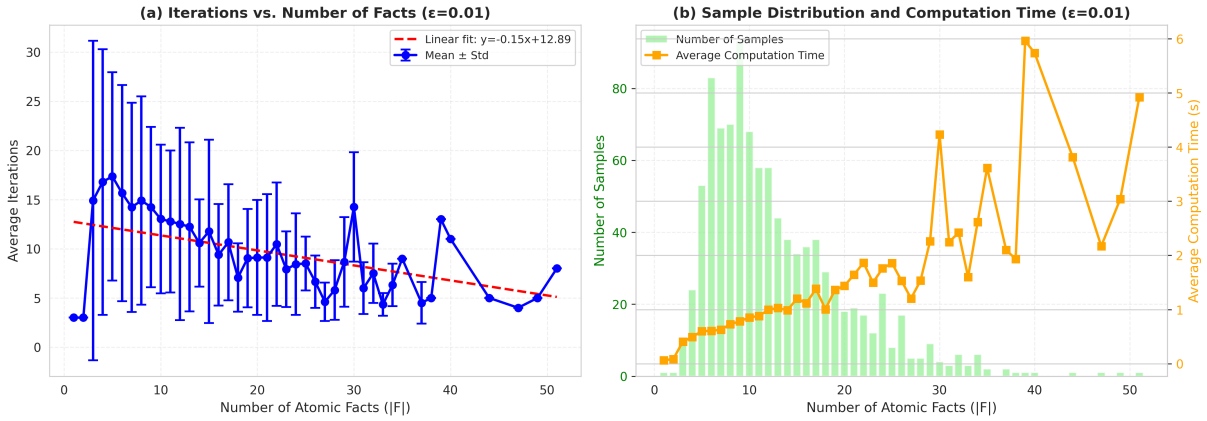


Figure 9: Efficiency analysis of Atomic Fact Shapley calculation. (a) Average iterations required for convergence vs. number of facts; (b) Distribution of samples and average computation time per fact set size.

Model Size (B)	Average Time (s)	Std Dev
1.7	1.03	0.82
3.0	6.27	5.18
8.0	10.47	8.84

Table 10: Impact of Model Scale on Time Complexity.

requirements (convergence tolerance), and model capacity. To ensure a representative evaluation, we randomly sampled 500 instances each from the training sets of CMB and MedQA for the tests.

Impact of the Number of Atomic Facts ($\mathcal{F}$). As illustrated in Figure 9, the computational cost of the SIG reward remains efficient even as the number of atomic facts increases. While the theoretical search space for Shapley values is $2^{|\mathcal{F}|}$, our batch-processing MC estimation reduces the empirical complexity. Specifically, Fig. (a) shows that the number of iterations required for convergence stays relatively stable (between 5 to 15 iterations) and even exhibits a slight downward trend as $|\mathcal{F}|$ increases. This suggests that in larger medical evidence sets, the marginal contributions often converge quickly. Fig. (b) confirms that the average computation time scales linearly with $|\mathcal{F}|$, typically remaining under 2 seconds. The spikes in computation time at higher $|\mathcal{F}|$ values are largely due to the increased prompt length and the linear growth of prefix evaluations within each batch. Moreover, as shown by the sample distribution in Figure (b), the number of facts per query rarely exceeds 30 in both CMB and MedQA datasets. This implies that high-latency cases are infrequent in practice, with most estimations concluding in under 2 seconds.

Impact of Convergence Tolerance (ϵ). The choice of tolerance ϵ presents a direct trade-off between estimation precision and latency. As shown in Table 9, setting $\epsilon = 0.00$ (full iteration) results in an average latency of 3.69s. However, by relaxing ϵ to 0.01, we achieve a 3.6× speedup (1.03s) with negligible impact on reward quality. A further increase to $\epsilon = 0.10$ reduces time to 0.34s, making it suitable for real-time interactive scenarios.

Impact of Model Scale. Table 10 details the impact of the model's parameter size on efficiency. While the 1.7B model is highly efficient (1.03s), moving to 8B parameters increases the latency. However, this growth remains near-linear relative to the model size. This scaling confirms that the time complexity is predictable and manageable, further underscoring the necessity of our batched inference strategy, which ensures that larger, more capable models can be integrated without incurring prohibitive, non-linear delays.

Overall, the results demonstrate that our batched Monte Carlo estimation of fact Shapley maintains high practical scalability. The detailed runtime analysis of integrating this reward into the overall framework is provided in Appendix F.

E SIG-Guided MCTS Sampling

E.1 Algorithm Details

To construct high-quality SFT interaction trajectories, we employ the SIG-Guided MCTS (Coulom, 2006) algorithm. Below, we provide detailed descriptions of the sampling process. Pseudo-codes for this process is provided in Algorithm 2.

Each MCTS run simulates a doctor-patient dialogue tree rooted at the initial clinical inquiry Q_p, where nodes represent interaction states. The search space is defined by the model's question

Algorithm 1 Batched Monte Carlo Fact Shapley

Require: Model M, Atomic Facts $\mathcal{F} = \{f_1, \ldots, f_n\}$, Question Q, Answer A, Max iterations K, Tolerance ϵ
Ensure: Estimated Shapley values $\{\phi(f_i)\}_{i=1}^n$

```
Initialize φ(f_i) ← 0 for all i
Compute v(S_0) ← log P_M(A | ∅, Q)
for k = 1 to K do
    Sample permutation π_k of {1, ..., n}
    Construct fact subsets S_j = {f_π(1), ..., f_π(j)}
    Batch compute {v(S_1), ..., v(S_n)} via parallel inference
    v_prev ← v(S_0)
    for j = 1 to n do
        i ← π(j)
        Δ ← v(S_j) − v_prev
        φ(f_i) ← ((k−1)/k) φ(f_i) + (1/k) Δ
        v_prev ← v(S_j)
        if |v(S_j) − v_F| < ε then
            break
        end if
    end for
    if (1/(k−1)) Σ_i |φ_i^(k) − φ_i^(k−1)| < ε then
        break    ▷ Convergence of estimation
    end if
end for
return {φ(f_i)}_{i=1}^n
```

generation distribution $\mathcal{M}(\cdot \mid \mathcal{H}_t)$, where $\mathcal{H}_t = \{(q_1, r_1), ..., (q_t, r_t)\}$ is the dialogue history up to step t. The MCTS search follows the process:

- **Selection.** Starting from the root node n_0, the algorithm selects a child node n' at each step that maximizes the Upper Confidence Bound for Trees (UCT) (Kocsis and Szepesvári, 2006):

$$\text{UCT}(n') = \bar{R}(n') + c \cdot \sqrt{\frac{\log N(n)}{N(n') + \epsilon}}, \quad (16)$$

  where $\bar{R}(n')$ is the average total reward of node n', $N(n)$ is the number of visits to parent node n, and c is an exploration coefficient. This process continues recursively until a leaf or unexpanded node is reached.
- **Expansion.** Given selected node n_{t-1}, the model decides to either:
  (a) Generate a follow-up question q_t, receive a response r_t from a simulated patient, and form new node $n_t = (q_t, r_t)$, or
  (b) Issue a final answer A' and terminate.

