RPRO: Ranked Preference Reinforcement Optimization for Enhancing Medical QA and Diagnostic Reasoning

Chia-Hsuan Hsu[1,2], **Jun-En Ding**[3,2], **Hsin-Ling Hsu**[4,2],
Chun-Chieh Liao[3,2], **Fang-Ming Hung**[2], **Feng Liu**[3]

[1]National Taiwan University of Science and Technology, [2]Far Eastern Memorial Hospital,
[3]Stevens Institute of Technology, [4]National Chengchi University

winnie920924@gmail.com, jding17@stevens.edu, hsinling.hsu@nccu.edu.tw,
fliu22@stevens.edu, philip@mail.femh.org.tw

Abstract

Medical question answering requires advanced reasoning that integrates domain knowledge with logical inference. However, existing large language models (LLMs) often generate reasoning chains that lack factual accuracy and clinical reliability. We propose Ranked Preference Reinforcement Optimization (RPRO), a novel framework that uniquely combines reinforcement learning with preference-driven reasoning refinement to enhance clinical chain-of-thought (CoT) performance. RPRO differentiates itself from prior approaches by employing task-adaptive reasoning templates and a probabilistic evaluation mechanism that aligns outputs with established clinical workflows, while automatically identifying and correcting low-quality reasoning chains. Unlike traditional pairwise preference methods, RPRO introduces a groupwise ranking optimization based on the Bradley–Terry model and incorporates KL-divergence regularization for stable training. Experiments on PubMedQA and MedQA-USMLE show consistent improvements over strong baselines. Remarkably, our 1.1B parameter model outperforms much larger 7B–13B models, including medical-specialized variants. These findings demonstrate that combining preference optimization with quality-driven refinement offers a scalable and effective approach to building more reliable, clinically grounded medical LLMs.

1 Introduction

Recent advances in large language models (LLMs) signal significant promise for medical applications, especially in clinical reasoning and diagnostic support systems Singhal et al. [2023a], Thirunavukarasu et al. [2023]. However, deploying AI in medicine requires exceptional reliability, making factual accuracy and thorough coverage of the data paramount. Medical reasoning introduces challenges distinct from those in general domains: clinical decisions require a comprehensive consideration of relevant factors, strict factual accuracy, and the avoidance of redundant information that could obscure critical insights.

Current approaches to medical reasoning primarily rely on chain-of-thought (CoT) prompting techniques Wei et al. [2022a], Kojima et al. [2022], which have shown effectiveness in improving reasoning capabilities across various domains. However, these methods typically generate single reasoning chains without systematic quality assessment or iterative refinement mechanisms. Moreover, existing evaluation frameworks often rely on overly simplistic binary metrics that fail to capture the

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

nuanced quality required for medical reasoning or require extensive human annotation, limiting their suitability for automated quality control where domain expertise is essential.

Recent work in preference learning and reinforcement learning from human feedback (RLHF) Ouyang et al. [2022], Christiano et al. [2017] shows promise in aligning AI systems with human preferences. Direct Preference Optimization (DPO) Rafailov et al. [2023] has shown promising results by learning from preference data. However, these methods typically rely on pairwise comparisons, which fail to capture all the ranking information in human preference datasets. Many real-world scenarios involve humans ranking multiple candidates, not just providing binary choices. For example, when they evaluate reasoning chains, mathematical solutions, or creative content, evaluators assign rankings that reflect nuanced quality differences. Reducing this rich data to pairwise comparisons can lose valuable training signals and lead to suboptimal solutions.

Specifically, these approaches face significant challenges in medical domains. First, the requirement for extensive expert annotation is prohibitively expensive. Additionally, preference judgments often lack the granularity needed for medical reasoning assessment. Furthermore, generic preference models fail to capture the domain-specific quality dimensions that are critical in healthcare applications.

To address these limitations, we introduce a novel framework that enhances medical reasoning through probabilistic refinement for quality assessment and RPRO (Ranked Preference Reinforcement Optimization), which optimizes over ranked groups of candidates using the Bradley-Terry model. We combine domain-adaptive reasoning templates with automated quality assessment and targeted revision mechanisms. Unlike existing methods, which use single reasoning chains or expensive human evaluation, our framework automatically generates multiple reasoning candidates. It evaluates their quality using probabilistic metrics and performs targeted improvements when needed. Our approach incorporates several key innovations:

- A probabilistic framework for automated medical reasoning quality assessment that captures multi-dimensional evaluation criteria through conditional probability modeling;
- Domain-specific CoT enhancement templates that adapt reasoning structures to different medical contexts;
- A novel linear reward learning framework that provides a stable policy gradient for ranked preference optimization;
- A method for generating high-quality preference data suitable for Ranked Preference Reinforcement Optimization (RPRO) fine-tuning, enabling continual improvement of medical reasoning capabilities.

2 Related Work

2.1 Medical LLMs: QA, Diagnosis, and EHR

Large language models (LLMs) have been used for medical tasks like knowledge-based question answering, diagnostic reasoning on clinical vignettes, and patient-specific reasoning over electronic health records (EHR) [Jin et al., 2019, 2021b, Shi et al., 2024, Zakka et al., 2024]. Domain-structured prompts and chain-of-thought (CoT) approaches consistently improve auditability and reduce unsupported claims by clearly separating decomposition and background from justification or diagnosis [Wei et al., 2022b, Singhal et al., 2023d,c]. In patient-specific contexts, retrieval-augmented generation (RAG) and citation-style evidence tracing help ensure faithfulness to chart data and guidelines Lewis et al. [2020], Shi et al. [2024], Zakka et al. [2024]. Recent work extends beyond answer accuracy, emphasizing multidimensional evaluation such as coverage of relevant findings and alignment with context, thereby better matching clinical expectations for completeness and correctness. In this work, we use domain-specific CoT templates (QA and diagnosis), automatically score rationales on coverage, factuality, and redundancy with an LLM judge, and make light score-guided revisions.

