MBR-PRO: Consensus-Aware Reward Modulation for Medical Reasoning

Anonymous ACL submission

Abstract

While outcome-only RLVR excels in medical question answering, it assigns identical learning signals to responses with the same final correctness, which can arise from very different reasoning processes. Consequently, this coarse feedback overlooks important information about reasoning stability and reliability. We propose MBR-PRO, a consensus-aware RLVR framework that leverages semantic agreement among multiple sampled reasoning trajectories as an auxiliary stability signal. MBR-PRO uses process-level semantic consensus to modulate outcome-based rewards, strengthening reinforcement for correct answers supported by stable reasoning patterns while updating low-consensus reasoning more conservatively. Experiments on five medical QA benchmarks show consistent accuracy improvements over corresponding backbone models, with gains of up to 5.7 points, and outperform outcome-only RLVR baselines on average. Moreover, Qwen3-8B trained with MBR-PRO can match or surpass substantially larger medical models.

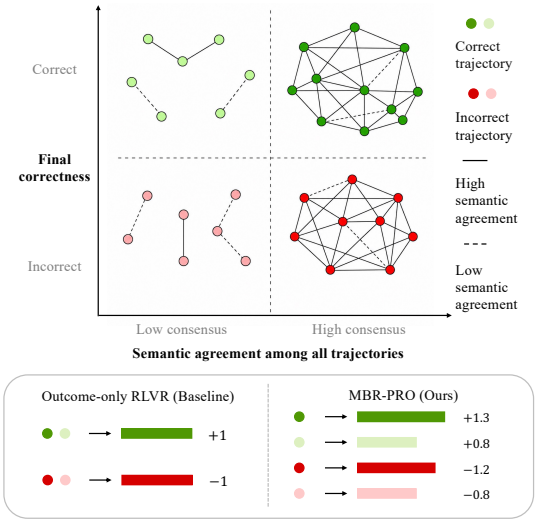

Figure 1: Multiple sampled reasoning trajectories exhibit diverse semantic agreement patterns beyond final-answer correctness. Unlike outcome-only RLVR, MBR-PRO uses process-level semantic agreement to adaptively reweight verifier rewards, amplifying updates for high-consensus trajectories while attenuating low-consensus ones. For clarity, only representative local semantic agreement connections are shown.

1 Introduction

Large language models have recently achieved strong performance on medical multiple-choice question answering (QA) benchmarks (Singhal et al., 2025; Chen et al., 2025). The presence of verifiable final answers makes these tasks a natural testbed for reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024), where model outputs can be automatically checked against ground-truth answers and used as training signals. Common outcome-only RLVR relies on final-answer correctness as the reward and have been shown to substantially improve performance on medical QA benchmarks (Chen et al., 2025).

Despite its empirical success, outcome-only RLVR suffers from an important limitation: it treats all responses with the same final correctness as equally informative. In medical reasoning tasks, however, final correctness alone is an incomplete proxy for reasoning quality. Empirically, when sampling multiple responses to the same question, we observe substantial variation in the underlying reasoning processes. Correct answers may arise from either stable, systematic reasoning patterns or fragmented cue matching that happens to reach the right option, while incorrect answers may reflect either random exploration failures or stable but systematically flawed reasoning schemas. By collapsing these qualitatively different behaviors into identical binary rewards, outcome-only RLVR fails to distinguish reliable reasoning from brittle or misleading patterns, resulting in coarse and potentially noisy training signals. This limitation suggests additional information beyond final cor-

rectness is needed to more effectively shape the model's reasoning behavior during training.