  If expanded, we update:

$$\mathcal{H}_t = \mathcal{H}_{t-1} \cup \{(q_t, r_t)\}, \quad (17)$$
$$U_t = \mathcal{M}_{\text{Understand}}(Q_p, \mathcal{H}_t), \quad (18)$$
$$R_{\text{local}}(q_t) = \text{SIG}(q_t). \quad (19)$$

- **Simulation.** The interaction proceeds recursively until the model issues a final answer A' or reaches a depth limit T_{max}. The trajectory reward is calculated as in Eq 7 in Section 3.3.
- **Backpropagation.** The final reward $R(\tau)$ is propagated to all nodes n along the selected path:

$$N(n) \leftarrow N(n) + 1, \quad (20)$$
$$W(n) \leftarrow W(n) + R(\tau), \quad (21)$$
$$\bar{R}(n) \leftarrow \frac{W(n)}{N(n)}. \quad (22)$$

E.2 Complexity Analysis

We analyze the computational complexity of the above algorithm and the practical strategies used to ensure efficient offline data generation.

Theoretical Complexity. The computational cost of Algorithm 2 is primarily determined by the number of simulations N and the maximum search depth T_{max}. For each simulation, the complexity is $O(N \cdot T_{\text{max}} \cdot C_{\text{LLM}})$, where N is the number of simulations, T_{max} is the maximum dialogue depth, and C_{LLM} represents the cost of model generation.

Practical Efficiency and Implementation Strategy. In practice, we implement several strategies to ensure efficiency. We adopt a *greedy-first* approach: MCTS is prioritized for challenging cases where initial greedy sampling fails to produce a correct and efficient trajectory. Both sampling processes are accelerated using the vLLM inference engine combined with asynchronous multi-threading. By utilizing asynchronous calls, we can concurrently manage multiple dialogue states and model requests, maximizing GPU utilization. The detailed runtime analysis of MCTS implementation in the overall framework is provided in Appendix F.

F Framework Runtime Analysis

Our framework comprises several stages, including MCTS sampling, SFT and RL. The MCTS sampling and SIG reward calculation introduce additional computational overhead beyond traditional training. We provide detailed runtime breakdowns for each stage of our method, along with comparisons against alternative approaches.

Method	Sampling Time	Avg/Sample	SFT Time	Correct Samples	SFT Acc. (%)
Single Sample	3h55m	0.91s	41m	11,227	49.58
Best-of-5	9h16m	4.54s	47m	13,422	49.77
DialogT	**1h37m**	**0.38s**	**28m**	N/A	38.37
MCTS (ProMed S#1)	13h12m	7.88s	54m	**15,420**	**51.78**

Table 11: Comparison of SFT data construction strategies.

Method	Training Time	Acc. (%)
DPO	30h16m + 45m	47.87
GRPO	**8h39m**	51.43
ProMed S#2	12h45m	**59.33**

Table 12: Comparison of RL training strategies.

F.1 SFT Stage

We compare multiple data construction strategies under the same experimental setup: CMB-Exam (15,465 samples), DeepSeek-R1 API (concurrency=80), and LLaMA3.1-8B-Instruct training on 4×A800 GPUs for one epoch.

As shown in Table 11, MCTS incurs higher sampling cost than Best-of-N due to structured exploration and Shapley computation. However, it produces nearly all-correct trajectories, substantially improving supervision quality and downstream accuracy. Under the same rollout budget, MCTS is significantly more effective than Best-of-N.

F.2 RL Stage

We further compare different RL fine-tuning strategies using LLaMA3-8B on 4×A800 GPUs.

As shown in Table 12, ProMed Stage 2 introduces additional cost over GRPO due to SIG reward calculation, resulting in approximately 4 additional hours of training. This overhead yields a substantial **+7.9% absolute improvement** in accuracy. In contrast, DPO is significantly more expensive while underperforming both GRPO and ProMed.

Summary. Overall, our framework introduces additional **offline training cost** in exchange for significantly improved supervision quality and model performance. After training, the deployed model has **no additional inference-time overhead** compared to standard baselines. This design follows a common paradigm in foundation model training: trading additional offline computation for improved and reusable model capability. The cost–performance trade-off is suitable for medical applications, where reliability and effective information acquisition are critical.

G Interactive Medical Questioning Dataset

Below we detail the construction, statistic and usage of datasets. Datasets are utilized in accordance with their respective licenses and intended use.

G.1 Original Datasets

Experiments are conducted on large-scale medical multiple-choice benchmarks: **MedQA** (Jin et al., 2021) and **CMB** (Wang et al., 2024). Datasets are publicly available and fully anonymized, containing no personally identifiable information.

MedQA is a multilingual benchmark derived from real and mock United States Medical Licensing Examination (USMLE) questions, covering diagnostic reasoning and clinical problem-solving. It includes over 60K questions across English, Simplified Chinese, and Traditional Chinese. **In this work, we use only the English subset**, which contains approximately 12.7K questions, each grounded in patient-specific cases.

CMB (Chinese Medical Benchmark) is a comprehensive Chinese benchmark featuring over 280K multiple-choice questions across six clinical domains and 28 subcategories. Unlike MedQA, not all questions are grounded in patient cases. We therefore apply a filtering strategy (described below) to extract the case-based subset suitable for interactive patient-doctor simulations.

These benchmarks are selected due to their scale, clinical coverage, and diversity in question types, which make them well-suited for evaluating interactive medical reasoning under partial information.