2.2 Reinforcement-Learned Medical LLMs

Reinforcement learning from human feedback (RLHF) and related preference-based methods have become central to aligning LLM behavior beyond supervised fine-tuning Christiano et al. [2017],

Ziegler et al. [2019], Ouyang et al. [2022]. In reasoning-heavy domains, variants that regularize to a reference policy while optimizing rewards derived from preference ordering or task-specific outcomes have shown improved stability and faithfulness.

In the medical domain, reinforcement-learned LLMs have been applied to tasks such as clinical question answering and diagnostic reasoning Singhal et al. [2023b], Nori et al. [2023]. These methods typically employ reinforcement learning techniques such as policy optimization with preference feedback Christiano et al. [2017], Ouyang et al. [2022] to guide models toward multi-step clinical reasoning, integration of biomedical knowledge, and reliable elimination of alternative diagnoses. Nevertheless, prior studies also emphasize persistent challenges in maintaining stability and factual grounding in high-stakes clinical scenarios, indicating that preference-optimized training plays an important role in reducing hallucination risks Singhal et al. [2023b].

Building on these advances, we introduce Ranked Preference Reinforcement Optimization (RPRO), a ranking-based extension of preference optimization. RPRO samples multiple responses per prompt and computes relative advantages, thereby reducing variance through group-level aggregation and providing more robust training signals for reasoning-intensive tasks like medical question answering.

2.3 Ranking-based Optimization in Clinical Reasoning

Recent studies have expanded preference alignment for LLMs from pairwise to listwise settings. Most methods focus on pairwise comparisons. Yet, listwise approaches have shown effectiveness in other domains Liu et al. [2025], Pesaran zadeh et al. [2025], Cai et al. [2025]. For example, ALMupQA by Yang et al. [2024] integrates multi-perspective ranking alignment for code QA. IRPO, introduced by Wu et al. [2025], is a framework that incorporates graded relevance and positional importance. These works highlight the impact of ranking-based feedback on aligning LLM outputs with diverse user expectations. However, medical multiple-choice QA often needs more sophisticated explanatory discourse and clinical reasoning. In this work, we evaluate our framework on multiple medical reasoning benchmarks. We use a probabilistic refinement mechanism, generate preference data with RPRO, and combine linear rewards for ranked candidate optimization.

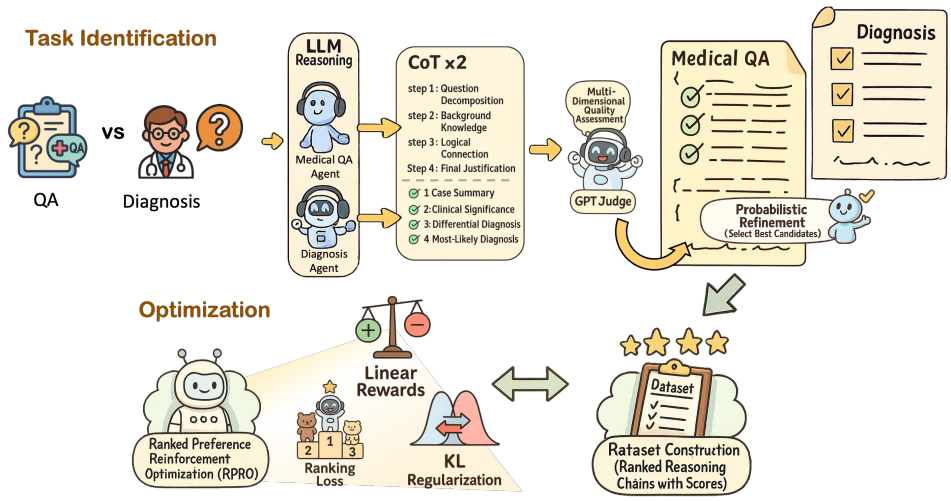

Figure 1: Overview of the proposed pipeline. The framework distinguishes between medical QA and diagnosis tasks, applies chain-of-thought (CoT) reasoning with multi-dimensional quality assessment and probabilistic refinement, and further improves through ranked preference reinforcement optimization (RPRO) with dataset construction.

3 Methodology

3.1 Overview

In this work, we provide an overview of a systematic framework for generating high-quality Chain-of-Thought (CoT) reasoning pairs optimized for Ranked Preference Reinforcement Optimization (RPRO) training in medical question answering. The framework follows a four-step pipeline (1) Based on medical task identification, we employ CoT reasoning for general medical Q&A and diagnostic tasks (2) For each chain instance, our process involves question decomposition, background knowledge integration, logical connection, and final justification for clinical assessment (3) We score each generated chain using probabilistic rules for assessment(4) After steps 1-3, we select high-quality reasoning chains and discard the lowest-quality onesThis results in a curated dataset for RPRO learning, enabling the model to learn from exemplary clinical reasoning patterns while avoiding reinforcement of suboptimal decision-making processes.

3.2 Problem Formulation

We begin by formulating medical questions into two complementary task types to enable domain-appropriate reasoning. We define τ_{QA} as general medical question-answering, which targets questions requiring factual medical knowledge and logical reasoning about biomedical concepts, treatments, or underlying mechanisms. Conversely, τ_{Diag} represents medical diagnosis tasks involving clinical scenarios that necessitate systematic diagnostic reasoning through detailed patient presentation analysis and comprehensive differential diagnosis consideration. This dual-task framework establishes our task space $\mathcal{T} = \{\tau_{\text{QA}}, \tau_{\text{Diag}}\}$. Let $\mathcal{X}$ and $\mathcal{Q}$ represent the spaces of structured medical records and natural-language prompts, respectively. Each task instance is an input pair $z = (x, q) \in \mathcal{X} \times \mathcal{Q}$, where x is a medical record and q is the corresponding prompt. A task classifier determines the appropriate task $\tau \in \mathcal{T}$ for each instance.

