PCQC: PRIVILEGED COUNTERFACTUAL QUESTION CREDIT FOR MULTI-TURN MEDICAL DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have made substantial progress on medical question-answering, yet effective medical dialogue also requires learning to ask questions that uncover relevant patient information. To train such dialogue policies, a common pipeline combines supervised fine-tuning with reinforcement learning (RL). Existing RL objectives derive supervision from realized or selectively expanded interactions: outcome rewards evaluate complete consultations, turn-level rewards evaluate executed questions, and tree-based methods estimate values along expanded branches. Under these objectives, other questions available at the same dialogue state receive no direct supervision unless their future branches are explored. We introduce PCQC (Privileged Counterfactual Question Credit), which learns from *questions never asked* by using privileged patient facts to answer alternative questions at the same dialogue state. A frozen diagnostic scorer evaluates each resulting question–answer pair, and within-state comparisons yield relative question credit for both executed and unexecuted questions without rolling the alternatives forward. At deployment, the policy requires neither privileged patient facts nor the diagnostic scorer. Across four medical benchmarks, PCQC achieves 63.10% mean diagnostic accuracy, outperforming GRPO and ATPO by 4.38 and 4.21 percentage points, respectively, while using 33.1% fewer inquiry turns than GRPO.

1 INTRODUCTION

In real-world clinical consultations, patients often come with incomplete information, and accurate diagnosis depends on follow-up questions that progressively uncover relevant symptoms and medical history. Large language models (LLMs) have demonstrated strong performance on medical examinations and case-based diagnosis (Singhal et al., 2023; 2025). Building on these capabilities, recent approaches use prompting (Li et al., 2024) and supervised fine-tuning (Tu et al., 2025) to extend medical language models to interactive consultations.

However, effective medical dialogue requires more than the ability to conduct multi-turn interactions: models must learn which questions to ask to acquire diagnostically useful information. Each question determines which patient information becomes available for subsequent inquiries and the final diagnosis, making question selection central to the consultation.

Learning such a question-selection policy requires supervision that can distinguish the value of individual inquiry decisions. Supervised fine-tuning provides the basic capability for interactive consultation, while reinforcement learning (RL) can further optimize question selection through interaction and diagnostic feedback (Lai et al., 2026; Cao et al., 2026). Existing RL supervision, however, remains tied to realized or selectively expanded interactions. Terminal rewards evaluate the consultation as a whole; finer-grained rewards evaluate information obtained along executed dialogue turns (Ding et al., 2026; Wang et al., 2026; Xie et al., 2026); and tree-based methods derive policy supervision from branches selected for expansion (Cao et al., 2026). Under these objectives, the other questions available at the same dialogue state receive no direct question-level supervision unless their future branches are explored.

Our key observation is that these alternatives need not be rolled forward to be evaluated during training: patient facts beyond the observed dialogue provide privileged information about what each candidate question would reveal. We exploit this information by learning from *questions never*

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

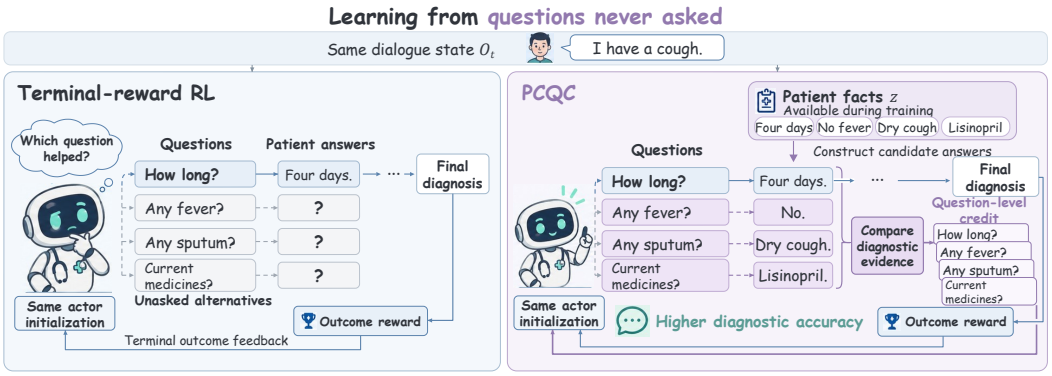

Figure 1: **Learning from questions never asked.** Terminal-reward RL (left) learns from final diagnoses. PCQC (right) additionally uses patient facts to answer unexecuted questions at the same dialogue state, turning comparisons of candidate question–answer pairs into relative question credit.

asked (Figure 1). A patient responder uses these facts to answer candidate questions at the same dialogue state, including those that are never selected for rollout. This produces same-state counterfactual question–answer pairs that expose the diagnostic evidence associated with both executed and unexecuted questions, without further dialogue rollouts.

We turn these counterfactual comparisons into learning signals with Privileged Counterfactual Question Credit (PCQC). A frozen diagnostic scorer evaluates how strongly each candidate question–answer pair supports the correct diagnosis, and their utilities are compared within the same dialogue state to derive relative question credit. This credit directly supervises question selection across both executed and unexecuted candidates, extending learning beyond the branches explored during rollout. We combine this question-level supervision with terminal GRPO, enabling the policy to learn both which information to acquire and how to use it for the final diagnosis. At deployment, the learned policy requires neither full patient information nor the diagnostic scorer.

Across MedQA, MedicalExam, MedMCQA, and iCRAFT-MD, PCQC achieves 63.10% mean diagnostic accuracy, surpassing GRPO (58.72%) and ATPO (58.89%) while requiring 33.1% fewer inquiry turns than GRPO. With the same frozen diagnostic scorer, PCQC exceeds executed-question feedback by 5.36 points in three-benchmark mean accuracy. At a matched training budget, PCQC further exceeds an executed-only Same-State-Q1 control by 3.07 points in four-benchmark mean accuracy, while diagnostic utility identifies candidates with higher downstream diagnostic success.

The main contributions of this work are as follows:

- We introduce learning from *questions never asked* for medical dialogue policies. Privileged patient information exposes the diagnostic evidence elicited by unexecuted questions, enabling question-level supervision without further dialogue rollouts.
- We propose PCQC, which compares the diagnostic utilities of candidate question–answer pairs at the same dialogue state to derive relative question credit. This credit directly supervises both executed and unexecuted questions and complements terminal GRPO to jointly train question selection and final diagnosis.
- Across four medical dialogue benchmarks, PCQC achieves 63.10% mean diagnostic accuracy, outperforming GRPO and ATPO by more than four percentage points while requiring 33.1% fewer inquiry turns than GRPO. Controlled comparisons show that directly supervising unexecuted questions adds value beyond executed-question feedback, and continuation experiments validate diagnostic utility as a predictor of downstream diagnostic success.

2 Related Work

2.1 Medical Dialogue Agents

Recent work has extended medical LLMs from static question answering to interactive clinical consultation. MediQ studies question asking under incomplete information, while AMIE and Agent-Clinic investigate diagnostic agents that acquire information through sequential patient interactions (Li et al., 2024; Tu et al., 2025; Schmidgall et al., 2026). Building on these interactive settings, DoctorAgent-RL learns questioning policies through simulated consultations, and Doctor-R1 jointly optimizes clinical inquiry and decision making with reinforcement learning (Feng et al., 2025b; Lai et al., 2026). ALFA instead aligns question asking with fine-grained attributes such as clarity, answerability, medical accuracy, and diagnostic relevance (Li et al., 2025). These studies establish question selection as a learnable component of medical dialogue. PCQC directly assigns diagnostic credit among alternative inquiries available at the same dialogue state, including questions that are never executed in the consultation.