One promising insight is to harness the semantic agreement among multiple sampled reasoning trajectories as a measure of reasoning stability. High agreement among reasoning paths generally indicates a more systematic, stable approach, even if such consensus does not guarantee the answer is correct, whereas low agreement suggests more fragmented, ad-hoc reasoning. This perspective aligns with the principle of Minimum Bayes Risk (MBR) (Kumar and Byrne, 2004; Freitag et al., 2022), which prefers outputs that exhibit mutual agreement under a utility function. Recent studies have extended MBR-style consensus beyond decoding to consensus-aware training (Prasad et al., 2024; Li et al., 2024; Wu et al., 2024). However, these approaches primarily use consensus as a surrogate supervision signal for sample selection and preference construction during self-improvement, without leveraging explicit verifier supervision.

To address this gap, we propose **MBR-PRO**, a consensus-aware RLVR framework that combines process-level semantic consensus among trajectories with verifiable rewards during reinforcement learning. MBR-PRO computes a semantic consensus score over multiple sampled reasoning trajectories for the same question, measuring how consistently a trajectory aligns with other plausible reasoning paths at the process level. The resulting consensus signal is used to modulate outcome-based verifiable rewards during training. As a result, trajectories that are both correct and semantically stable receive stronger reinforcement, consistently flawed reasoning patterns are penalized more heavily, and low-consensus trajectories, whether correct or incorrect, are updated more conservatively. Importantly, MBR-PRO does not impose a single canonical reasoning path or require step-level correctness annotation. Unlike existing process-aware reinforcement learning approaches that rely on external medical knowledge bases or dense intermediate supervision, our method leverages agreement structure among the model's own sampled trajectories, preserving the direction of verifiable supervision while reshaping the optimization landscape to favor reliable and systematic reasoning.

We conduct extensive experiments on five medical QA benchmarks, using MedQA as the training-aligned benchmark and evaluating generalization on PubMedQA, MedMCQA, GPQA-M, and MMLU-Pro-M. Experiments across five representative 7B–8B backbones, including both general-purpose and medical-specialized models, show that MBR-PRO consistently improves average accuracy over corresponding backbone models, with gains of up to $+\mathbf{5.7}$ points. Beyond accuracy improvements, our analyses further suggest that semantic consensus captures complementary information about reasoning organization beyond final-answer correctness, helping distinguish stable reasoning patterns from fragmented or systematically biased trajectories. Furthermore, on representative backbones including Qwen2.5-7B and Llama-3.1-8B, MBR-PRO consistently outperforms outcome-only RLVR baselines on average, highlighting the effectiveness of consensus-aware reward modulation. Notably, Qwen3-8B trained with MBR-PRO reaches $\mathbf{74.6}$ average accuracy, matching or slightly exceeding a substantially larger 32B baseline (74.4), highlighting the compute-efficiency of process-aware training signals relative to parameter scaling.

In summary, our contributions are threefold:

- We study the role of process-level semantic consensus beyond final-answer correctness, exploring its use as an auxiliary signal combined with verifiable rewards in RLVR.
- We propose MBR-PRO, a lightweight, consensus-aware RLVR framework that improves reward separability within identical binary outcomes without requiring dense process annotations, external knowledge bases, or explicit process reward models.
- We validate MBR-PRO across 7B–8B backbones on multiple medical QA benchmarks, showing consistent improvements over corresponding backbone models and outcome-only RLVR baselines, enabling them to match or surpass medical models up to four times larger.

2 Related Work

2.1 Medical Reasoning Models

Enhancing medical reasoning in large language models is critical for complex clinical decision-making. Early approaches rely on continual pre-training or prompting (Luo et al., 2022; Singhal et al., 2025; Nori et al., 2023; Labrak et al., 2024), without explicit reasoning-oriented training objectives. Recent work introduces explicit

supervision over reasoning. One line distills structured reasoning traces from teachers or knowledge sources (Chen et al., 2025; Huang et al., 2026; Wu et al., 2025), or enhances reasoning through iterative refinement and self-correction (Yang et al., 2025b), improving controllability but requiring costly annotation, refinement pipelines, or external supervision. Another line shows that answer-level supervision can elicit strong reasoning via outcome-only RLVR (Shao et al., 2024), with medical adaptations ranging from pure RLVR methods such as AlphaMed (Liu et al., 2025) to hybrid SFT-RL approaches such as BioMed-R1 (Thapa et al., 2026).