G.2 Dataset Construction Process

To build datasets for interactive doctor–patient questioning, we apply the following pipeline:

- For **MedQA**, since questions are already constructed around specific patient cases, we retain all items as they naturally match the interactive scenario.
- For **CMB**, we filter the CMB-Exam questions using the *Judge Patient Prompt* in Appendix K,

Dataset	Split	# Questions	Avg. Atomic Facts
MedQA	Train	10178	15.92
	Val	1272	9.25
	Test	1273	9.54
	Total	12723	14.58
CMB	Train	15465	8.89
	Val	1940	8.86
	Test	1935	8.82
	Total	19340	8.88

Table 13: Dataset statistics for MedQA and CMB.

keeping only those questions judged as based on patient cases.

- Next, for all retained questions from both datasets, we apply the *Atomic Fact Decomposition Prompt* in Appendix K to break down the full question stem into a set of atomic facts—each a minimal, self-contained piece of patient information.
- We then construct partial-information inputs: for MedQA, we feed only the patient's chief complaint as the partial input; for CMB, we randomly sample about half of the atomic facts as the partial context.

All prompt-based processing steps above are executed using `Qwen2.5-32B-Instruct`, ensuring high-quality and medically consistent outputs.

G.3 Dataset Statistics

In Table 13, we report the number of questions and the average number of atomic facts per question in each split. For **MedQA**, we reuse the development and test splits from prior work (MEDIQ (Li et al., 2024)) for fair comparisons, while the training set is newly processed in this study. For **CMB**, we perform full-scale filtering and processing, and then randomly split the resulting examples into training, validation, and test sets.

Dataset Examples. To help understand the structure of our interactive medical questioning task, Table 14 provides representative examples from MedQA and CMB. Each example includes a partial information question that simulates the limited patient information initially available to the doctor, the corresponding full set of atomic facts decomposed from the original question stem, and the final answer. These examples demonstrate the clinical richness and granularity of our dataset construction, as well as the challenges under partial information.

G.4 OOD Evaluation Datasets

For OOD evaluation, we use two additional datasets with distinct characteristics. The **CMB-Clin** (Wang et al., 2024) dataset contains 208 multiple-choice questions covering various clinical reasoning scenarios, with an average of 40.31 atomic facts per question. The **Abg-CoQA** (Guo et al., 2021) dataset consists of 1,055 question-answer pairs, of which 123 questions (11.66%) are ambiguous and 932 questions (88.34%) are non-ambiguous. These datasets allow us to assess the robustness and generalization ability of the models in settings that differ from the training distribution.

H Implementation Details

All experiments are conducted on a Ubuntu 20.04 server equipped with two NVIDIA A800 GPUs. We implement our framework using Python 3.10 and PyTorch. All codes used in this work are utilized in accordance with their respective licenses and intended use.

H.1 MCTS Configuration

During data sampling, we set the outcome-level and question-level reward weights to $\alpha = 2$ and $\beta = 1$, respectively. The MCTS is configured with an exploration weight of 2.2, maximum width of 8, number of iterations set to 5, and a maximum search depth of 10.

H.2 Training Configuration

We adopt LoRA for all model training stages, with the LoRA rank set to 8. The SFT and DPO stages are trained for 1 epoch with a batch size of 64, using learning rates of 5×10^{-5} for Qwen3 models and 1×10^{-4} for LLaMA models (SFT), and 5×10^{-6} for DPO. These two stages are implemented using the `LLaMAFactory` framework(Zheng et al., 2024).

For GRPO, we set the outcome-level and question-level reward weights to $\alpha = 4$ and $\beta = 2$. In outcome reward distribution, the reward is allocated to the answer and questions with weights $\lambda_a = 3$ and $\lambda_q = 1$, respectively. Weights are selected based on validation performance. The GRPO stage is trained on our developed training framework for 200 steps, with a batch size of 1 and 4 rollouts per case.

H.3 Baselines

For existing baselines, we follow the best practices reported in their papers to ensure fair comparisons.

- **MEDIQ** (Li et al., 2024): We use the official implementation from the public GitHub repository[1]. We follow their prompt framework, using an abstention module prompting the model to decide if there is enough evidence. We replace the system prompt with our doctor model system prompt.
- **UoT** (Hu et al., 2024): We use their publicly available code[2] and adapt the prompts to our task setting.
- **DialogT** (Liu et al., 2025b) reformulates QA pairs as dialogues and fine-tunes the model. For our experiments, we strictly follow their prompt templates to construct multi-turn dialogue data from our task datasets and fine-tune the model accordingly.

H.4 Evaluation Protocol

We adopt accuracy as the primary evaluation metric to objectively assess the correctness of final answers. During evaluation, we report the mean and standard deviation of the accuracy metric via bootstrap resampling over prediction outputs, following standard practices to measure performance variability. Metric details are as follows:

- **Exact Match (EM) for Multiple-Choice Questions.** To evaluate LLMs on multi-task medical multiple-choice questions, we instruct models to provide only the correct answer and adopt the widely used **Exact Match (EM)** metric (Jiang et al., 2025; Ding et al., 2024). An answer is considered correct if it exactly matches all entries in the ground truth. EM is computed as:

$$EM = \frac{\text{Number of Correct Answers}}{\text{Total Number of Answers}} \times 100\%.$$

  In our main experiments, EM is used as the primary evaluation metric for all multiple-choice tasks.
- **Open-ended Medical QA Accuracy.** For open-ended MedQA tasks, we assess answer correctness by comparing the model-generated response against the reference using an LLM judge, which evaluates whether the response satisfies the standard answer.
- **BLEU-4 and ROUGE-R for CMB-Clin** We employ BLEU-4 (Papineni et al., 2002) and ROUGE-R (Lin, 2004) to quantify model response quality and coverage compared to expert reference answers. **BLEU-4** measures the 4-gram precision of generated answers and captures fluency through higher-order n-gram consistency:

$$\text{BLEU-N} = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N}\log p_n\right),$$

  where p_n is the precision of n-grams and BP is the brevity penalty:

$$BP = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \leq r \end{cases},$$

  with c and r denoting the lengths of the generated and reference responses, respectively.
  ROUGE-R emphasizes recall of relevant content, measuring how comprehensively the generated response covers the reference:

$$\text{ROUGE-R} = \frac{|R \cap G|}{|G|},$$

  where $|R \cap G|$ is the number of overlapping n-grams between the generated response R and reference G, and $|G|$ is the total n-grams in G.
  We compute ROUGE scores using the `rouge` package and BLEU scores (BLEU-4) using the `nltk` module, applying smoothing for BLEU and default settings for ROUGE.
- **Action-Level Accuracy for Abg-CoQA.** For Abg-CoQA, we measure **action-level accuracy**. If a question is ambiguous, a model-generated clarifying question is considered correct; if the question is non-ambiguous, directly providing the correct answer is counted as correct.