2.2 Fine-Grained Credit Assignment

A central challenge in multi-turn agent training is assigning credit to individual decisions when supervision is dominated by delayed trajectory outcomes. One line of work addresses this problem through finer-grained value estimation and structured comparisons of intermediate decisions. VinePPO estimates intermediate state values from Monte Carlo continuations, while GiGPO groups actions associated with the same anchor state across collected trajectories to construct step-level relative advantages (Kazemnejad et al., 2025; Feng et al., 2025a). In medical dialogue, ATPO uses adaptive tree expansion and critic-based value estimates to guide branch exploration and propagate supervision through the search tree (Cao et al., 2026). IGRPO similarly uses the informativeness of intermediate states to allocate rollout budget within tree-structured exploration (Zhang et al., 2026). These approaches obtain finer-grained learning signals from future returns, repeated states, or selectively expanded trajectories.

A complementary line of work constructs dense supervision from the information revealed during interaction. ProMed rewards newly acquired patient information according to its Shapley-weighted diagnostic value (Ding et al., 2026). IGPO measures turn-level information gain through changes in the policy's likelihood of the correct answer, while TIPS uses a teacher model to construct analogous likelihood-based rewards (Wang et al., 2026; Xie et al., 2026). InfoPO instead measures how observed feedback changes the agent's subsequent action distribution relative to a masked-feedback counterfactual (Kong et al., 2026). These methods derive fine-grained supervision from information obtained through realized or explicitly explored interactions. PCQC extends this supervision to alternative questions that are never executed: privileged patient information allows a responder to answer multiple candidates at the same dialogue state, so their diagnostic utilities can directly supervise question selection without further dialogue rollouts. Using privileged facts to instantiate answers to alternative questions is related to asymmetric learning approaches (Pinto et al., 2018; Chen et al., 2020). The deployed PCQC policy, however, acts only on the observed dialogue.

3 Privileged Counterfactual Question Credit

Figure 2 summarizes the PCQC training framework. The following subsections formalize the clinical inquiry setting, construct and evaluate same-state candidate interactions, derive relative question credit, and define the joint optimization with terminal GRPO.

3.1 Clinical Inquiry as Selective Information Acquisition

We formulate clinical inquiry as sequential information acquisition for diagnosis. For each consultation, z denotes the full set of patient facts recorded in the case, which remains fixed throughout the dialogue. At turn t, the doctor policy π_θ observes the dialogue state O_t, consisting of the initial patient information and all preceding question–answer exchanges. The clinical task and answer options are also provided to the policy and diagnostic scorer but are omitted from the notation for brevity.

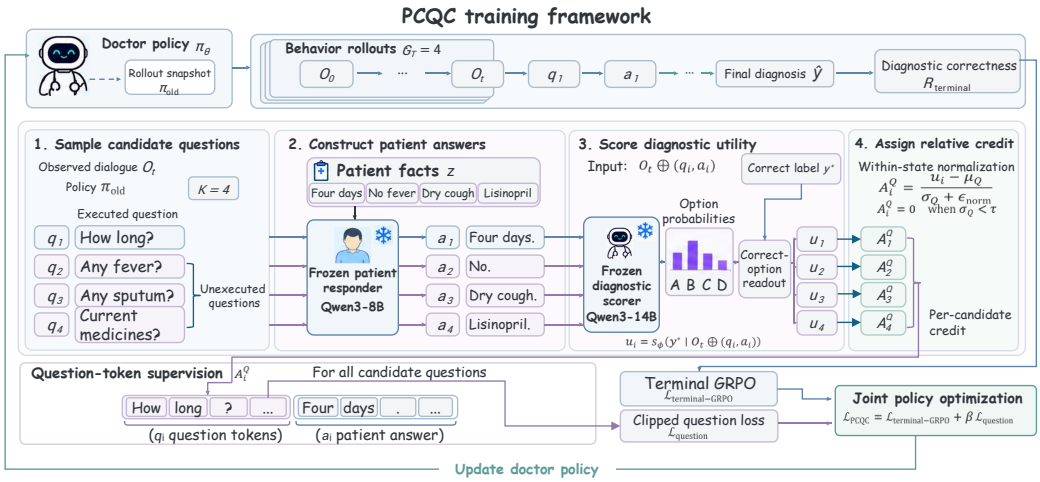

Figure 2: **PCQC training framework.** The policy samples candidate questions at O_t. A frozen patient responder answers them using patient facts z, and a frozen diagnostic scorer assigns diagnostic utilities u_i. Within-state normalization yields relative question credit A_i^Q for candidate question tokens, alongside terminal GRPO. Only (q_1, a_1) continues the consultation.

Given O_t, the policy either asks a follow-up question or makes a final diagnosis. When it asks a question q_t, a patient responder $\mathcal{E}$ generates an answer a_t conditioned on the patient facts z:

$$\begin{aligned} q_t &\sim \pi_\theta(\cdot \mid O_t), \qquad a_t \sim \mathcal{E}(\cdot \mid z, q_t), \\ O_{t+1} &= O_t \oplus (q_t, a_t). \end{aligned} \tag{1}$$

Here, $\oplus$ appends the question–answer pair to the dialogue state. At the end of the consultation, the final diagnosis $\hat{y}$ is evaluated against the reference answer y^* to determine the terminal reward.

3.2 Learning from Questions Never Asked

During training, patient facts in z that have not yet been revealed in O_t provide *privileged patient information*. The doctor policy conditions only on O_t, whereas a frozen patient responder can use the complete case facts z to answer candidate questions. This allows PCQC to construct multiple question–answer interactions from the same observed dialogue state, including questions that are never executed in the continued consultation.

At each question-asking state selected from the behavior consultation, the rollout policy π_{old} samples $K = G_Q \geq 2$ responses x_i, from which we extract candidate questions q_i. One sampled response, x_1, is used to continue the consultation and defines the executed question q_1; the remaining responses provide unexecuted candidates $q_{2:K}$. The patient responder answers every candidate using the same patient facts z, forming the same-state comparison group

$$\begin{aligned} x_i &\sim \pi_{\text{old}}(\cdot \mid O_t), \\ a_i &\sim \mathcal{E}(\cdot \mid z, q_i), \qquad i = 1, \ldots, K, \\ \mathcal{G}_Q(O_t) &= \{(O_t, q_i, a_i)\}_{i=1}^{K}. \end{aligned} \tag{2}$$

Only (q_1, a_1) is appended to the dialogue state and advances the consultation. The remaining pairs $(q_i, a_i)_{i=2}^{K}$ are *counterfactual interactions*: they provide answers to the unexecuted questions without requiring further dialogue rollouts. Appendix B.4 provides details of candidate construction and question-token supervision.

We evaluate each interaction with a frozen diagnostic scorer s_ϕ, which maps the augmented dialogue $O_t \oplus (q_i, a_i)$ to a distribution over answer options. The probability assigned to the correct option y^* defines the candidate's *diagnostic utility*:

$$u_i = s_\phi(y^* \mid O_t \oplus (q_i, a_i)). \tag{3}$$

The scorer never observes the patient facts z directly: it receives only the observed dialogue, the candidate question–answer pair, and the fixed clinical task and answer options. The label y^* is used after scoring to read the corresponding option probability.