Moreover, recent process-aware RL approaches explicitly evaluate intermediate reasoning steps, either via search-based credit assignment (Jiang et al., 2025) or external knowledge-guided scoring (Yun et al., 2025). While more fine-grained, such approaches introduce additional models or supervision pipelines. In contrast, we operate within the RLVR paradigm without explicit process supervision. Our key idea is to use similarity-based consensus as a training signal to modulate rewards, implicitly shaping reasoning trajectories without requiring step-level labels or auxiliary reward models.

2.2 From MBR Decoding to Consensus-Aware Training

While Minimum Bayes Risk (MBR) has traditionally been used as a test-time decoding strategy for selecting candidates with the highest expected utility (Kumar and Byrne, 2004; Freitag et al., 2022; Jinnai et al., 2025), recent work has further connected MBR-style consensus signals to uncertainty estimation (Vashurin et al., 2025), as well as self-improvement and training. Self-consistency (Wang et al., 2022), a simplified answer-level MBR variant, has been used to construct preference pairs for verifiable reasoning tasks via ScPO (Prasad et al., 2024). To capture signals beyond answer-level agreement, Wu et al. (2024) applies MBR reranking to construct training targets for SFT and DPO-based self-distillation in instruction-following tasks, while SEALONG (Li et al., 2024) further applies MBR-style scoring over reasoning trajectories for long-context self-improvement. However, existing approaches primarily use consensus as a surrogate quality signal for reranking, sample selection, or preference construction. In contrast, we integrate process-level semantic consensus directly into RLVR as a dynamic reward modulation signal coupled with verifiable rewards, enabling finer-grained optimization beyond final-answer correctness alone.

3 Method

3.1 Problem Setup

We consider reinforcement learning with verifiable rewards for medical reasoning, where supervision is obtained by checking the correctness of the final answer. Given a medical query q, the policy π_θ generates a group of reasoning trajectories $\mathcal{O} = \{o_1, \ldots, o_N\}$. A rule-based verifier assigns a binary outcome reward

$$R_{\text{ver},i} = \begin{cases} 1, & \text{if Extract}(o_i) = y \\ -1, & \text{otherwise} \end{cases} \tag{1}$$

where y is the ground-truth option and $\text{Extract}(\cdot)$ extracts the final answer option from the response. While reliable, outcome-only supervision assigns identical rewards to trajectories with the same final-answer correctness even when their reasoning processes differ substantially, leading to coarse-grained credit assignment during policy optimization. This motivates using the consensus structure among sampled trajectories to refine the relative strength of verifier-guided updates.

3.2 Semantic Consensus Estimation

Following the MBR intuition that samples with higher agreement under the model distribution are more reliable, we estimate a trajectory-level *consensus score* for each sampled response. As shown in Figure 2, each trajectory o_i is first decomposed into a sequence of reasoning segments $P_i = \{p_{i,1}, \ldots, p_{i,K_i}\}$ using paragraph boundaries. Subsequently, each segment is encoded with a text encoder Ψ to obtain a sequence of semantic embeddings $T_i = [\Psi(p_{i,1}), \ldots, \Psi(p_{i,K_i})]$.

To compare trajectories with different lengths, we compute pairwise semantic similarity using Dynamic Time Warping (DTW) (Mul, 2007) with cosine distance as the local matching cost:

$$S_{i,j} = \exp\left(-\frac{D(T_i, T_j)}{\sigma}\right), \tag{2}$$

where $D(T_i, T_j)$ denotes the length-normalized DTW distance between trajectories T_i and T_j. We set σ as a fixed temperature in all experiments.