I Intermediate Question Evaluation Protocol

In main experiments, we primarily evaluate model performance using accuracy on MCQ benchmarks, and the correctness of the final answer in open-ended settings. Such outcome-oriented metrics are insufficient for assessing *interactive medical consultation*, where the quality of intermediate questions posed by the model plays a critical role in information acquisition and clinical reasoning. To this end, we conduct a complementary evaluation focusing on the **quality of model-generated questions during the consultation process**. Specifically, we perform both **automated evaluation** using LLMs as evaluators and a **human evaluation** conducted by licensed clinicians.

[1] https://github.com/stellalisy/mediQ
[2] https://github.com/zhiyuanhubj/UoT

I.1 Evaluation Criteria.

Each model-generated interaction trajectory is evaluated along the following five dimensions, using a **5-point Likert scale** (1 = very poor, 5 = excellent):

- **Helpfulness.** Whether the question is likely to elicit information that is useful for reaching a correct and confident diagnosis or clinical decision.
- **Medical Relevance.** Whether the question aligns with standard clinical reasoning and focuses on medically significant factors related to the patient's condition.
- **Logical Flow.** Whether the sequence of questions follows a coherent and natural diagnostic progression, rather than being disjointed, redundant, or repetitive.
- **Clarity and Specificity.** Whether the question is clearly phrased, unambiguous, and sufficiently specific for a patient to understand and answer accurately.
- **Safety and Appropriateness.** Whether the question is appropriate for the given consultation context and aligns with standard clinical interviewing practices (e.g., avoiding premature conclusions or unnecessary invasiveness).

For automated evaluation, we additionally compute the **total Shapley Information Gain (SIG) reward** accumulated over the entire interaction trajectory, which quantitatively measures the clinical value of information acquired through questioning.

I.2 Automated Evaluation

For automated evaluation, we employ **DeepSeek-V3.2** as an LLM-based evaluator to assess the quality of model-generated interaction trajectories. Each trajectory, including the patient context and the sequence of intermediate questions, is provided to the evaluator along with detailed scoring instructions corresponding to the five evaluation dimensions described above. The evaluator is prompted to assign a score from 1 to 5 for each dimension. We also compute the total Shapley Information Gain reward of the trajectory. The evaluation prompt is provided in Appendix K.5

I.3 Human Evaluation Protocol

We recruit **five board-certified clinicians** from collaborating hospitals, each with formal clinical training and practical experience. All annotators are fluent in English. From the MedQA test results, we randomly sample **25 test cases per model**. All instances are **randomly shuffled and anonymized** such that evaluators are blind to model identity.

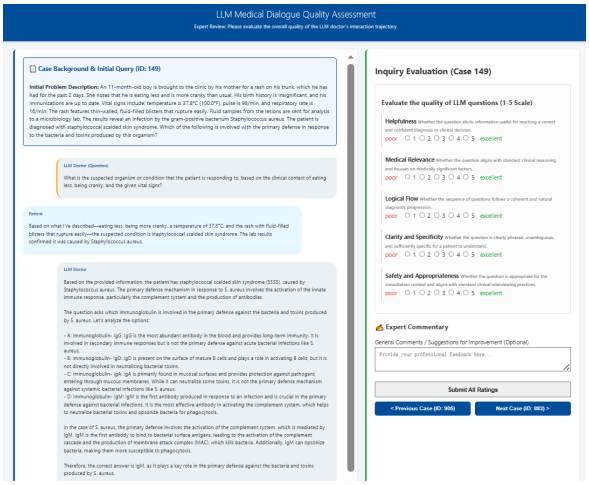


Figure 10: Screenshot of the web-based interface used for human evaluation of model-generated intermediate questions.

The evaluation is conducted through a web-based annotation platform, which displays the patient context together with the interaction trajectory. A screenshot of the annotation interface is shown in Figure 10. Annotators are asked to assess the quality of the intermediate questions according to the predefined evaluation criteria and assign a score for each dimension. No discussion or coordination between annotators is allowed during the evaluation process.

All annotators were compensated in compliance with local labor regulations, with payment rates exceeding the applicable minimum hourly wage. All participants consented to the use of their annotations in this study, and the experimental protocol was approved by the relevant ethics review board.

J Case Study

To demonstrate how our method improves the targeted information-seeking ability of LLMs in clinical contexts, we present representative cases from MedQA and MIMIC-IV datasets on `LLaMA3.1-8B`.

MedQA Case Study. As shown in Figure 11, the model before ProMed optimization operates under a reactive paradigm, immediately predicting the wrong rheumatoid arthritis based solely on age and symptom chronicity, without seeking further clarifying information. This reactive behavior leads to an incorrect diagnosis and highlights a critical risk in medical applications: making premature decisions under insufficient information. In contrast, the ProMed-optimized model proactively asks a

high-value question about nail changes, which successfully reveals a key clinical feature, nail pitting, which is essential for correctly identifying psoriatic arthritis (Eastmond and Wright, 1979; Jha et al., 2019).

MIMIC-VI Outcome Prediction Case Study. As shown in Figure 12, the GRPO-trained model exhibits a superficially guided reasoning pattern, focusing on relatively stable vital signs. It reinforces this bias by querying low-impact details (e.g., GCS eye score) and over-weighting procedure success, ultimately leading to an incorrect prediction of a positive clinical outcome. This behavior reflects a common failure mode in complex clinical settings: over-reliance on non-decisive features, while neglecting latent indicators of poor prognosis. In contrast, the ProMed-optimized model adopts a more targeted querying strategy, explicitly probing for respiratory status and overall comorbidity burden. These queries uncover critical factors, including prior respiratory failure, significant multimorbidity (e.g., COPD/asthma, CKD, CAD), and functional decline, which collectively dominate outcome risk despite localized signs of stabilization. ProMed demonstrates a stronger capacity to identify the true drivers within MIMIC-IV's complex, multi-source data, leading to a correct prediction. This case highlights the importance of guided information acquisition in mitigating misleading correlations.