3.3 Relative Question Credit

Given the diagnostic utilities in Eq. 3, PCQC compares candidates within each dialogue state. The group mean and sample standard deviation are

$$\mu_Q = \frac{1}{K}\sum_{j=1}^{K} u_j, \qquad \sigma_Q = \sqrt{\frac{1}{K-1}\sum_{j=1}^{K}(u_j - \mu_Q)^2}. \tag{4}$$

The *relative question credit* for candidate i is

$$A_i^Q = \begin{cases} \dfrac{u_i - \mu_Q}{\sigma_Q + \epsilon_{\text{norm}}}, & \sigma_Q \geq \tau, \\ 0, & \sigma_Q < \tau, \end{cases} \tag{5}$$

where $\epsilon_{\text{norm}} = 10^{-6}$ stabilizes normalization and $\tau = 10^{-4}$ suppresses credit when candidate utilities are nearly indistinguishable. For groups with $\sigma_Q \geq \tau$, candidates above the group mean receive positive credit, whereas those below it receive negative credit.

Relative diagnostic gain. Because all candidates share the same pre-interaction dialogue state O_t, the same credit can equivalently be computed from diagnostic gains relative to a common baseline. Let

$$b(O_t) = s_\phi(y^* \mid O_t), \qquad \Delta_i = u_i - b(O_t). \tag{6}$$

Since the baseline is shared across candidates, $\bar{\Delta} = \mu_Q - b(O_t)$ and $\sigma_\Delta = \sigma_Q$. For $\sigma_Q \geq \tau$, Eq. 5 therefore becomes

$$\begin{aligned} A_i^Q &= \frac{\Delta_i - \bar{\Delta}}{\sigma_\Delta + \epsilon_{\text{norm}}} \\ &= \frac{1}{K(\sigma_Q + \epsilon_{\text{norm}})}\sum_{j=1}^{K}(u_i - u_j). \end{aligned} \tag{7}$$

Thus, PCQC compares each candidate's diagnostic gain with those of the other questions sampled at the same dialogue state. The implementation uses this equivalent diagnostic-gain form (Appendix B.3).

Information-theoretic interpretation. The relative comparison also connects question credit to the scorer's reduction in correct-option surprisal. For positive correct-option probabilities, this reduction after candidate i is

$$g_i = -\log b(O_t) + \log u_i = \log \frac{u_i}{b(O_t)}. \tag{8}$$

Since all candidates share $b(O_t)$, for $\sigma_Q \geq \tau$,

$$A_i^Q > A_j^Q \quad \Longleftrightarrow \quad u_i > u_j \quad \Longleftrightarrow \quad g_i > g_j. \tag{9}$$

Relative question credit therefore preserves the ordering induced by reductions in correct-option surprisal, while its magnitude is determined by normalized diagnostic gains on the probability scale. The full derivation is given in Appendix C.

3.4 Joint Optimization of Inquiry and Diagnosis

PCQC combines supervision at two levels. Terminal GRPO provides trajectory-level feedback from G_T complete consultations for the same patient case, whereas the relative credit in Eq. 5 provides question-level supervision at individual dialogue states. The latter is applied only to tokens generated by the doctor policy for candidate questions; patient responses provide scoring context and receive

no policy gradient. Both the question credit A_i^Q and rollout policy π_{old} are fixed during each policy update.

Let $\mathcal{B}_Q$ index candidate responses in an update batch, with $x_i = (w_{i,1}, \ldots, w_{i,L_i})$ sampled at $O_{t(i)}$. The binary mask $m_{i,\ell}$ selects the tokens that receive question credit (Appendix B.4). For each token, the policy probability ratio is

$$r_{i,\ell}(\theta) = \frac{\pi_\theta(w_{i,\ell} \mid O_{t(i)}, w_{i,<\ell})}{\pi_{\text{old}}(w_{i,\ell} \mid O_{t(i)}, w_{i,<\ell})}. \tag{10}$$

We optimize these tokens with a PPO-style clipped objective:

$$\mathcal{L}_{\text{question}} = -\frac{1}{N_Q} \sum_{i \in \mathcal{B}_Q} \sum_{\ell=1}^{L_i} m_{i,\ell} \min\Big\{ r_{i,\ell}(\theta) A_i^Q, \; \text{clip}(r_{i,\ell}(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})\, A_i^Q \Big\}, \tag{11}$$

where $N_Q = \sum_{i \in \mathcal{B}_Q} \sum_{\ell=1}^{L_i} m_{i,\ell}$ is the number of selected tokens in the update and $\epsilon_{\text{clip}} = 0.2$ (Schulman et al., 2017). The loss is zero when $N_Q = 0$. The same objective applies to both executed and unexecuted candidates; their additive decomposition is given in Appendix D.2.

We combine this question-level objective with terminal GRPO (Shao et al., 2024):

$$\mathcal{L}_{\text{PCQC}} = \mathcal{L}_{\text{terminal-GRPO}} + \beta\, \mathcal{L}_{\text{question}}. \tag{12}$$

The coefficient β balances trajectory-level outcome supervision and question-level credit, with each objective normalized by its own selected-token count. Reference-policy KL regularization is included in the actor update (Appendix A.2).

Together, the two objectives train the policy to acquire diagnostically useful information while optimizing the final diagnostic outcome. At deployment, the policy requires only the observed dialogue, clinical task, and answer options, without privileged patient information, counterfactual candidate construction, or the diagnostic scorer.

4 Experiments

4.1 Experimental Setup

Task and interaction environment. We evaluate medical dialogue policies in the interactive setting defined in Section 3.1. Each case provides initial patient information, fixed patient facts, and a clinical decision question with answer options. A Qwen3-8B doctor policy acquires information through interaction with a frozen Qwen3-8B patient responder before making its final diagnosis. Each consultation allows at most 10 dialogue turns, including the final-answer turn.

Training data and evaluation benchmarks. Following ATPO's released training pipeline (Cao et al., 2026), all RL methods use the same 14,256-case training set, comprising 9,400 MediQ/MedQA cases and 4,856 MedMCQA cases. We likewise follow ATPO's released evaluation protocol and use its MedQA dialogue split (1,268 cases) as the development benchmark for method development and checkpoint selection. The selected checkpoints are then evaluated on MedicalExam (150 cases), MedMCQA (536 cases), and iCRAFT-MD (139 cases) without further checkpoint selection. The iCRAFT-MD cases form an independently sourced dermatology benchmark released with MediQ (Li et al., 2024); no exact or thresholded lexical overlap is detected against the SFT or RL training cases. Dataset details are provided in Appendix A.1.

Methods and training configuration. We compare Shared SFT, GRPO, ATPO, and PCQC under the same interaction and evaluation protocol. All RL methods initialize from the same Shared SFT policy. GRPO learns from terminal diagnostic outcomes over complete consultations, whereas ATPO derives action-level supervision through adaptive tree expansion and value traceback. PCQC augments terminal GRPO with same-state relative question credit for both executed and unexecuted questions. GRPO samples $G_T = 32$ complete consultations per case; PCQC uses $G_T = 4$ terminal rollouts together with $G_Q = 4$ candidate responses at each supervised dialogue state and sets $\beta = 1$. A frozen Qwen3-14B model serves as the diagnostic scorer for PCQC. Further training details are provided in Appendices A and B, with full-system training costs reported in Appendix A.4.