3.3 Task-Adaptive Chain-of-Thought Generation

Given z and τ, a conditional policy $\pi_\theta(\cdot \mid z, \tau)$ generates $K = 5$ candidate CoTs $C = \{c_k\}_{k=1}^{5}$ and $c_k \sim \pi_\theta(\cdot \mid z, \tau)$ for each medical question. Then, we present each candidate c_k four reasoning chain steps as:

$$c_k = \big(u_k^{(1)},\, u_k^{(2)},\, u_k^{(3)},\, u_k^{(4)}\big), \tag{1}$$

where the j-th component $u_k^{(j)}$ depends on the task category τ and is defined as

$$u_k^{(j)} = \begin{cases} s_k^{(j)}, & \text{if } \tau = \tau_{QA}, \\ t_k^{(j)}, & \text{if } \tau = \tau_{Diag}, \end{cases} \qquad j = 1, 2, 3, 4. \tag{2}$$

For $\tau = \tau_{QA}$, the components include four clinical factor $s^{(1)}$: Question Decomposition, $s^{(2)}$: Background Knowledge, $s^{(3)}$: Logical Connection and $s^{(4)}$: Final Justification to determine generated sequence medical QA chain is correct. Similarly, the $\tau = \tau_{Diag}$ focus on diagnosis task include $t^{(1)}$: Case Summary, $t^{(2)}$: Clinical Significance, $t^{(3)}$: Differential Diagnosis and $t^{(4)}$: Most Likely Diagnosis. Those criteria can better align the reasoning process with standard diagnostic procedures, thereby tailoring the CoT representation to the requirements of clinical decision-making. Thus, the unified formulation $c_k = (u_k^{(1)}, u_k^{(2)}, u_k^{(3)}, u_k^{(4)})$ provides a task-agnostic representation of CoT candidates, while retaining the ability to specialize the semantic meaning of each reasoning step according to the task label τ. After generation, the set of five candidates is reduced to $M = 4$ by retaining the top four according to subsequent evaluation and discarding the lowest-ranked one.

3.4 Medical Prompt Instruction and Reasoning

Previous research shows that effective prompt template instructions can improve model performance on CoT reasoning Zhu et al. [2024], Wei et al. [2022a], Kojima et al. [2022]. We use structured reasoning templates based on established clinical and biomedical reasoning patterns. Table 3 presents

a general medical QA template that applies a four-step reasoning process, using binary classification at each step according to the validation criteria from Section 3.3. Specifically, background knowledge retrieval gathers relevant biomedical facts and principles from medical literature. Logical connection links the retrieved knowledge to the question context. The justification step synthesizes the reasoning to support the conclusion.

In parallel, our diagnostic reasoning template mirrors the clinical diagnostic workflow used by medical professionals. The process begins with a case summary, which extracts and organizes key clinical findings from patient presentations. Clinical significance analysis then examines symptom patterns and diagnostic indicators to understand their medical relevance. Differential diagnosis systematically considers alternative diagnoses and establishes criteria for elimination. Finally, the most likely diagnosis step concludes with an evidence-supported diagnostic decision based on the available clinical information.

3.5 Reasoning Refinement Method

3.5.1 Multi-Dimensional Quality Assessment Evaluation

To evaluate reasoning quality across three key dimensions, we propose a joint-probability-based evaluation framework for assessing multiple candidate reasoning paths and selecting the optimal subset from $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$. For each candidate $c \in \mathcal{C}$, we define three scoring functions $S_i : \mathcal{C} \to \mathcal{S}$ where $i \in \{\text{cov}, \text{fact}, \text{red}\}$, corresponding to coverage $S_{\text{cov}}(c)$, factual accuracy $S_{\text{fact}}(c)$, and redundancy $S_{\text{red}}(c)$, each evaluated on a 5-point scale. These raw scores are mapped into probabilistic indicators $p_i(c) = \pi\big(S_i(c)\big)$ via a normalization mapping $\pi : \mathcal{S} \to [0, 1]$.

3.5.2 Probabilistic Reasoning Refinement

Unlike traditional additive scoring approaches, our objective is to simultaneously ensure adequate coverage of the source content and maintain factual accuracy, while also suppressing redundancy during generation. To achieve this, we formulate the decoding objective as the maximization of a multiplicative acceptance function that integrates these three criteria. Specifically, the each question of optimal candidate c^* is defined as

$$c^* = \arg\max_{c} \ \Big[p_{\text{cov}}(c) \cdot p_{\text{fact}}(c) \cdot \big(1 - p_{\text{red}}(c)\big) \Big], \tag{3}$$

where $p_{\text{cov}}(c)$ denotes the probability of coverage, $p_{\text{fact}}(c)$ denotes the probability of factual accuracy, and $p_{\text{red}}(c)$ measures the likelihood of redundancy in the generated content. Next, we introduce the selection operator with *acceptance function* $P_{\text{accept}}(c)$ as:

$$\Phi_k(\mathcal{C}) = \underset{c \in \mathcal{C}}{\text{Top-}k} \ P_{\text{accept}}(c), \tag{4}$$

where $\Phi_k(\mathcal{C})$ is extracts the top k candidates with the highest acceptance probabilities to be used for subsequent training or decision-making. Intuitively, we can effectively unifies multi-dimensional quality assessment into a probabilistic formulation and provides a mechanism for ranking and optimal subset selection in reasoning tasks.

4 Training Framework

4.1 Ranking Preference Optimization Algorithm

In this section, we propose RPRO, a flexible optimization method that differs from traditional approaches such as DPO and PPO-RLHF, which rely solely on pairwise comparisons (i.e., $c_1 \succ c_2$). Our proposed method leverages complete ranking information (i.e., $c_1 \succ c_2 \succ c_3 \succ \ldots \succ c_n$) to optimize model performance. Our goal is to learn a policy $\pi_\theta(\cdot \mid z, \tau)$ that assigns higher likelihood to preferred responses. For each candidate CoT sequence c_j given medical question z and temperature τ, we define the preference score as the conditional log-probability:

$$s_j = \log \pi_\theta(c_j \mid z, \tau) = \frac{1}{|c_j|} \sum_{t=1}^{|c_j|} \log \pi_\theta(c_{j,t} \mid z, \tau, c_{j,<t}) \tag{5}$$

where $|c_j|$ denotes the sequence length of the CoT, and $c_{j,t}$ represents the token at position t in the reasoning chain. Following the Bradley-Terry model Bradley and Terry [1952], the probability that CoT sequence c_i is preferred over sequence c_j is:

$$P(c_i \succ c_j \mid z, \tau) = \sigma\left(\frac{s_i - s_j}{\tau_{BT}}\right) = \frac{1}{1 + \exp\left(-\frac{s_i - s_j}{\tau_{BT}}\right)} \tag{6}$$

where $\tau_{BT} > 0$ is a Bradley-Terry temperature parameter (distinct from the generation temperature τ) controlling the sharpness of the preference distribution. Then, we can construct the ranking loss for the $K = 5$ candidate CoT sequences extends the Bradley-Terry model to handle complete expert rankings. The $\mathcal{L}_{CoT-rank}$ can be rewritten as:

$$\mathcal{L}_{CoT-rank} = \frac{1}{\binom{K}{2}} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \log\left(1 + \exp\left(-\frac{s_i - s_j}{\tau_{BT}}\right)\right). \tag{7}$$