MIMIC-VI Readmission Prediction Case Study. As shown in Figure 13, the GRPO-trained model only asks for patient conditions while ignoring the patient's social support system. It focuses narrowly on the presence of unresolved medical issues and incorrectly predicts a high likelihood of readmission. In contrast, the ProMed-optimized model proactively initiates a multi-step inquiry to capture a more comprehensive clinical picture. While it recognizes the high disease burden, its targeted questions regarding social support and discharge specifics reveal a critical turning point: the patient was transitioned to home hospice for comfort-focused care. By identifying that the management strategy had shifted from acute intervention to symptom control, ProMed correctly predicts that the patient will not be readmitted. This case underscores ProMed's broader evidence-gathering ability to recall and integrate diverse latent factors.

Summary. These cases demonstrate how ProMed enhances the model's ability to detect missing but diagnostically salient information, acquire it through targeted follow-up questions and apply clinically informed reasoning to make accurate decisions. ProMed effectively shifts medical LLMs from a reactive to a proactive paradigm, and achieves improved diagnostic accuracy and safer, guideline-aligned decision-making in complex clinical scenarios.

K Prompts

K.1 Doctor System Prompt.

We design a task-specific system instruction that guides the LLM to act as a clinical decision maker. It organizes the partial information question and the instructions to guide the LLM to proactively ask follow-up questions when the given information is insufficient for an accurate prediction, and to output the final answer once it has gathered enough evidence. The prompt explicitly encourages iterative questioning and targeted information seeking.

Partial Information Question	A 67-year-old man with a 16-year history of arthritis presents with right knee swelling and pain. What is the most likely diagnosis?
Options	**A: Psoriatic arthritis (correct)** | B: Arthritis mutilans | C: Rheumatoid arthritis | D: Mixed connective tissue disease
Missing Facts	1. His left knee was swollen a few weeks ago. 2. Now both joints are affected, and he has difficulty walking. 3. He feels frustrated. 4. He has back pain. 5. The back pain makes it extremely difficult to move around. 6. His pain improves with rest. 7. He suffers from dandruff. 8. He uses special shampoos for his dandruff. 9. **Physical examination reveals pitting of his nails** .

Base Model Output

Model Answer: C (Incorrect)

The patient is a 67-year-old man with a 16-year history of arthritis with right knee swelling and pain. Key considerations include:

1. Psoriatic Arthritis (A): Often associated with psoriasis, this condition commonly involves distal joints and can present with asymmetric joint involvement. The chronic nature of arthritis (16 years) and the absence of explicit mention of psoriasis or distal joint involvement make this less likely.

3. Rheumatoid Arthritis (C): Common in older adults, RA often presents with symmetric joint involvement (e.g., hands, knees) and can lead to chronic joint swelling and pain. The patient's long-standing arthritis aligns with RA, and knee involvement is a common site.

> *The most likely diagnosis is Rheumatoid Arthritis (C), as it aligns with the chronicity of symptoms, knee involvement, and the patient's age.* **[answer: C]**

ProMed Model Output

> **Model Question:** Does the patient have a history of psoriasis, and are there any nail changes (e.g., pitting, onycholysis) or dactylitis?
>
> **Patient Response:** I don't know about psoriasis, but I do have pitting of my nails.

Model Answer: A (Correct)

The patient is a 67-year-old man with a 16-year history of arthritis, presenting with right knee swelling and pain. **Key findings include nail pitting and no reported psoriasis history.**

— *Psoriatic arthritis (A)* is strongly supported by the presence of nail pitting, a hallmark feature. While psoriasis is not explicitly confirmed, the combination of arthritis and nail changes aligns with this diagnosis.

— *Arthritis mutilans (B)* is a severe, deforming variant of psoriatic arthritis, but the absence of reported dactylitis or severe joint destruction makes this less likely.

— *Rheumatoid arthritis (C)* typically involves symmetric small-joint involvement and is less likely in this case, though it cannot be excluded without joint-specific exam findings.

— *Mixed connective tissue disease (D)* is less probable given the lack of systemic symptoms (e.g., Raynaud's, myositis) and the absence of other rheumatologic features.

> *The nail changes and chronic arthritis history most strongly suggest psoriatic arthritis, even without confirmed psoriasis.* **[answer: A]**

Figure 11: MedQA case study.

Partial Information Question	**Partial information:** age 77 male; GCS total 14 (motor 6); SBP 116; HR 69; glucose 88; temperature 36.83°C; LOS ≈ 146.1 days. Discharge note fragment includes NSTEMI transfer, COPD/asthma exacerbation with respiratory failure, and CABG after cardiac catheterization. Based on the patient information above (some vital signs, lab values, and a portion of the discharge note), predict whether this patient will have a positive clinical outcome. You may ask the patient for more details if needed.
Options	A: True (positive clinical outcome) | **B: False (correct)** (not a positive clinical outcome)
Missing Facts	1. Very prolonged hospitalization (~146 days) despite eventual stabilization. 2. Recent acute hypercarbic hypoxemic respiratory failure requiring intubation before transfer. 3. High-risk multimorbidity burden: COPD/asthma, diabetes, hypertension, CKD, coronary artery disease . 4. Cardiac disease not limited to one clean signal: NSTEMI plus diffuse CAD and CABG-level intervention. 5. Functional decline and frailty concerns emerged during course, not fully reflected by isolated "alert/oriented" snippets.

GRPO Model Output

Model Question: What was the patient's GCS eye response score?

Patient Response: GCS eye response score was 3.

Model Question: Was the patient's cardiac catheterization and CABG procedure successful, and did it improve his cardiac function ?

Patient Response: Pre-bypass LVEF was 55% with normal global systolic function; post-bypass biventricular function was preserved and no major immediate surgical complication was highlighted.

Model Answer: A (Incorrect)

> *Predicted positive clinical outcome.* **[answer: A]**

ProMed Model Output

Model Question: What is the patient's respiratory and functional status given prior COPD/asthma respiratory failure ?