Table 1: **Diagnostic accuracy (%) across four medical dialogue benchmarks.** Per-benchmark results for GRPO, ATPO, and PCQC are reported as mean ± sample standard deviation over three independent RL runs. Avg. denotes the unweighted mean across benchmarks. Best results are **bold**.

Method	MedQA	MedicalExam	MedMCQA	iCRAFT-MD	**Avg.**
Shared SFT	46.69	52.67	41.23	57.55	49.54
GRPO	56.28 ± 0.66	61.56 ± 0.38	49.19 ± 0.47	67.87 ± 4.22	58.72
ATPO	56.70 ± 1.04	61.11 ± 4.07	50.12 ± 3.61	67.63 ± 1.25	58.89
PCQC (ours)	**62.54** ± 1.64	**61.78** ± 1.54	**54.48** ± 1.13	**73.62** ± 0.83	**63.10**

Table 2: **Mean inquiry turns per consultation (Q/ep.) across four medical dialogue benchmarks.** Per-benchmark results for GRPO, ATPO, and PCQC are reported as mean ± sample standard deviation over three independent RL runs. Avg. denotes the unweighted mean across benchmarks.

Method	MedQA	MedicalExam	MedMCQA	iCRAFT-MD	**Avg.**
Shared SFT	1.519	1.373	1.466	1.482	1.460
GRPO	2.280 ± 0.237	2.358 ± 0.311	2.142 ± 0.228	2.228 ± 0.206	2.252
ATPO	1.196 ± 0.088	1.162 ± 0.086	1.098 ± 0.069	1.173 ± 0.092	1.157
PCQC (ours)	1.508 ± 0.074	1.549 ± 0.048	1.560 ± 0.047	1.410 ± 0.123	1.507

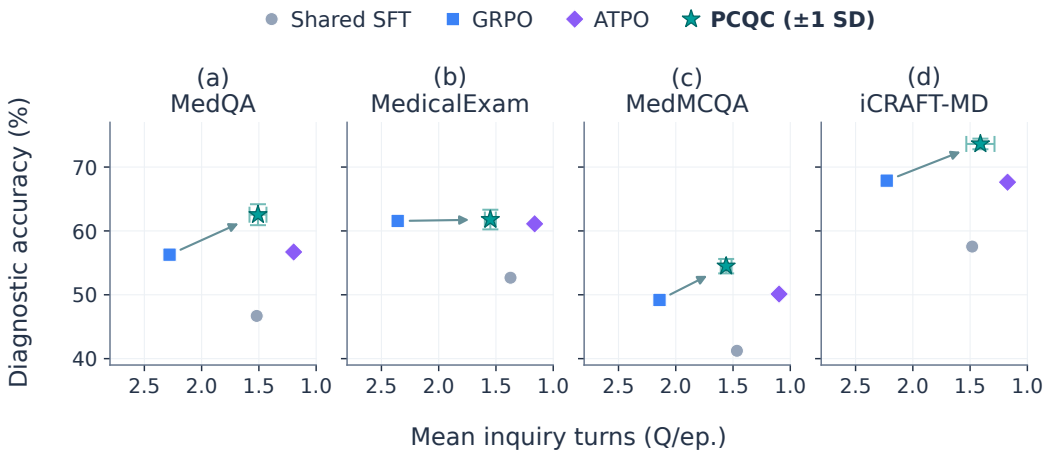

Figure 3: **Diagnostic accuracy versus inquiry efficiency across four benchmarks.** Arrows connect GRPO to PCQC; upward indicates higher diagnostic accuracy and rightward indicates fewer inquiry turns. RL results are means over three independent runs; error bars show one sample standard deviation for PCQC.

Evaluation and metrics. Each policy completes one consultation per evaluation case. We report *diagnostic accuracy*, the percentage of consultations with a correct final diagnosis, and *mean inquiry turns* (Q/ep.), the average number of non-final interaction attempts per consultation. For GRPO, ATPO, and PCQC, both metrics are reported as mean and sample standard deviation across three independent RL runs. All methods use the same interaction protocol; decoding, action parsing, and metric computation are detailed in Appendix A.3.

4.2 Diagnostic Performance and Inquiry Efficiency

PCQC reaches 63.10% mean accuracy across the four benchmarks (Table 1), improving over GRPO (58.72%) and ATPO (58.89%) by 4.38 and 4.21 percentage points, respectively. Question-level

Table 3: **Ablation of question-level supervision with matched terminal group size ($G_T = 4$).** Executed-local and PCQC use the same frozen diagnostic scorer but supervise the executed question and all candidates, respectively; Terminal-only uses no question-level objective. PCQC reports mean $\pm$ sample standard deviation over three independent RL runs.

	MedQA		MedicalExam		MedMCQA		Avg.
Method	Acc. (%) ↑	Q/ep.	Acc. (%) ↑	Q/ep.	Acc. (%) ↑	Q/ep.	Acc. (%) ↑
Terminal-only	52.92	2.241	49.33	2.273	45.90	2.047	49.38
Executed-local	57.18	2.142	56.67	2.220	48.88	2.183	54.24
PCQC	**62.54** $\pm$1.64	1.508 $\pm$0.074	**61.78** $\pm$1.54	1.549 $\pm$0.048	**54.48** $\pm$1.13	1.560 $\pm$0.047	**59.60** $\pm$1.25

counterfactual supervision therefore improves diagnosis beyond both terminal outcome optimization and tree-based action supervision.

PCQC's gain does not rely on longer consultations. Relative to Shared SFT, it raises mean accuracy by 13.56 percentage points while leaving mean inquiry turns nearly unchanged (1.460 versus 1.507). Compared with terminal GRPO, PCQC is 4.38 percentage points more accurate while using 33.1% fewer inquiry turns on average, with reductions of 27.2–36.7% across the four benchmarks (Table 2). Figure 3 summarizes this joint improvement: PCQC achieves higher mean diagnostic accuracy while using fewer inquiry turns than GRPO across all four benchmarks. Per-run results and additional interaction analyses are provided in Appendix D.1.

Generalization to independently sourced clinical cases. PCQC generalizes its performance–efficiency advantage to iCRAFT-MD, an independently sourced dermatology benchmark with no detected overlap against the SFT or RL training cases. It achieves $(73.62 \pm 0.83)\%$ diagnostic accuracy, compared with $(67.87 \pm 4.22)\%$ for GRPO and $(67.63 \pm 1.25)\%$ for ATPO, while using 36.7% fewer inquiry turns than GRPO (1.410 versus 2.228). This result demonstrates that the joint improvement carries to clinical cases outside the primary ATPO evaluation suite.

4.3 Value of Counterfactual Candidate Supervision

Same-state counterfactual credit outperforms executed-question feedback. We isolate the source of PCQC's gains with three configurations sharing the same terminal group size ($G_T = 4$ complete consultations per case). Terminal-only training uses final diagnostic correctness. Executed-local adds diagnostic gain from the frozen Qwen3-14B scorer but applies it only to the executed question. PCQC instead derives relative credit from same-state candidate groups and applies it to both executed and unexecuted questions. Because Executed-local retains the same candidate generation pathway and frozen scorer, this comparison isolates the value of same-state candidate supervision beyond simply adding scorer-based local feedback.

Executed-local raises mean diagnostic accuracy across MedQA, MedicalExam, and MedMCQA from 49.38% to 54.24%, confirming the value of question-level feedback. PCQC raises it further to $(59.60 \pm 1.25)\%$, exceeding Executed-local by 5.36 percentage points and Terminal-only by 10.22 percentage points. It also improves accuracy on every benchmark while using fewer inquiry turns than either control. These results show the value of learning from same-state alternatives beyond terminal correctness and executed-question feedback. Full objective definitions are provided in Appendices B.4–B.5 and A.2.