4.2 Linear Rewards From Rank

GRPO is one of the widely used algorithms that employs entropy regularization to stabilize policy gradient–based learning processes. However, it still exhibits sensitivity to parameter tuning, which can lead to unstable convergence performance or even divergence during training Leahy et al. [2022], Schulman et al. [2017], Mnih et al. [2016]. To translate the complete ranking of candidate reasoning chains into effective training signals, we adopt a *linear reward shaping scheme*. We then define the advantage value a_j of candidate c_j as:

$$a_j = (K - r_j) - \frac{K-1}{2}, \quad j = 1, \ldots, K \quad \text{s.t.} \sum_{j=1}^{K} a_j = 0. \tag{8}$$

where each candidate is assigned a rank $r_j \in \{1, \ldots, K\}$, with $r_j = 1$ denoting the highest-quality candidate and $r_j = K$ the lowest. Intuitively, the linear reward can improve the model by incentivizing it to shift probability mass from poorly ranked reasoning paths toward higher-quality ones. Compared with pairwise-only preference optimization, RPRO efficiently incorporates full ranking information within each group of candidates, while retaining simplicity and stability during optimization. We then incorporate these linear rewards to increase the probability of higher-ranked candidates (positive a_j) and decrease the probability of lower-ranked ones (negative a_j) can be present as:

$$\mathcal{L}_{\text{Linear}} = -\frac{1}{K} \sum_{j=1}^{K} a_j \, s_j. \tag{9}$$

4.3 KL-Divergence Regularization for Medical Reasoning

To prevent the medical reasoning policy from deviating too far from a reference model $\pi_{ref}(\cdot \mid z, \tau)$, we incorporate KL-divergence regularization. The sequence-level KL divergence for CoT sequence c_j is:

$$\mathcal{D}_{KL}(c_j \mid z, \tau) = \frac{1}{|c_j|} \sum_{t=1}^{|c_j|} \mathcal{D}_{KL,t} \tag{10}$$

The KL regularization term across all K candidates is:

$$\mathcal{L}_{KL} = \beta \cdot \frac{1}{K} \sum_{k=1}^{K} \mathcal{D}_{KL}(c_k \mid z, \tau) \tag{11}$$

where $\beta \geq 0$ controls the regularization strength for medical reasoning stability. Here, the $\mathcal{D}_{KL,t}$ is token-level KL divergence at position t in the CoT sequence be expressed as:

$$D_{KL,t} = \sum_{v \in \mathcal{V}} \pi_\theta(v \mid z, \tau, c_{<t}) \log \frac{\pi_\theta(v \mid z, \tau, c_{<t})}{\pi_{ref}(v \mid z, \tau, c_{<t})} \tag{12}$$

where $\mathcal{V}$ is the set of all possible tokens in language model. During training, we combine the $\mathcal{L}_{CoT-rank}$ and $\mathcal{L}_{KL}$ to form $\mathcal{L}_{RPRO-CoT}$ so that the final objective becomes:

$$\mathcal{L}_{total} = \mathcal{L}_{RPRO-CoT} + \mathcal{L}_{KL} + \mathcal{L}_{\text{Linear}} \tag{13}$$

5 Experiments

5.1 Experimental Setting

Datasets. We evaluate our approach on two benchmark medical QA datasets. PubMedQA Jin et al. [2019] contains biomedical research questions derived from PubMed abstracts, requiring yes/no/maybe answers based on scientific evidence. MedQA-USMLE Jin et al. [2021a] consists of multiple-choice diagnostic questions from the United States Medical Licensing Examination, testing clinical reasoning and diagnostic capabilities. These datasets represent complementary aspects of medical reasoning: research-oriented knowledge synthesis and clinical diagnostic decision-making.

Implementation Details. Our base model is LLaMA 1.1B, chosen for computational efficiency while maintaining competitive performance. We generate five CoT reasoning variants per question and select the four highest-ranking candidates based on their acceptance probabilities for RPRO training. The acceptance threshold τ is set to 0.6 based on preliminary experiments. All models are trained with AdamW optimizer using a learning rate of 5e-5, batch size of 16, and 3 training epochs. Training is performed on NVIDIA A100 GPUs with mixed precision.

5.2 Baselines

We compare against several established language models across different scales and medical specializations. General-purpose models include Gemma 2B and 7B Mesnard et al. [2024], and Mistral 7B Jiang et al. [2023]. Medical-specialized models include MedAlpaca 7B and 13B Han et al. [2023], fine-tuned specifically for medical tasks, and BioMistral 7B Labrak et al. [2024], trained on biomedical corpora. These baselines represent the current state-of-the-art in both general and domain-specific language modeling for medical applications.

5.3 Comparison with DPO and SFT

We compare our RPRO approach against two established training paradigms. Supervised Fine-Tuning (SFT) employs standard next-token prediction on the highest-quality reasoning chains. DPO utilizes pairwise preference learning between high- and low-quality reasoning pairs. RPRO implements group-wise ranking optimization across multiple quality levels, enabling more nuanced preference learning from ranked reasoning chains.