Patient Response: Respiratory failure required intubation earlier in the course; currently improved but still in a complex recovery context with substantial functional decline.

Model Question: What are the overall comorbid conditions and their impact on current status ?

Patient Response: Significant burden from HTN, DM, COPD/asthma, CKD, and CAD, with cumulative impact on recovery risk.

Model Answer: B (Correct)

> *Predicted no positive clinical outcome.* **[answer: B]**

Figure 12: MIMIC-IV outcome prediction case study.

Dataset	Original Question	Decomposed Results	Partial Information Question	Options	Answer
MedQA	A 70-year-old man presents with hematuria, lower abdominal pain, urinary frequency, and urgency. He recently completed chemotherapy for non-Hodgkin lymphoma. Which medication in the chemotherapy regimen most likely caused his symptoms?	**Atomic Facts:** The patient is male. The patient is 70 years old. The patient reports blood in his urine. The patient reports lower abdominal pain. The patient is concerned about urinary frequency. The patient is concerned about urinary urgency. The patient recently completed chemotherapy for non-Hodgkin lymphoma. **Atomic Question:** Which medication in the chemotherapy regimen most likely caused his symptoms?	A 70-year-old man presents with hematuria, lower abdominal pain, urinary frequency, and urgency. Which medication in the chemotherapy regimen most likely caused his symptoms?	A: Cytarabine B: Methotrexate C: Rituximab D: Cyclophosphamide E: Prednisone	D
CMB	**CN:** 男性，25岁，被热油烧伤，总面积60%，血压10/8kPa，中心静脉压0.294kPa。表明该病人存有什么问题? **EN:** Male, 25 years old, suffered 60% burn from hot oil, BP 10/8kPa, CVP 0.294kPa. What condition does this suggest?	**Atomic Facts:** **CN:** 患者是男性。 患者年龄25岁。 被热油烧伤。 烧伤面积达60%。 血压为10/8kPa。 中心静脉压为0.294kPa。 **EN:** The patient is male. The patient is 25 years old. The patient was burned by hot oil. The total burn area is 60%. The patient's blood pressure is 10/8 kPa (75/60 mmHg). The patient's central venous pressure is 0.294 kPa (3 cmHO). **Atomic Question:** **CN:** 表明该病人存有什么问题? **EN:** What condition does this suggest?	**CN:** 患者是男性。患者年龄25岁。被热油烧伤。表明该病人存有什么问题? **EN:** The patient is male. The patient is 25 years old. The patient was burned by hot oil.What condition does this suggest?	**CN:** A: 血容量不足 B: 心功能不全 C: 血容量相对过多 D: 血容量严重不足 E: 容量血管过度收缩 **EN:** A: Mild hypovolemia B: Cardiac insufficiency C: Relative hypervolemia D:Severe hypovolemia E: Excessive constriction of capacitance vessels	D

Table 14: Examples from the interactive medical questioning datasets. CMB samples show both Chinese and English translations.

Partial Information Question	**Partial information:** age 60 male, GCS 15, SBP 138, HR 125, temperature 35.56°C, LOS ~28.7 days; history includes laryngeal cancer, metastatic NSCLC, brain metastasis/subdural hematoma, admission for dyspnea. Based on the patient information, predict whether this patient will be readmitted within the follow-up period.
Options	A: True (will be readmitted) | **B: False (correct)** (will not be readmitted)
Missing Facts	1. Discharge disposition: home hospice / comfort-focused care . 2. Functional status declined significantly with advanced metastatic disease burden. 3. Family/caregiver involvement was active in discharge planning. 4. Ongoing management centered on symptom control (e.g., morphine/Ativan), not recurrent acute intervention.

GRPO Model Output

Model Question: Was the patient discharged with any ongoing medical issues or treatments that would require follow-up care?
Patient Response: Yes. Multiple serious diagnoses and unresolved follow-up items were listed (DKA/cardiac/lab/imaging concerns and specialist follow-up).
Model Answer: A (Incorrect)

Predicted readmission. **[answer: A]**

ProMed Model Output

Model Question: What are the patient's history of prior hospitalizations/readmissions and current disease burden?
Patient Response: Multiple prior admissions for metastatic cancer-related complications and neurologic events.
Model Question: What is the patient's current functional status and ability to perform daily activities ?
Patient Response: Advanced decline; fatigue and reduced independence with high symptom burden.
Model Question: What is the patient's social support / caregiver setup ?
Patient Response: Family is actively involved in care decisions and support.
Model Question: What are the discharge plans and ongoing home care services ?
Patient Response: Discharged to home hospice with comfort-care medications and symptom-focused management.
Model Answer: B (Correct)

Predicted no readmission. **[answer: B]**

Figure 13: MIMIC-IV readmission prediction case study.

Algorithm 2 SIG-Guided MCTS Sampling

Require: Initial inquiry Q_p, atomic facts $\mathcal{F}$, ground-truth answer A^*, model $\mathcal{M}$, simulated patient $\mathcal{P}_{\text{sim}}$, maximum depth $T_{\max}$, number of simulations N

Ensure: Answer-correct optimal trajectory $\tau^* = \{Q_p, (q_1, r_1), \ldots, (q_T, r_T), A'\}$