Direct supervision of unexecuted questions. Directly training on unexecuted candidates adds value beyond using them only to estimate relative credit. Same-State-Q1 and PCQC compute credit from the same $G_Q = 4$ candidate group, but only PCQC applies the question loss to the unexecuted candidate tokens. At the same training budget of 222 updates, PCQC raises four-benchmark mean diagnostic accuracy from 59.95% to 63.02% and reduces mean inquiry turns from 2.153 to 1.573. On MedQA, it reaches 61.59% accuracy versus 58.20% for Same-State-Q1, while reducing inquiry turns from 2.096 to 1.574. This comparison shows the additional benefit of applying credit

directly to unexecuted candidate tokens, beyond using those candidates only as a reference group. Appendix D.2 reports the full comparison.

4.4 Diagnostic Utility Predicts Downstream Success

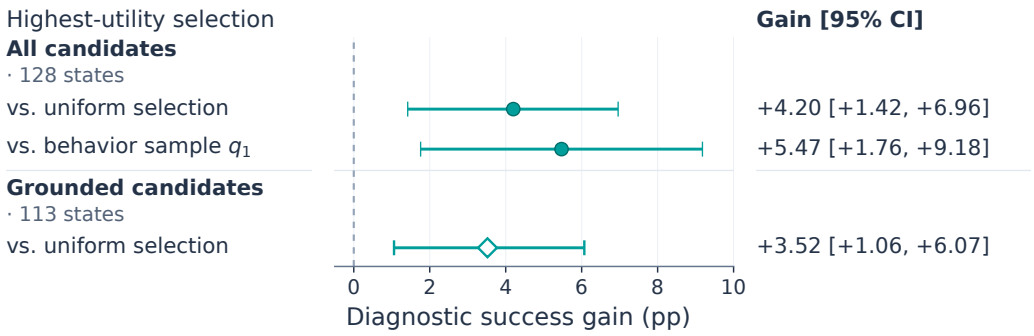

Figure 4: **Downstream diagnostic gains from highest-utility candidate selection.** Markers show gains in downstream diagnostic success (pp), with bars denoting paired 95% case-bootstrap confidence intervals. Success is estimated from eight continuations per candidate under PCQC Run A; results are shown for all candidates (128 MedQA states) and grounded candidates (113 states).

Candidate-selection evaluation. We test whether diagnostic utility from a single question–answer interaction predicts downstream value. We sample four initial-state candidate questions for each of 128 MedQA cases, answer and score each candidate, and generate eight independent continuations, yielding 4,096 trajectories. A candidate's *downstream diagnostic success* is the fraction of continuations reaching the correct diagnosis. We compare highest-utility selection with uniform selection and the executed question q_1, measuring the gain beyond the policy's own sampling. Selection rules and paired analyses are detailed in Appendix D.3.

Downstream diagnostic success. Selecting the highest-utility candidate (Figure 4) yields 65.14% downstream diagnostic success, compared with 60.94% under uniform selection, an improvement of 4.20 percentage points (95% CI $[+1.42, +6.96]$). It also exceeds the 59.67% success rate obtained from the executed question by 5.47 percentage points (95% CI $[+1.76, +9.18]$). Thus, diagnostic utility computed from a single question–answer interaction identifies candidates that are more likely to support successful downstream diagnosis.

Diagnostic utility remains predictive with grounded answers. We further restrict the comparison to answers that are supported by patient facts, confined to the requested information, and free of decision-target disclosure. Among the 113 cases with at least two such candidates, highest-utility selection improves downstream diagnostic success by 3.52 percentage points over uniform selection (95% CI $[+1.06, +6.07]$). The utility–success relationship therefore holds when selection is restricted to grounded candidate interactions. Appendix D.4 provides the full analysis.

5 Conclusion

We introduced Privileged Counterfactual Question Credit (PCQC), which turns *questions never asked* into direct supervision for medical dialogue policies. Using privileged patient facts, PCQC compares candidate question–answer interactions at the same dialogue state and assigns relative credit to both executed and unexecuted questions without full counterfactual rollouts. PCQC reaches 63.10% mean diagnostic accuracy, outperforming GRPO and ATPO by more than four percentage points while using 33.1% fewer inquiry turns than GRPO. Same-scorer controls show additional gains from supervising unexecuted candidates, while continuation experiments confirm that diagnostic utility predicts downstream success. More broadly, when an environment's hidden state is available during training, the evidence revealed by unexecuted actions can directly supervise action selection without estimating a complete future return for every alternative.

Reproducibility Statement

The experimental setup and evaluation protocol are described in Section 4.1, with further details provided in the appendix, including model and training configurations, data provenance, sampling budgets, checkpoint selection, patient-responder and diagnostic-scorer prompts, information access, action parsing, question-token supervision, and statistical procedures.

To facilitate reproducibility, we will publicly release the training and evaluation code, experiment configurations, data-processing and evaluation adapters, model prompts, checkpoint-selection scripts, and analysis code used to reproduce the reported tables and figures. We will also release the derived evaluation artifacts and per-run outputs supporting the reported results where permitted by the licenses of the underlying datasets.

Ethics Statement

PCQC is studied in benchmark-based simulated medical dialogue and is not evaluated for clinical deployment. Our experiments use existing clinical benchmark cases and involve no prospective patient interaction or collection of new patient data. The reported results characterize diagnostic accuracy and inquiry behavior within these experimental settings and should not be interpreted as evidence of clinical safety or effectiveness. The appendix documents the information available to the doctor policy, patient responder, and diagnostic scorer.

AI Use Statement

Generative AI tools were used to assist with manuscript writing and editing, literature discovery, and implementation of data-processing, evaluation, and plotting scripts. They were also used as model judges in the answer-grounding analysis described in Appendix D.4. All AI-assisted outputs used in this work were reviewed by the authors, who take responsibility for the scientific claims, experimental results, and final manuscript.

References

Ruike Cao, Shaojie Bai, Fugen Yao, Liang Dong, Jian Xu, and Li Xiao. ATPO: Adaptive tree policy optimization for multi-turn medical dialogue. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.02216. arXiv:2603.02216.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 66–75. PMLR, 2020.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Xinke Jiang, Jinyang Zhang, Yinghao Zhu, Zheng Li, Liantao Ma, Junfeng Zhao, and Yasha Wang. ProMed: Shapley information gain guided reinforcement learning for proactive medical LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32481–32515. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1500. URL https://aclanthology.org/2026.acl-long.1500/.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46375–46408. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1544. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Yichun Feng, Jiawei Wang, Lu Zhou, and Yixue Li. DoctorAgent-RL: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue, 2025b. URL https://arxiv.org/abs/2505.19630v1. Preliminary version accepted at ICASSP 2026.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M. Van Allen, David Kim, Roxana Daneshjou, and Pranav Rajpurkar. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31:77–86, 2025. doi: 10.1038/s41591-024-03328-5.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590. PMLR, 2025.

Fanqi Kong, Jiayi Zhang, Mingyi Deng, Chenglin Wu, Yuyu Luo, and Bang Liu. InfoPO: Information-driven policy optimization for user-centric agents, 2026. URL https://arxiv.org/abs/2603.00656.

Yunghwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-R1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.04284. arXiv:2510.04284.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28858–28888. Curran Associates, Inc., 2024. doi: 10.52202/079017-0908. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/32b80425554e081204e5988ab1c97e9a-Abstract-Conference.html.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Pedram Hosseini, Bryceton G. Thomas, Jessica M. Sin, Bing Ren, Jonathan S. Ilgen, Yulia Tsvetkov, and Maarten Sap. ALFA: Aligning LLMs to ask good questions a case study in clinical reasoning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=12u7diwku0.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Robotics: Science and Systems XIV*, 2018. doi: 10.15607/RSS.2018.XIV.008.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9:499, 2026. doi: 10.1038/s41746-026-02674-7.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943–950, 2025. doi: 10.1038/s41591-024-03423-7.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642:442–450, 2025. doi: 10.1038/s41586-025-08866-7.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qkWP6phrvZ.

Yutao Xie, Nathaniel Thomas, Nicklas Hansen, Yang Fu, Li Erran Li, and Xiaolong Wang. TIPS: Turn-level information-potential reward shaping for search-augmented LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.22293. arXiv:2603.22293.

Yijun Zhang, Fan Xu, Jiaxin Ding, Yule Xie, Shiqing Gao, Xin Ding, Haoxiang Zhang, Luoyi Fu, and Xinbing Wang. Information gain-based rollout policy optimization: An adaptive tree-structured rollout approach for multi-turn LLM agents, 2026. URL https://arxiv.org/abs/2607.06223.

A Datasets and Experimental Setup

A.1 Datasets

Training data. All methods use the released ATPO collection (Cao et al., 2026): 4,157 conversations for supervised initialization and 14,256 RL cases, comprising 9,400 MediQ/MedQA and 4,856 MedMCQA cases. Conversion preserves the source messages and target responses.

Evaluation data. The ATPO dialogue sets contain 1,268 MedQA, 150 MedicalExam, and 536 MedMCQA cases. We use the complete released MedQA and MedicalExam test sets and the released MedMCQA held-out split. Conversion preserves initial patient information, decision questions, options, labels, and atomic patient facts. iCRAFT-MD supplies an independent dermatology case source released with MediQ (Li et al., 2024), based on CRAFT-MD (Johri et al., 2025). We retain 139 of its 140 cases, excluding one with conflicting answer text and option index before evaluation. The 139-case set and all evaluated policy checkpoints are fixed before iCRAFT-MD evaluation, which uses the same patient responder and interaction protocol.

Training–evaluation overlap. No evaluation case is an exact normalized full-prompt match to the RL training set. A broader component-level audit identifies two exact MedQA case-and-problem matches and lexical near-duplicate candidates spanning 50 cases across the three ATPO evaluation sets. For iCRAFT-MD, comparisons against both the SFT conversations and RL cases find no exact or thresholded lexical matches (character 5-gram cosine 0.85 and word 1–2 gram cosine 0.90).

A.2 Training Configuration

Table 4 summarizes the models and sampling settings.

Table 4: **Core experimental configuration.** Model identities and sampling groups for the reported comparisons.

Item	Setting
Actor backbone	Qwen3-8B
Actor initialization	Shared protocol-SFT, 3 epochs / step 96
Patient responder	Frozen Qwen3-8B protocol-SFT, 1 epoch / step 32
PCQC diagnostic scorer	Frozen Qwen3-14B
Maximum dialogue turns	10, including the final answer
Cases per RL update	128
GRPO terminal group G_T	32
PCQC terminal group G_T	4
PCQC question group G_Q	4; question states in the first consultation trajectory
PCQC question-loss weight β	1
Independent RL training runs	GRPO, ATPO, and PCQC: three per method

Supervised initialization. The shared Qwen3-8B actor receives assistant-token supervision for three epochs (96 updates), with BF16 precision, batch size 128, learning rate 10^{-5}, and weight decay 0.01. The frozen patient responder is a separate Qwen3-8B checkpoint trained for one epoch (32 updates) on the same conversations.

Policy optimization. GRPO and PCQC use full-parameter BF16 training, learning rate 10^{-6}, weight decay 0.01, one PPO epoch per update, clipping ratio 0.2, token-mean losses, and reference-policy KL coefficient 0.001. PCQC uses AdamW with a constant learning rate and gradient clipping at 1.0. Terminal rewards are binary diagnostic correctness; PCQC additionally optimizes question credit with $\beta = 1$. Candidate construction and token masks are specified in Appendix B.4.

ATPO follows the U1+U2 configuration (Cao et al., 2026), with tree budget $M = 128$, four candidates, variance threshold 1.2, and pruning. Its critic uses learning rate 10^{-5}, four warmup updates, and GAE with $\gamma = \lambda = 1$. Actor learning uses rate 10^{-6}, KL coefficient 0.01, clipping bounds $0.20/0.28$, and per-sequence token-mean losses. Rewards are $+3$ for a correct final answer, 0 for an incorrect valid answer, and -1 without a valid final answer.

Development and evaluation protocol. We adopt the MedQA dialogue split configured as the validation stream in ATPO's released RL pipeline (Cao et al., 2026). This split originates from the MedQA test split released with MediQ (Li et al., 2024) and is used for method development and checkpoint selection. The selected checkpoints are then evaluated on MedicalExam, MedMCQA, and iCRAFT-MD without further checkpoint selection. GRPO, ATPO, and PCQC each use three independent RL training runs for the main comparison. PCQC Run A anchors the matched question-supervision, continuation, and full-system cost analyses, each of which requires a single fixed policy checkpoint and per-case artifacts.

Question-supervision controls. The controls in Table 3 retain PCQC's initialization, data, and $G_T = 4$. Terminal-only sets $\beta = 0$ and uses the checkpoint at update 75. Executed-local retains the same responder, scorer, $G_Q = 4$, $\beta = 1$, and Run A training configuration; the common selection criterion selects its update-100 checkpoint from 222 training updates. All candidates are generated, answered, and scored, but only the executed question receives local supervision. Appendices B.5 and D.2 define Executed-local credit and Same-State-Q1, respectively.

A.3 Evaluation and Metrics

Each policy produces one consultation per case, using the complete option set, including MedicalExam's 4–8 options. Actor decoding uses temperature 1.0, top-p 0.8, no top-k truncation, thinking enabled, and a 512-token action limit. Prompt and rollout-response limits are 1,024 and 4,096 tokens. Case-level sampling schedules are shared across methods. A valid final answer takes precedence over a question and must belong to the case's option set.

Diagnostic accuracy measures correct final labels; missing or invalid answers count as incorrect. Evaluation does not use the diagnostic scorer. Q/ep. is the mean number of non-final interaction attempts per consultation. Mean actor-generated tokens measure generation length.

Statistical analysis. For GRPO, ATPO, and PCQC, means and sample standard deviations of diagnostic accuracy and inquiry turns summarize three RL runs. Paired 95% confidence intervals use 10,000 percentile bootstrap resamples of cases, preserving both methods' outcomes within each case. Exact two-sided McNemar tests use discordant predictions. These intervals describe case-level uncertainty at fixed checkpoints.

A.4 Full-System Training Cost

GPU-hours in Table 5 equal recorded RL training time multiplied by all eight allocated H800 GPUs through the reported checkpoint, including policy optimization and rollout, patient responses, and diagnostic scoring. Shared SFT and separate benchmark evaluations are excluded.