5.4 Evaluation Metrics

We use two complementary metrics to assess model performance: accuracy, which measures the proportion of correctly answered questions for direct factual assessment, and macro F1 score, which computes the unweighted average of F1 scores across answer categories to ensure balanced evaluation and address class imbalance. This dual-metric approach captures both overall performance and category-specific reasoning quality.

5.5 Quantitative Results

Main Results. Table 1 presents the performance comparison across all evaluated models. Our method achieves substantial improvements over baseline models on both datasets. On PubMedQA, we obtain 61.95% accuracy and 50.88% Macro F1, outperforming the best baseline (Gemma 7B) by 3.87% and 2.86%, respectively. On MedQA-USMLE, our approach reaches 27.92% accuracy and 20.41% Macro F1, surpassing the strongest baseline (MedAlpaca 7B) by 3.91% accuracy. Remarkably, our

Table 1: Comparison of different models on PubMedQA and MedQA-USMLE. Best results are in bold.

Models	**Method**	**PubMedQA**		**MedQA-USMLE**	
		ACC (%)	Macro F1 (%)	ACC (%)	Macro F1 (%)
Gemma 2B	-	51.03	52.61	17.78	9.21
Gemma 7B	-	58.08	48.02	20.37	10.18
Mistral 7B	-	51.43	47.11	22.66	11.78
MedAlpaca 7B	-	44.83	42.41	24.01	16.37
BioMistral 7B	-	38.89	37.98	19.89	7.47
MedAlpaca 13B	-	52.61	46.16	23.15	**22.42**
Our Method (LLaMA 1.1B, RPRO)	RPRO	**61.95**	**50.88**	**27.92**	20.41

Table 2: Ablation study comparing different training methods and refinement strategies on PubMedQA and MedQA-USMLE. Best results are in bold.

Method	**PubMedQA**		**MedQA-USMLE**	
	ACC (%)	Macro F1 (%)	ACC (%)	Macro F1 (%)
SFT (with Refinement)	57.46	41.37	24.81	13.52
DPO (with Refinement)	60.90	41.92	25.83	15.65
RPRO (No Refinement)	61.43	38.11	23.33	13.27
Our Method: RPRO (with Refinement)	**61.95**	**50.88**	**27.92**	**20.41**

1.1B parameter model achieves better overall performance than much larger models, including 7B and 13B variants, demonstrating the effectiveness of our reasoning refinement approach.

Ablation Studies. Table 2 analyzes the contribution of key components in our framework. The comparison between "No Refinement + RPRO" and "RPRO (with Refinement)" reveals that reasoning refinement provides substantial improvements, particularly for Macro F1 (38.11% $\rightarrow$ 50.88%), while maintaining similar accuracy levels. This suggests that refinement enhances the quality of reasoning beyond mere correctness. Comparing different training paradigms, RPRO with refinement outperforms both SFT and DPO variants, validating the effectiveness of group-wise preference optimization for multi-candidate reasoning scenarios.

In Fig. 2, we investigate the impact of different regularization coefficients β on performance, revealing that moderate β values can enhance Macro F1 while maintaining accuracy, thereby achieving a balance between factual correctness and reasoning comprehensiveness. Excessively low β values cause the model to deviate from the reference distribution and lose stability, while excessively high β values over-constrain the model and suppress reasoning diversity. These findings validate the critical

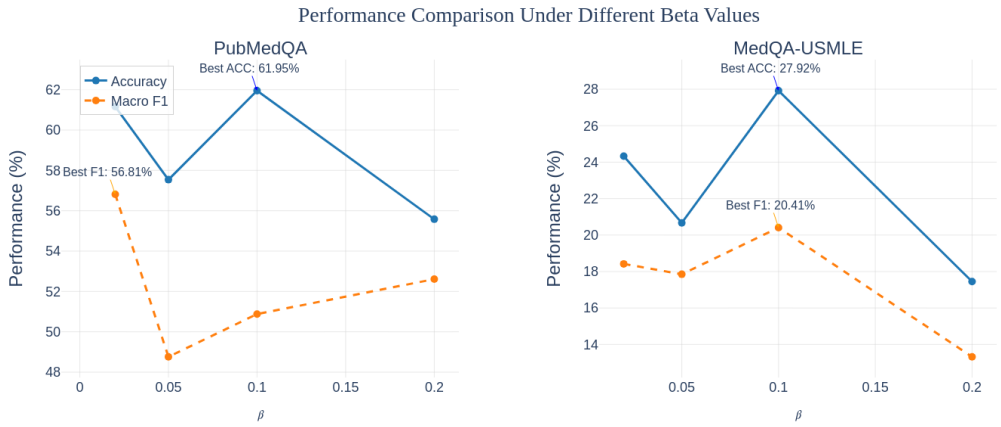

Figure 2: Performance comparison on PubMedQA and MedQA-USMLE under different β values. Accuracy (solid blue line) and Macro F1 (dashed orange line) are reported.

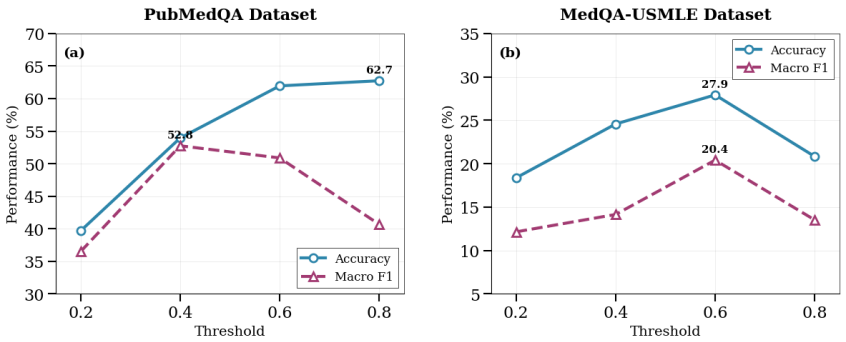

Figure 3: Performance across different confidence thresholds on (a) PubMedQA and (b) MedQA-USMLE datasets. Accuracy (solid blue line) and Macro F1 (dashed purple line) are reported.