Initialize root node $n_0 \leftarrow Q_p$
Initialize best trajectory $\tau^* \leftarrow \emptyset$, best reward $R^* \leftarrow -\infty$
for $i = 1$ to N **do**
    Initialize path $\mathcal{P} \leftarrow [n_0]$, history $\mathcal{H}_0 \leftarrow \emptyset$
    ▷ Selection
    while n is fully expanded **and** not terminal **do**
        $n \leftarrow \arg\max_{n'} \mathrm{UCT}(n')$
        $\mathcal{P} \leftarrow \mathcal{P} \cup \{n\}$
    end while
    ▷ Expansion
    if n is not terminal **then**
        Generate question $q_t \leftarrow \mathcal{M}(Q_p, \mathcal{H}_{t-1})$
        Simulate response $r_t \leftarrow \mathcal{P}_{\text{sim}}(q_t, \mathcal{F})$
        $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(q_t, r_t)\}$
        Add node $n_t = (q_t, r_t)$ to $\mathcal{P}$
        Compute local reward $R_{\text{local}}(q_t) \leftarrow \mathrm{SIG}(q_t)$
    else
        Predict final answer $A' \leftarrow \mathcal{M}(Q_p, \mathcal{H}_{t-1})$
    end if
    ▷ Simulation
    while not terminal **and** $t < T_{\max}$ **do**
        Generate question q_t, simulate response r_t
        $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(q_t, r_t)\}$
    end while
    Predict answer $A' \leftarrow \mathcal{M}(Q_p, \mathcal{H}_t)$
    Compute trajectory reward $R(\tau)$
    ▷ Backpropagation
    for all $n \in \mathcal{P}$ **do**
        $N(n) \leftarrow N(n) + 1$
        $W(n) \leftarrow W(n) + R(\tau)$
        $\bar{R}(n) \leftarrow W(n)/N(n)$
    end for
    if $A' = A^*$ **and** $R(\tau) > R^*$ **then**
        $\tau^* \leftarrow \tau$
        $R^* \leftarrow R(\tau)$
    end if
end for
return τ^*

Doctor System Prompt

You are a professional doctor with excellent reasoning and analytical skills in diagnosing medical conditions, as well as strong abilities in clinical inquiry and patient evaluation.
Your task is to answer a problem based on patient information. **The information you are given may be incomplete**. You should rely on your medical knowledge, the patient's current status, and the clinical question to **ask follow-up questions and obtain necessary supplementary information**.

Below is a {question_type} problem based on patient information:
Problem: {question}
Options: {option_str}

Please analyze the problem thoroughly using your professional medical knowledge.
During each round of dialogue, if you believe the current patient information is insufficient to determine the correct answer, you should analyze the options and **ask a targeted question to gather essential information that will help you make the correct diagnosis**.
If you think the available information is sufficient to answer the question, please **combine all relevant medical knowledge and patient data to perform a detailed analysis and provide the correct answer**.

Important instructions:
1. Each of your responses must follow one of the two formats below:
a. If you need to ask a question, start your response with **"question:"** followed by the specific question you want to ask based on the options and current patient information;
b. If you are ready to give the final answer, start with **"answer:"**, then provide your detailed reasoning, and end with your chosen option in the format: [answer: XXX].
2. If there is uncertainty due to incomplete patient information, you must ask follow-up questions to gather more data.
3. In each round, you may only ask one question or provide the final answer.
4. You may ask up to 10 questions; after that, you must provide your final answer.

K.2 Patient Prompt.

To simulate realistic patient responses, we construct a system prompt that instructs the LLM to role-play as the patient. Given a set of atomic ground-truth facts $\mathcal{F}$, the model is asked to respond faithfully and concisely to each doctor-issued question using only the available facts and output "I don't know" when no facts are applicable. This ensures alignment with the underlying clinical condition and prevents hallucinated or overly informative answers:

Simulated Patient Prompt

You are a patient undergoing a medical consultation. Your basic health condition is entirely based on the atomic facts provided below. You will interact with the doctor by answering the questions they ask, using only the information given. You must not reveal that you are a language model; instead, treat the provided information as your actual health status.

Your information is as follows:
{atomic_facts}

During your interaction with the doctor, please adhere to the following guidelines:
1. Your responses must be strictly based on the provided facts. Do not add, assume, or fabricate any information beyond what is explicitly stated.
2. If you are unable to answer a question based on the facts, respond with "I don't know" or another appropriate expression of uncertainty.
3. Do not mention or imply that your responses are drawn from predefined records or external data. Your expressions should feel natural, as if they reflect your own experiences and conditions.
4. Do not state or imply that you are simulating or playing the role of a patient. Assume the identity of someone who is genuinely experiencing these symptoms.

K.3 Reward Calculation Prompts.

Doctor Understanding Prompt. To accurately compute the Shapley Information Gain reward for a candidate question q_t, we require an intermediate representation of the model's current understanding of the patient's condition. Specifically, we design a prompt to elicit the LLM's implicit reasoning state, denoted as U_t, which is dynamically constructed based on the initial inquiry and the accumulated dialogue history up to time step t. This serves as the context for evaluating the marginal information gain introduced by q_t. The prompt instructs the LLM to act as a professional physician and generate a comprehensive and structured summary of the patient's medical condition, grounded in the provided facts and prior interactions:

Doctor Understanding Prompt

You are a professional physician. Your task is to **provide a comprehensive understanding and summary of the patient's current condition** based on the provided patient information and doctor-patient dialogue. Your summary should reflect a clear grasp of the patient's medical history, current symptoms, relevant diagnostic information, test results, and possible diagnostic directions.

Known patient information:
{patient_information}
Doctor-patient dialogue:
{dialogue}

Based on the above information, please provide your overall understanding of the patient. You must include all explicit information and reasonable inferences based on the available data. Do not make any unfounded guesses or fabricate facts.

Your summary may include:
1. Basic patient information and medical history overview, such as age, gender, past medical history, family history, and allergy history.
2. The patient's chief complaint and current symptoms, identifying the most prominent discomforts or symptoms.
3. Summary of physical signs and test findings, describing relevant signs and abnormal test results based on the available data and dialogue.
4. Possible diagnoses, suggesting plausible diagnoses at the current stage.

Please ensure your summary is medically professional and logically coherent, and avoid omitting any important information.

Fact Checker Prompt. The *Fact Checker Prompt* during the computation of SIG reward. It checks whether each atomic fact is entailed by the model's current understanding U_t. It formulates a binary (True/False) query for each fact given the current context, enabling us to measure the information gained from a candidate question q_t as the number of facts newly verified as `True`.

Fact Checker Prompt

Answer the question about patient information based on the given context.

Context: {context}
Input: {fact} True or False?
You should only reply True or False, no other information should be outputted.
Output:

K.4 Dataset Construction Prompts.

Judge Patient Prompt. To construct interactive medical questioning datasets, we filter out questions that do not involve patient-specific scenarios. While the CMB dataset covers a wide range of medical topics, many items reflect general medical knowledge rather than patient-centered consultation. To address this, we introduce a *Judge Patient Prompt*, which instructs the LLM to determine whether a given question is based on the analysis of a specific patient's medical condition. This binary classification helps us retain only those questions suitable for interactive doctor-patient dialogues.