On MedQA, PCQC Run A ($G_T = 4$) achieves 61.59% diagnostic accuracy at 70.62 H800 GPU-hours, compared with 56.07% at 72.68 GPU-hours for terminal GRPO ($G_T = 32$). The 5.52-percentage-point improvement comes with 44.5% fewer complete consultation trajectories: 113,664 versus 204,800.

Table 5: **PCQC improves MedQA accuracy by 5.52 percentage points at comparable full-system training cost while using 44.5% fewer complete consultation trajectories.** The comparison uses PCQC Run A with $G_T = 4$ and GRPO with $G_T = 32$. GPU-hours include all allocated devices for policy training, patient responses, and diagnostic scoring.

Method	Acc. (%) ↑	H800 GPU-hours	Complete trajectories	Q/ep.
GRPO	56.07	72.68	204,800	2.483
PCQC	**61.59**	**70.62**	**113,664**	**1.574**

B Question-Credit Implementation

B.1 Information Access

Table 6 summarizes the inputs to the policy, patient responder, and diagnostic scorer. The correct label indexes the scorer's output probability when constructing credit.

Table 6: **Information access during question-credit construction.** Patient facts supply candidate answers; the scorer evaluates observed question–answer evidence. The correct label indexes its output probabilities.

Component	O_t	q_i	Patient facts z	Task	y^*
Actor	Yes	Generates	No	Yes	No
Patient responder	No	Yes	Yes	No	No
Diagnostic scorer	Yes	Yes	No	Yes	No
Utility extraction	—	—	—	—	Index

Task denotes the clinical decision question and answer options. The patient responder receives the patient facts and candidate question; the scorer receives $O_t \oplus (q_i, a_i)$ and the clinical task. After scoring, y^* selects the correct-option probability.

B.2 Interaction Protocol and Prompts

The actor receives initial patient information, the clinical decision question, answer options, and the observed dialogue. It responds with `Question: [question text]` or `Final Answer: [option label]`.

The frozen Qwen3-8B patient responder receives the question and fixed patient facts, using the following instruction:

```
You are a medical information assistant. Your role is to help doctors by providing
    information strictly from patient data.

INSTRUCTIONS:
1. Search through the provided atomic facts for information that directly answers the doctor'
    s question
2. If you find relevant atomic facts, provide the answer using ONLY that information
3. Do NOT add any medical analysis, inference, interpretation, or external knowledge
4. Do NOT make assumptions or draw conclusions beyond what is explicitly stated
5. If no atomic fact directly answers the question, respond with exactly this phrase: "The
    patient cannot answer this question."

Patient atomic facts:
{atomic facts joined by newline}

Doctor's question:
{doctor question}

Your response:
```

Patient decoding uses temperature 0.8, top-p 1.0, a 256-token limit, and disabled thinking. The generated response supplies the patient answer.

B.3 Diagnostic Utility Construction

The frozen Qwen3-14B scorer receives the initial information, observed exchanges, and the candidate question–answer pair in the clinical-evidence field below. Each exchange uses the fields `Doctor question:` and `Patient response:`.

```
Clinical evidence:
{initial information and observed exchanges}
Question: {clinical decision question}
Options:
{label}. {option text}
...

Answer:
```

Let $\mathcal{Y}$ denote the answer-option labels. For each appended label $c \in \mathcal{Y}$, the scorer reads its final token log probability ℓ_c and normalizes across options:

$$s_\phi(c \mid O) = \frac{\exp(\ell_c)}{\sum_{d \in \mathcal{Y}} \exp(\ell_d)}. \tag{13}$$

Our training cases use $\mathcal{Y} = \{A, B, C, D\}$. The correct-option probability defines diagnostic utility. The implementation subtracts the shared pre-exchange probability $b(O_t)$ before normalization; this gives the same relative question credit as Equation 5, as shown in Appendix C.

B.4 Candidate Construction and Question-Token Optimization

Actions and rewards. The actor either asks a question or provides a final answer. A correct valid final answer receives terminal reward 1; other terminal outcomes receive 0.

Candidate groups. The first of the $G_T = 4$ consultations per case supplies supervised states. The actor samples $G_Q = 4$ responses at each state, retaining a group when the first response parses as a question. The same (q_1, a_1) pair is used for scoring and continuing the consultation; the other candidates do not advance it. Utilities are normalized across the sampled responses. A response without a parsed question supplies no question text to the responder.

Question-token masks. Question credit applies to actor tokens following the first `Question:` marker; responses without this marker have a zero question mask. The marker, preceding tokens, and patient answers receive no question credit. Terminal and question losses are normalized separately by their selected token counts across the update batch, then combined at $\beta = 1$ with reference-policy KL regularization.

B.5 Executed-Question Local Credit

Executed-local retains PCQC's candidate generation and frozen scorer. For every valid executed-question event e in the current update, its diagnostic gain is

$$d_e = s_\phi(y_e^* \mid O_e \oplus (q_{e,1}, a_{e,1})) - s_\phi(y_e^* \mid O_e). \tag{14}$$

The gains are standardized across these events, including different cases and dialogue states, using the same sample-standard-deviation threshold and normalization constant as PCQC. Credit is zero for fewer than two events or standard deviation below 10^{-4}. It supervises executed-question tokens with the mask above. Unexecuted candidates are still generated, answered, and scored, but receive no local gradient and do not enter the normalization used by Executed-local. The clipped local objective follows Equation 11, normalized by its own selected executed-question tokens, and is combined with terminal GRPO at $\beta = 1$.

C Properties of Relative Question Credit

Shared-baseline invariance. For a fixed dialogue state, let $\Delta_i = u_i - b(O_t)$ be the diagnostic gain of candidate i. Subtracting the shared baseline shifts the mean but preserves the sample variance:

$$\begin{aligned} \bar{\Delta} &= \mu_Q - b(O_t), \\ \sigma_\Delta^2 &= \frac{1}{K-1} \sum_{i=1}^{K} (\Delta_i - \bar{\Delta})^2 = \sigma_Q^2. \end{aligned} \tag{15}$$

Utilities and gains therefore give identical threshold decisions and normalized credit. For $\sigma_Q \geq \tau$,

$$\begin{aligned} A_i^Q &= \frac{\Delta_i - \bar{\Delta}}{\sigma_Q + \epsilon_{\text{norm}}} \\ &= \frac{u_i - K^{-1} \sum_{j=1}^{K} u_j}{\sigma_Q + \epsilon_{\text{norm}}} \\ &= \frac{1}{K(\sigma_Q + \epsilon_{\text{norm}})} \sum_{j=1}^{K} (u_i - u_j). \end{aligned} \tag{16}$$

This is the normalized mean of pairwise utility differences; positive credit indicates an above-average diagnostic gain.

Correct-option cross-entropy and surprisal. Let δ_{y^*} be the point-mass distribution on the correct option. For positive correct-option probabilities,

$$\begin{aligned} h_\phi(O) &= H(\delta_{y^*}, s_\phi(\cdot \mid O)) = -\log s_\phi(y^* \mid O) \\ &= D_{\mathrm{KL}}(\delta_{y^*} \,\|\, s_\phi(\cdot \mid O)), \end{aligned} \tag{17}$$

where $H(\delta_{y^*}) = 0$ gives the equality with KL divergence.