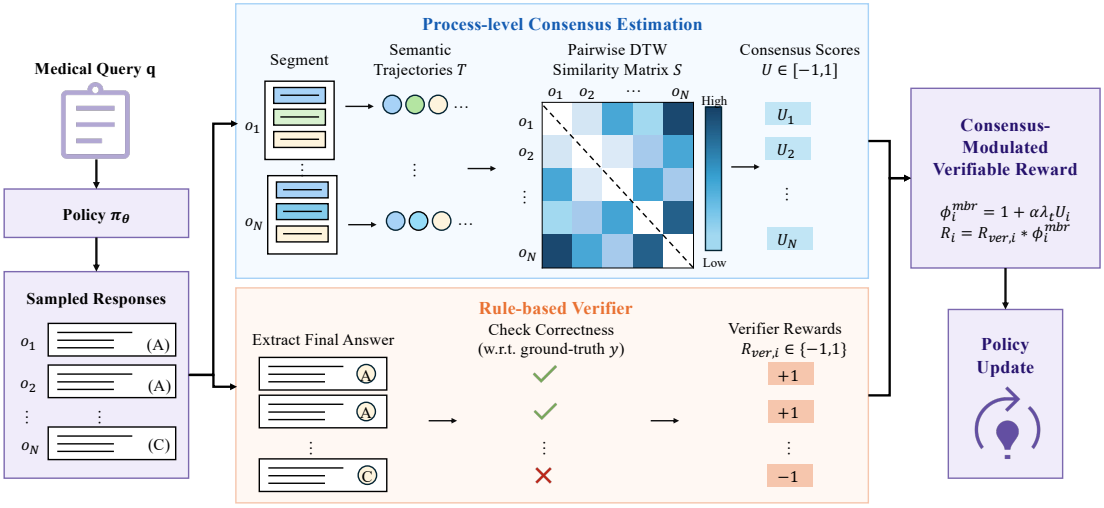

Figure 2: Overview of MBR-PRO. The framework estimates process-level semantic consensus among multiple sampled reasoning trajectories and uses it to modulate verifier rewards during RL training.

The raw consensus score $\bar{S}_i$ for trajectory o_i is defined as its average semantic similarity to the remaining samples within the group:

$$\bar{S}_i = \frac{1}{N-1} \sum_{j \neq i} S_{i,j}. \tag{3}$$

To eliminate scale variance across different queries and construct a zero-centered relative signal, we apply intra-group min–max normalization to project the score into a $[-1, 1]$ interval, yielding the final consensus utility score U_i:

$$U_i = \frac{2(\bar{S}_i - \bar{S}_{\min})}{\bar{S}_{\max} - \bar{S}_{\min} + \epsilon} - 1, \tag{4}$$

where

$$\bar{S}_{\min} = \min_k \bar{S}_k, \qquad \bar{S}_{\max} = \max_k \bar{S}_k, \tag{5}$$

and $\epsilon > 0$ is a small constant introduced for numerical stability. Intuitively, U_i captures how consistently a reasoning trajectory aligns with other sampled trajectories for the same question under stochastic generation.

3.3 Consensus-Modulated Verifiable Reward

To better distinguish trajectories with identical outcome rewards, we use the consensus score as a trajectory-level stability signal to adaptively modulate the verifier reward.

Given the intra-group consensus utility score U_i, we construct a modulation factor ϕ_i^{mbr}:

$$\phi_i^{\text{mbr}} = 1 + \alpha \cdot \lambda_t \cdot U_i, \tag{6}$$

where the warm-up coefficient λ_t gradually increases during training: $\lambda_t = \min\left(\frac{t}{t_{\text{warm_up}}}, 1\right)$. The hyperparameter α controls the strength of consensus-based modulation. By constraining $\alpha \in [0, 1]$, the modulation factor is within a bounded range: $\phi_i^{\text{mbr}} \in [1-\alpha, 1+\alpha] \subseteq [0, 2]$. The final trajectory reward is computed as:

$$R_i = R_{\text{ver},i} \cdot \phi_i^{\text{mbr}}. \tag{7}$$

This formulation preserves the verifier reward as the primary optimization objective while improving reward separability under binary supervision according to process-level semantic consensus. As a result, high-consensus correct trajectories receive stronger positive updates, high-consensus incorrect trajectories receive stronger penalties, and low-consensus trajectories are updated more conservatively regardless of their final correctness. The modulated reward R_i is then directly utilized by the DAPO algorithm (Yu et al., 2025) to optimize the policy, replacing the baseline verifier reward without altering the underlying training objective.