Judge Patient Prompt

Please refer to the examples and **determine whether the following question is based on the analysis of a patient's medical record**. Only output "Yes" or "No" as the answer; do not include any additional text:

Examples:
Question: A 30-year-old male fell from the third floor and injured his left abdomen. He sustained fractures of the 6th, 7th, and 8th left ribs, splenic rupture, and intestinal rupture. Upon admission, he was tense, had a temperature of 38.5°C, pale complexion, cold extremities, rapid thready pulse at 110 bpm, blood pressure 130/100 mmHg, and reduced urine output. Which of the following examinations is currently inappropriate?
Answer: Yes

Question: In a certain region, the average life expectancy of women in 2005 was 72.24 years, and in 2009 it was 75.47 years. The two years' life expectancies can be compared because the life table indicator...
Answer: No

Question: {Question}
Answer:

Atomic Fact Decomposition Prompt. In the original MedQA and CMB datasets, each clinical question typically presents all patient information at once, which does not align with the partial-information setting required for simulating interactive medical consultations. To bridge this gap, we introduce an *Atomic Fact Decomposition Prompt* that transforms the full question stem into a set of atomic facts, where each fact represents a minimal, self-contained piece of patient information. This decomposition allows us to create realistic interaction scenarios in which the model gradually acquires information through questioning, and provides the foundation for computing fact-based Shapley information gain rewards.

Atomic Fact Decomposition Prompt

Please refer to the example and **decompose the following clinical question stem into atomic facts** about the patient.
Each atomic fact should be a complete sentence. You should only output the atomic facts, one sentence per line.
Do not output any extra content:

Example:
Question:
Male, 55 years old. He experienced upper abdominal discomfort and vomiting for the past 2 days. The vomitus contained sour-smelling food residue and symptoms were relieved after vomiting. Physical examination revealed visible gastric peristalsis.

Answer:
The patient is male.
The patient is 55 years old.
The patient experienced upper abdominal discomfort for the past 2 days.
The patient experienced vomiting for the past 2 days, and the vomitus contained sour-smelling food residue.
The patient's symptoms were relieved after vomiting.
Physical examination revealed visible gastric peristalsis.
Physical examination revealed visible peristaltic waves.

Question:{Question}
Answer:

K.5 Evaluation Prompt

Question Quality Evaluation Prompt We use the following prompt to instruct an LLM evaluator to assess the quality of model-generated questioning trajectories along multiple clinically grounded dimensions.

LLM-based Question Quality Evaluation Prompt

You are a fair evaluator of medical dialogues.
Below is a simulated doctor–patient conversation (trajectory). Your task is to rate the **entire conversation trajectory** on the following five dimensions, each scored from 1 (very poor) to 5 (excellent):

- **Helpfulness**: Whether the questions are likely to elicit information useful for reaching a correct and confident diagnosis or clinical decision.
- **Medical Relevance**: Whether the questions align with standard clinical reasoning and focus on medically significant factors.
- **Logical Flow**: Whether the sequence of questions follows a coherent and natural diagnostic progression.
- **Clarity and Specificity**: Whether the questions are clearly phrased, unambiguous, and sufficiently specific for patient understanding.
- **Safety and Appropriateness**: Whether the questions are appropriate for the consultation context and adhere to standard clinical interviewing practices.

Evaluation Guidelines:

- Evaluate the entire trajectory as a whole rather than individual questions.
- Be fair and balanced; give credit for reasonable clinical reasoning even if the final answer is incorrect.
- Use the full score range (1–5), but assign very low scores (1–2) only for clearly poor performance.

Conversation:
{dialogue}
Correct Answer: {gold_answer}
Final Answer Given: {final_answer}
Answer Correct: {answer_correct}
Output Format (JSON only):
{ "helpfulness": X, "medical_relevance": Y, "logical_flow": Z, "clarity_and_specificity": W, "safety_and_appropriateness": V }
Do not include any commentary or explanation.

Open-Ended Answer Accuracy Evaluation Prompt. For open-ended questions in MedQA, we use the following prompt instructing an LLM to judge whether the model-generated answer correctly matches the golden reference.

LLM Evaluation Prompt

You are a medical exam evaluator. Your task is to determine if the model's answer correctly matches the golden answer.
Question: {question}
Golden Answer: {gold_answer}
Model Answer: {model_answer}
Please refer to the question and the golden answer, decide if the model answer correctly answers the question. Reply only with one word: "yes" or "no".

L Code and Data Availability

To support reproducibility and facilitate future research, we will publicly release all code and processed datasets upon publication. For reference and transparency, the complete code is also provided in https://github.com/hxxding/ProMed.

M Use of Large Language Models

In this work, LLMs are used in a supportive role to assist with language refinement and programming debugging tasks. All LLM-assisted outputs are carefully reviewed, verified, and revised by the authors before inclusion. The core research ideas, methodological design, experimental setup, and result analysis are conceived and carried out entirely by the authors.

N Notations Table

This section summarizes the key notations used in the ProMed framework for interactive medical questioning.

Symbol	Description
$\mathcal{D} = \{\mathcal{X}_i\}_{i=1}^{N}$	Dataset of N patient cases
$\mathcal{X}_i$	Patient case: initial question Q, atomic facts $\mathcal{F}$, ground-truth answer A^*
$Q_p = (F_p, Q)$	Partial-information question ($F_p \subset \mathcal{F}$)
A'	Model-generated answer
$\mathcal{H}_t$	Dialogue history up to turn t
s_t	Model belief state; U_t interpretable proxy
q_t, r_t	Model question and patient response at turn t
$IG(q_t)$	Raw information gain (new facts)
$v(S)$	Value of fact subset S (log-prob of A^*)
$\phi(f_i), \tilde{\phi}_i$	Shapley value and softmax-normalized importance
SIG(q_t)	Shapley Information Gain reward
τ	Interaction trajectory of length T
$R(\tau)$	Trajectory-level reward
$R(q_t), R(A')$	Action-level reward
$\{x_i\}$	Token sequence of τ
$r(x_i)$	Token-level reward
$\mathcal{G}$	Group of trajectories for GRPO
$\mathcal{R}_\mathcal{G}$	Group-relative token rewards for advantage

Table 15: Key notations in ProMed for interactive medical questioning, SIG reward, SFT, and RL (compressed version).