Taking $b(O_t) = s_\phi(y^* \mid O_t) > 0$, define the correct-option surprisal reduction after the candidate question–answer exchange as

$$\begin{aligned} g_i &= h_\phi(O_t) - h_\phi(O_t \oplus (q_i, a_i)) \\ &= \log u_i - \log b(O_t) = \log \frac{u_i}{b(O_t)}. \end{aligned} \tag{18}$$

Equivalently, $u_i = b(O_t)e^{g_i}$, so the diagnostic gain satisfies

$$\Delta_i = b(O_t)(e^{g_i} - 1), \qquad \frac{\mathrm{d}\Delta_i}{\mathrm{d}g_i} = b(O_t)e^{g_i} > 0. \tag{19}$$

Thus, for $\sigma_Q \geq \tau$, group normalization preserves the ordering by correct-option surprisal reduction:

$$A_i^Q > A_j^Q \iff u_i > u_j \iff g_i > g_j. \tag{20}$$

Credit magnitudes are determined by normalized probability-scale diagnostic gains.

D Additional Results and Question-Selection Analyses

D.1 Benchmark Results and Interaction Metrics

Independent RL training runs. Table 7 reports diagnostic accuracy for the three update-222 PCQC checkpoints under the same case-level decoding schedule. The main comparison averages all four benchmarks, yielding 63.10% mean diagnostic accuracy. For the question-supervision comparisons on MedQA, MedicalExam, and MedMCQA, the per-run means are 59.33%, 58.50%, and 60.96%, yielding $(59.60 \pm 1.25)\%$. Within each averaging scope, we first average benchmark accuracies for each run and then compute the across-run mean and sample standard deviation.

Table 7: **Diagnostic accuracy (%) across independent PCQC training runs.** All update-222 checkpoints preserve the same performance pattern across the three primary benchmarks and the independent iCRAFT-MD evaluation.

Benchmark	Run A	Run B	Run C	Mean ± sample SD
MedQA	61.59	61.59	64.43	62.54 ± 1.64
MedicalExam	62.67	60.00	62.67	61.78 ± 1.54
MedMCQA	53.73	53.92	55.78	54.48 ± 1.13
Mean (3 benchmarks)	59.33	58.50	60.96	$\mathbf{59.60 \pm 1.25}$
iCRAFT-MD	74.10	74.10	72.66	73.62 ± 0.83
Mean (4 benchmarks)	63.02	62.40	63.89	$\mathbf{63.10 \pm 0.74}$

Generation length. Table 8 supplements the inquiry results in Table 2 with mean actor-generated tokens per consultation. Using fixed checkpoints for both methods, PCQC Run A reduces generation length by 33.7–37.6% relative to GRPO across MedQA, MedicalExam, and MedMCQA, complementing the turn-level reductions in the main results.

Table 8: **PCQC Run A reduces mean actor-generated tokens by 33.7–37.6% relative to GRPO.** The comparison uses the fixed GRPO and PCQC Run A checkpoints.

Benchmark	GRPO	PCQC	Reduction
MedQA	473.0	313.7	33.7%
MedicalExam	483.1	301.5	37.6%
MedMCQA	459.1	304.1	33.8%

Question-supervision controls. Table 9 reports the paired MedQA comparisons using PCQC Run A and the controls in Table 3. The comparisons use the same 1,268 cases and the paired statistical protocol in Appendix A.3. PCQC also shortens generation relative to Executed-local: mean actor-generated tokens decrease from 353.85 to 321.06 on MedQA, 348.67 to 321.43 on MedicalExam, and 350.85 to 317.93 on MedMCQA, corresponding to reductions of 9.3%, 7.8%, and 9.4%, respectively. PCQC values are means over three independent runs.

Table 9: **Paired MedQA comparisons for question supervision.** The PCQC comparison uses Run A.

Comparison	Δ Acc. (pp)	Paired 95% CI	McNemar p
PCQC − Executed-local	$+4.42$	$[+1.26, +7.49]$	0.0067
Executed-local − Terminal-only ($\beta = 0$)	$+4.26$	$[+1.10, +7.33]$	0.0083

D.2 Decomposing Candidate Question Supervision

Same-State-Q1 retains all four candidates when computing relative question credit but directly supervises only the executed question. Splitting Equation 11 by executed and unexecuted tokens gives $\mathcal{L}_{\text{question}} = \mathcal{L}_Q^{\text{exec}} + \mathcal{L}_Q^{\text{unexec}}$. The control keeps the executed-question terms and averages them over their token count N_{exec}; PCQC averages over all candidate question tokens N_Q. It shares PCQC Run A's initialization, case order, optimizer, sampling budgets, patient responder, and scorer, retaining $\beta = 1$, terminal GRPO, and the original KL objective. Table 10 compares both methods at 222 training updates across all four benchmarks.

Table 10: **Candidate supervision at a matched training budget.** PCQC Run A and Same-State-Q1 each use one training run and 222 updates. The mean equally weights all four benchmarks.

	Accuracy (%)		Q/ep.	
Benchmark	PCQC	Same-State-Q1	PCQC	Same-State-Q1
MedQA	61.59	58.20	1.574	2.096
MedicalExam	62.67	62.67	1.580	2.207
MedMCQA	53.73	51.31	1.590	2.136
iCRAFT-MD	74.10	67.63	1.547	2.173
Mean	**63.02**	59.95	**1.573**	2.153

D.3 Downstream Evaluation of Candidate Questions

The study uses 128 MedQA initial states, each with four valid candidate questions selected before continuation generation. The frozen PCQC Run A policy generates both the candidates and eight continuations from each fixed question–answer pair, yielding 4,096 trajectories. A candidate's downstream diagnostic success is its fraction of correct final diagnoses. We compare the highest-utility candidate, the executed question q_1, and uniform selection, whose expected success averages all four candidates. The paired bootstrap in Appendix A.3 resamples cases after aggregating each candidate's eight outcomes.

D.4 Answer-Grounded Continuation Analysis

We evaluate selection rules on candidate answers that are supported by patient facts, confined to the requested information, and free of decision-target disclosure.

Manual review of prioritized candidates and model-judge agreement for the remainder identify 395 eligible candidates. Selection comparisons reuse the original eight continuations per candidate. Among 113 cases with at least two eligible candidates, highest-utility selection improves downstream diagnostic success by 3.52 percentage points over uniform selection from the same eligible set (95% CI $[+1.06, +6.07]$). Table 11 also compares selection with the executed question q_1 on the corresponding eligible case sets.

Table 11: **Question selection with grounded answers.** Success uses eight continuations per candidate under PCQC Run A; Δ is selected minus comparator success (pp). Eligibility is defined in Appendix D.4.

Comparison	N cases	Selected success (%)	Comparator success (%)	Δ (pp)	95% CI
Original highest-utility selection vs. executed question q_1	108	66.44	61.11	+5.32	$[+2.08, +8.68]$
Highest-utility vs. uniform selection (eligible candidates)	113	65.82	62.30	+3.52	$[+1.06, +6.07]$
Highest-utility selection vs. executed question q_1 (eligible candidates)	104	67.07	61.42	+5.65	$[+2.28, +9.13]$

Row 1 preserves the original highest-utility selection and executed question q_1 and includes cases where both candidates are eligible. Row 2 includes cases with at least two eligible candidates; uniform selection is the mean success of all candidates in that eligible set. Row 3 further requires the executed question q_1 to be eligible and compares the highest-utility eligible candidate with q_1.
The paired bootstrap in Appendix A.3 resamples the eligible cases in each row, with one dialogue state per case. Success rates and differences are rounded independently.