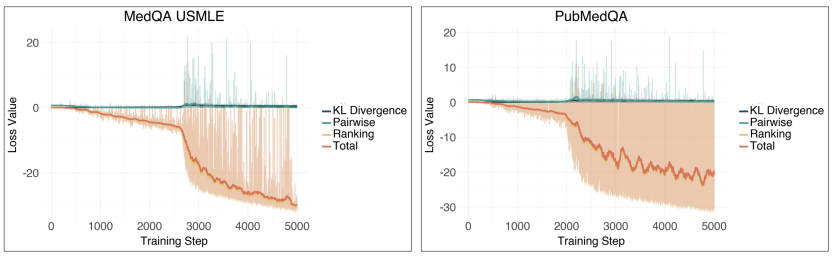

Figure 4: Training loss curves on MedQA-USMLE (left) and PubMedQA (right) datasets. The plots show KL divergence, pairwise loss, ranking loss, and total loss across training steps.

importance of appropriate KL regularization for enhancing robustness and generalization capabilities in clinical reasoning tasks.

Threshold and Convergency Analysis. Figure 3 examines the impact of the acceptance probability threshold τ on model performance. Lower thresholds (0.2, 0.4) result in excessive refinement, potentially degrading reasoning quality through over-correction. The optimal threshold of 0.6 balances the necessity for refinement with the preservation of reasoning. Higher thresholds (0.8) show marginal accuracy improvements but decreased Macro F1, suggesting that conservative refinement may miss opportunities for quality enhancement while maintaining correctness.

To evaluate the robustness of our training, we further analyze our RPRO training loss as shown in Fig. 4. The results demonstrate that RPRO exhibits stable and convergent training loss curves across both MedQA-USMLE and PubMedQA datasets, evidencing the robust training stability of the ranking-based preference optimization approach in medical reasoning tasks. Furthermore, the group-level ranking signals effectively mitigate the high variance issues inherent in pairwise preference methods.

Reasoning Results from PubMedQA and MedQA-USMLE Tasks. The reasoning examples shown in Table 3 in Appendix A.2 provide representative cases generated by our model using the proposed prompt templates. These examples demonstrate how our approach enables structured, interpretable, and domain-specific reasoning across both biomedical and clinical question answering tasks.

6 Conclusion and Future Work

In this paper, we present a novel approach for enhancing medical reasoning by combining ranked preference reinforcement optimization with probabilistic reasoning refinement. Our method introduces task-adaptive Chain-of-Thought generation, employs domain-specific reasoning templates, and incorporates a probabilistic quality assessment framework that enables targeted self-reflection and refinement. We evaluate our 1.1B parameter model on PubMedQA and MedQA-USMLE, demonstrating substantial improvements over much larger baseline models. These results validate that

quality-driven reasoning enhancement is more effective than simple parameter scaling. Ablation studies further highlight the critical roles of reasoning refinement and ranked preference optimization.

In future work, we plan to extend our approach to additional medical reasoning tasks, integrate external medical knowledge bases, and explore optimal trade-offs between model scale and reasoning refinement quality. Moreover, conducting human evaluation studies with domain experts and real-world clinical assessments will provide valuable insights into practical utility and safety implications, ultimately advancing the deployment of AI in healthcare settings.

References

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Shihao Cai, Chongming Gao, Yang Zhang, Wentao Shi, Jizhi Zhang, Keqin Bao, Qifan Wang, and Fuli Feng. K-order ranking preference optimization for large language models. *arXiv preprint arXiv:2506.00441*, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Mladen Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. Medalpaca: An open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Albert Jiang, Timothée Lacroix, Guillaume Lample, Arthur Mensch, Jean-Baptiste Alayrac, and other contributors. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of EMNLP-IJCNLP*, pages 2567–2577, 2019.

Takeshi Kojima, Sherry Shi Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

James-Michael Leahy, Bekzhan Kerimkulov, David Siska, and Lukasz Szpruch. Convergence of policy gradient for entropy regularized mdps with neural network approximation in the mean-field regime. In *International Conference on Machine Learning*, pages 12222–12252. PMLR, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. Retrieval-augmented generation for knowledge-intensive nlp. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. Lipo: Listwise preference optimization through learning-to-rank. In *Proceedings of NAACL-HLT 2025*, 2025.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, and Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016.

Harsha Nori, Nicholas King, Scott M McKinney, Daniel Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Fatemeh Pesaran zadeh, Yoojin Oh, and Gunhee Kim. Lpoi: Listwise preference optimization for vision language models. In *Proceedings of ACL 2025*, 2025.

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

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

Wenqi Shi, Jing Yang, et al. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of EMNLP*, 2024.

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(7972):172–180, 2023a.

Karan Singhal, Shekoofeh Azizi, Timo Tu, Sara Sepahi Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Anil Kumar Tanwani, Heather Cole-Lewis, Mayur Gambhir, et al. Towards expert-level medical question answering with large language models. *Nature Medicine*, 29(3):454–465, 2023b.

Karan Singhal, Shekoofeh Azizi, Tsung-Yen Tu, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023c.

Karan Singhal, Tsung-Yen Tu, Johannes Gottweis, et al. Large language models encode clinical knowledge. *Nature*, 2023d. Advance online publication.

Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b.

Junda Wu, Rohan Surana, Zhouhang Xie, Yiran Shen, Yu Xia, Tong Yu, Ryan A Rossi, Prithviraj Ammanabrolu, and Julian McAuley. In-context ranking preference optimization. *arXiv preprint arXiv:2504.15477*, 2025.

Hongyu Yang, Liyang He, Min Hou, Shuanghong Shen, Rui Li, Jiahui Hou, Jianhui Ma, and Junda Zhao. Aligning llms through multi-perspective user preference ranking-based feedback for programming question answering. *arXiv preprint arXiv:2406.00037*, 2024.