4 Experiments

4.1 Experimental Setup

Baselines. We compare our models with two types of LLMs: 1) General base LLMs: Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen-2.5-7B-Instruct (Qwen et al., 2025), Qwen3-8B (Yang et al., 2025a); and 2) Medical-specific LLMs: HuatuoGPT-o1 (Chen et al., 2025), m1 (Huang et al., 2026), AlphaMed (Liu et al., 2025), MedREFL-8B (Yang et al., 2025b), BioMed-R1 (Thapa

Model	In-Domain	Out-of-Domain				AVG
	MedQA	PUBMEDQA	MedMCQA	GPQA	MMLU-Pro	MedBench
General 7B-8B models						
Qwen2.5-7B-Instruct	62.1	75.4	55.5	48.5	63.4	61.0
Llama-3.1-8B-Instruct	68.4	77.4	59.3	44.5	61.2	62.2
Qwen3-8B	80.1	75.5	65.8	57.3	76.4	71.0
Medical Specific 7B-8B models						
AlphaMed-7B‡§	71.6	73.3	61.5	53.9	66.8	65.4
m1-7B†§	75.8	62.5	**75.8**	53.1	65.9	66.6
HuatuoGPT-o1-8B†§	72.0	78.8	62.5	52.6	63.1	65.8
Med-REFL-8B	73.7	80.0	64.7	56.8	65.0	68.0
AlphaMed-8B‡§	76.2	**80.4**	64.5	58.4	66.7	69.2
BioMed-R1-8B†‡§	81.7	78.4	67.9	53.3	71.3	70.5
Medical Specific Larger models (>30B)						
m1-32B†§	83.5	77.6	67.3	66.7	76.9	74.4
AlphaMed-70B‡§	87.5	80.9	75.1	77.5	79.6	80.1
MBR-PRO on Different Backbones (7-8B)						
Qwen2.5-7B + MBR-PRO	72.0	74.3	61.2	53.3	68.5	65.9(+4.9)
AlphaMed-7B + MBR-PRO	73.8	76.8	61.8	52.8	69.7	67.0(+1.6)
Llama3.1-8B + MBR-PRO	75.6	78.3	64.4	55.4	65.6	67.9(+5.7)
AlphaMed-8B + MBR-PRO	77.5	79.9	67.0	60.3	68.5	71.0(+1.8)
Qwen3-8B + MBR-PRO	**84.8**	76.4	68.6	**63.8**	**79.5**	**74.6**(+3.6)

Table 1: Comparison on five medical QA benchmarks. "In-Domain" and "Out-of-Domain" are defined relative to the post-training data. The best results among 7B-8B models are highlighted in **bold**, and the second best are underlined. Distilled-CoT supervision from stronger models(†), reinforcement learning with verifier rewards (‡), trained with external datasets beyond MedQA (§).

et al., 2026). Comparison results are obtained either from published papers or from evaluating publicly available checkpoints.

Benchmarks. Following HuatuoGPT-o1 (Chen et al., 2024), we evaluate our MBR-PRO on standard medical benchmarks. The benchmark consists of five multi-choice medical question answer datasets, including MedQA (USMLE test set) (Jin et al., 2021), MedMCQA (validation set) (Pal et al., 2022) and PubMedQA (test set) (Jin et al., 2019), GPQA (medical subset) (Rein et al., 2024) and MMLU-Pro (medical subset) (Wang et al., 2024). We use "MedBench" to denote the average answer accuracy across the five medical QA datasets.