Cyril Zakka, Joseph Cho, Gracia Fahed, et al. Almanac copilot: Towards autonomous electronic health record navigation. *arXiv preprint arXiv:2405.07896*, 2024.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *World Wide Web*, 27(5):58, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Prompt Templates and Reasoning Outputs

A.1 Prompt Templates

To improve the interpretability and reliability of model outputs, we introduce task-adaptive prompt templates that guide the generation of structured reasoning traces. Instead of relying on a one-size-fits-all format, each task type is paired with a reasoning scaffold that reflects its unique characteristics and objectives. This design choice serves two purposes: first, it enforces a step-by-step decomposition of the reasoning process, which makes the model's decision pathway easier to trace and evaluate; second, it provides an inductive bias that steers the model towards domain-appropriate styles of justification.

The overarching principle behind these templates is to encourage multi-step reasoning that remains interpretable to human evaluators, while ensuring both factual accuracy and alignment with domain-specific norms. By providing explicit structures for reasoning, such as decomposition, background recall, logical connection, and justification, the templates reduce ambiguity in how the model organizes its answers. In turn, this structured approach facilitates both automated scoring and human assessment, making the reasoning process not only more transparent but also more consistent across different tasks.

General QA (PubMedQA-style)

This template is specifically tailored for biomedical question answering tasks such as PubMedQA, where the objective goes beyond producing a short answer. Instead, the emphasis is placed on generating an explicit reasoning trace that makes the answer both interpretable and verifiable. Biomedical questions often involve complex study designs, statistical results, or methodological considerations, and thus require reasoning that can bridge raw evidence with a clear conclusion.

By decomposing the reasoning into four stages: question decomposition, background knowledge, logical connection, and final justification, the prompt provides a structured scaffold that encourages the model to move progressively from understanding the question to constructing a justified answer. This explicit structure reduces the risk of shallow or incomplete responses and ensures that each step in the reasoning process contributes to the overall coherence of the final output.

Such a design is not only beneficial for improving factual accuracy, but also for enhancing transparency: human evaluators can more easily trace how the model arrived at its decision, identify possible reasoning gaps, and assess the quality of the justification. In this way, the prompt template serves as an inductive bias that steers the model toward producing consistent, domain-appropriate, and interpretable reasoning outputs.

Prompt Template (PubMedQA style)

```
You are a medical reasoning expert. Based on the question, context,
and final answer, produce a 4-step reasoning process:

1. Question Decomposition: Identify what the question is asking.
2. Background Knowledge: Recall relevant biomedical facts.
3. Logical Connection: Link knowledge to the question logically.
4. Final Justification: Conclude the answer with reasoning support.

Ensure factual consistency, no hallucination.
```

Clinical QA (USMLE-style, as in MedQA)

For clinical question answering tasks that involve USMLE-style medical exam questions, such as those featured in the MedQA dataset, the prompt template is explicitly designed to emulate the diagnostic reasoning process of physicians. Unlike general QA, clinical cases demand not only factual recall but also structured interpretation of patient data, prioritization of differential diagnoses, and justification of the most plausible outcome. The reasoning scaffold is therefore organized into four stages: case summarization, where essential patient demographics and presenting symptoms are distilled; evaluation of clinical significance, which highlights the findings most relevant for decision-making; differential diagnosis, where possible explanations are systematically compared; and justification of the most likely diagnosis, which consolidates evidence in support of a single conclusion. By mirroring this stepwise diagnostic workflow, the template provides the model with an inductive bias towards producing reasoning that is aligned with clinical practice. This design facilitates both transparency and evaluability, enabling human experts to trace how the model arrives at its conclusions and to assess whether the reasoning follows medically sound logic.

Prompt Template (MedQA-USMLE style)

```
You are a clinical reasoning expert. Given the patient case and
final diagnosis, produce a 4-step diagnostic reasoning process:

1. Case Summary: Summarize the key patient information.
2. Clinical Significance: Explain the important findings.
3. Differential Diagnosis: Consider possible alternatives.
4. Most Likely Diagnosis: Justify the final diagnosis.

Ensure medical accuracy, no hallucination.
```

A.2 Model-Generated Reasoning Cases

The following reasoning cases, shown in Table 3, present representative outputs generated by our model using the proposed prompt templates. These cases demonstrate how our approach enables structured, interpretable, and domain-specific reasoning across both biomedical and clinical question answering tasks.

B Scoring Examples

We illustrate how candidate CoTs are scored in our pipeline. Each candidate is systematically evaluated on three complementary criteria, using a 0–5 scale for each dimension: **Coverage**, which measures how comprehensively the reasoning process addresses the key aspects of the question; **Factual Accuracy**, which assesses whether the statements made are medically or scientifically correct; and **Redundancy**, which captures the extent of unnecessary repetition or irrelevant content that may obscure the clarity of the reasoning.

This tripartite scoring framework is designed to balance both the completeness and reliability of model-generated reasoning. By combining a positive dimension (coverage), a correctness dimension

(accuracy), and a negative dimension (redundancy), we obtain a holistic view of quality that goes beyond whether the final answer is correct. In particular, the redundancy criterion ensures that verbose or circular explanations are not mistakenly rewarded, while the factual accuracy criterion penalizes hallucinations or incorrect claims even if the answer happens to be correct.

To make the evaluation process more concrete, we provide illustrative examples from two widely used benchmark datasets, **PubMedQA** and **MedQA-USMLE**. For each dataset, we present three representative CoTs: one *high-quality*, one *medium-quality*, and one *low-quality*. These examples are synthetic and included solely to demonstrate how our scoring framework operates, rather than being actual model outputs. Through these examples, it becomes clear that evaluating reasoning requires attention not only to the final answer but also to the intermediate steps that support it.

B.1 PubMedQA

In PubMedQA, models are required to answer biomedical research questions and justify their predictions with evidence-based reasoning. To illustrate how our scoring framework operates, we present synthetic examples that highlight the distinction between comprehensive, factually grounded reasoning and superficial or vague chains of thought.

Example 1: High-quality CoT

Question: Systematic use of patient-rated depression severity monitoring: is it helpful and feasible in clinical psychiatry? **Answer:** Yes
Generated CoT:

1. Measurement-based care is key for achieving depression remission.
2. PHQ-9 was implemented in 19 psychiatric practices.
3. 93% of 6,096 patient contacts influenced by PHQ-9, 40% led to treatment changes.
4. Thus, PHQ-9 is both helpful and feasible in clinical psychiatry.

Scores: Coverage = 4.7, Factual Accuracy = 4.9, Redundancy = 0.1

Example 2: Medium-quality CoT

Question: Systematic use of patient-rated depression severity monitoring: is it helpful and feasible in clinical psychiatry? **Answer:** Yes
Generated CoT:

1. PHQ-9 is sometimes used in psychiatry.
2. It measures depression symptoms.
3. Some patients changed treatment after using it.
4. Therefore, it is helpful.

Scores: Coverage = 3.2, Factual Accuracy = 3.5, Redundancy = 0.6

Example 3: Low-quality CoT

Question: Systematic use of patient-rated depression severity monitoring: is it helpful and feasible in clinical psychiatry? **Answer:** Yes
Generated CoT:

1. Depression is bad.
2. Patients answer questions.
3. Doctors see results.
4. So it helps.

Scores: Coverage = 1.2, Factual Accuracy = 1.5, Redundancy = 0.1

B.2 MedQA-USMLE

For MedQA-USMLE, the task involves answering clinical questions in the USMLE style with precise diagnostic reasoning. To illustrate our scoring framework, we present examples showing that the same final answer can be supported by reasoning of very different quality. Our evaluation criteria capture whether the reasoning correctly contextualizes the case, enumerates differential diagnoses, and avoids vague or trivial statements.

Example 4: High-quality CoT

Question: A 48-year-old man with chronic alcoholism presents with seizure, confusion, and severe hyponatremia. What is the most likely complication of urgent treatment? **Answer:** Osmotic myelinolysis
Generated CoT:

1. Case Summary: 48M, seizure, confusion, history of alcoholism, hyponatremia.
2. Clinical Significance: Hyponatremia with seizure suggests electrolyte disturbance.
3. Differential Diagnosis: Possible complications include cerebral edema, arrhythmia, Wernicke encephalopathy.
4. Most Likely Diagnosis: Rapid correction of sodium can cause osmotic myelinolysis.

Scores: Coverage = 4.8, Factual Accuracy = 5.0, Redundancy = 0.1

Example 5: Medium-quality CoT

Question: A 48-year-old man with chronic alcoholism presents with seizure, confusion, and severe hyponatremia. What is the most likely complication of urgent treatment? **Answer:** Osmotic myelinolysis
Generated CoT:

1. Patient had seizure and alcoholism.
2. Hyponatremia is dangerous.
3. Could be related to brain swelling or nerve problems.
4. So the answer is osmotic myelinolysis.

Scores: Coverage = 3.0, Factual Accuracy = 3.2, Redundancy = 0.5

Example 6: Low-quality CoT

Question: A 48-year-old man with chronic alcoholism presents with seizure, confusion, and severe hyponatremia. What is the most likely complication of urgent treatment? **Answer:** Osmotic myelinolysis
Generated CoT:

1. Patient sick.
2. Sodium problem.
3. Bad outcome.
4. Answer: osmotic myelinolysis.

Scores: Coverage = 1.0, Factual Accuracy = 1.5, Redundancy = 0.1

C Dataset Format (JSONL Example)

We adopt a JSONL format for our training data in the RPRO pipeline. Each entry corresponds to a single PubMedQA-style example, and contains the following fields:

- `id`: unique identifier of the example.

- `question`: the biomedical question to be answered.
- `context`: supporting abstract or passage (from PubMedQA).
- `answer`: the gold label.
- `ranked`: a list of candidate reasoning chains ranked from better to worse.
- `scores`: numerical quality scores corresponding to each reasoning chain.
- `breakdowns`: fine-grained evaluation dimensions (coverage, factual accuracy, redundancy).

Below is an illustrative JSONL example (truncated for brevity):

```
{
  "id": "PMQA_001",
  "question": "Can patient coaching reduce disparities?",
  "context": "This study evaluated whether structured patient coaching
      improves adherence and reduces disparities ...",

  "answer": "Yes",

  "ranked": [
    "1. Key clinical features suggest...\n4. Disparities were
       eliminated.",
    "1. Coaching is sometimes used...\n4. May not fully resolve
       disparities.",
    ...
  ],

  "scores": [
    0.90,
    0.65,
    ...
  ],

  "breakdowns": [
    {"coverage": 4.5, "factual_accuracy": 4.7, "redundancy": 0.2},
    {"coverage": 3.5, "factual_accuracy": 3.2, "redundancy": 0.7},
    ...
  ]
}
```

Table 3: Reasoning results for PubMedQA and MedQA-USMLE tasks.

Question	Reasoning & Prediction
PubMedQA	
Does body mass index (BMI) influence morbidity and long-term survival in gastric cancer patients after gastrectomy?	**Reasoning CoT:** 1. Question Decomposition: Determine if BMI is a predictor of morbidity and long-term survival in gastric cancer patients. 2. Background Knowledge: BMI is a predictor of morbidity and long-term survival in gastric cancer patients. 3. Logical Connection: If BMI is significantly associated with these outcomes, it can serve as a predictive factor. 4. Final Justification: Based on the results of the study, BMI is a predictor of morbidity and long-term survival in gastric cancer patients. **Answer: yes** **Final Prediction: yes**
MedQA-USMLE	
A 35-year-old man with itchy, watery eyes, recurrent in spring; conjunctival injection and watery discharge. Most appropriate treatment? **Options:** A. Erythromycin ointment B. Ketotifen eye drops C. Warm compresses D. Fluorometholone eye drops E. Latanoprost eye drops	**Reasoning CoT:** 1. Case Summary: 35M with itchy, watery eyes and sneezing, recurrent in spring, exam shows conjunctival injection with watery discharge. 2. Clinical Significance: These findings suggest allergic conjunctivitis rather than infection. 3. Differential Diagnosis: Bacterial conjunctivitis usually has purulent discharge, viral can be acute but less itchy, seasonal pattern fits allergy. 4. Most Likely Diagnosis: Allergic conjunctivitis, treatment is with antihistamine/mast cell stabilizer eye drops such as ketotifen. **Answer: Ketotifen eye drops (B)** **Final Prediction = Ground Truth: B**