Implementation Details. We conduct training on three general-purpose LLMs as well as two AlphaMed variants initialized from prior medical RLVR training, to evaluate the generality of our method across both generic and medically adapted backbones. Only MedQA (USMLE, four-option multiple-choice, 12,723 samples) is used for post-training, while all remaining benchmarks are reserved for evaluating generalization. Qwen3-8B-Embedding model serves as text encoder Ψ for semantic embedding. We adopted the Verl framework [1] and trained the model for 200 steps with a learning rate of $1e^{-6}$. The warm-up step number $t_{\text{warm_up}}$ is empirically set to half of the total training steps in all experiments. We sample 16 answers for each question and a batch size of 64 questions, mini batch size of 16 for policy update. The max response length was empirically set to 2048 for Qwen2.5-7B, Llama3.1-8B and AlphaMed variants, and 6144 for Qwen3-8B. The parameters clip_ratio_low and clip_ratio_high were set to 0.2 and 0.28 respectively. All experiments are conducted on a cluster of 4 NVIDIA A100 80G GPUs. For inference, we follow the default sampling configuration of each backbone model and report the average accuracy over three random seeds.

Metrics. Beyond answer accuracy, we analyze reasoning **process quality** using an LLM-based judge (GPT-4o) across four dimensions: evidence grounding, medical soundness, logical coherence, and answer alignment (definitions and prompts are provided in Appendix C). The mean score across these dimensions is reported as the process quality

[1] https://github.com/volcengine/verl

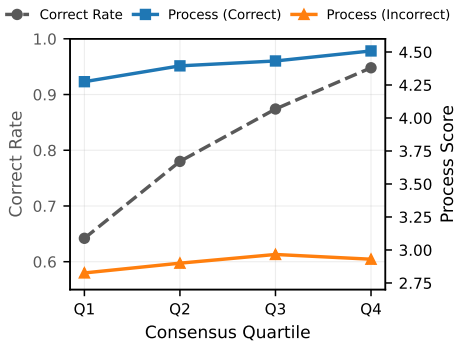

Figure 3: Process quality and correctness across consensus quartiles. Trajectories are partitioned into four consensus quartiles, from lowest (Q1) to highest (Q4). Higher-consensus bins consistently display both a higher proportion of correct answers and elevated process quality scores.

score, ranging from 1 to 5.

4.2 Main Results

Table 1 summarizes the main results on five medical QA datasets across model scales. Across all 7B–8B models, incorporating our MBR-PRO framework consistently improves performance, yielding gains of up to +5.7 average accuracy points over the corresponding backbones. Notably, these improvements are observed not only on the training-aligned MedQA dataset but also transfer reliably to diverse out-of-domain datasets, indicating enhanced generalization. On the strongest 8B backbone, Qwen3-8B + MBR-PRO achieves an average accuracy of 74.6, matching or exceeding substantially larger 32B medical models. This indicates that explicitly leveraging process-level consensus information during training can serve as a compute-efficient alternative to scaling model size for improving medical reasoning performance. Overall, the results demonstrate three consistent properties of MBR-PRO: (1) effectiveness across diverse medical QA datasets, (2) robustness under out-of-domain evaluation settings, and (3) scalability across both general-purpose and medical-specific backbones.

4.3 What Does Consensus Capture Beyond Correctness?

We next analyze what process-level semantic consensus captures beyond final-answer correctness.

Quantitatively, as illustrated in Figure 3, higher semantic consensus is generally associated with improved final-answer correctness and better process quality scores when conditioning on correctness outcomes. Within the correct subset, trajectories with higher consensus achieve higher process quality scores, indicating that semantic agreement captures fine-grained nuances masked by binary outcome rewards. Conversely, for incorrect trajectories, process quality initially improves with consensus before hitting a plateau, implying that high-consensus failures may stem from coherent, structured misconceptions rather than stochastic errors.

Qualitatively, we further present a representative medical question to better illustrate the reasoning patterns associated with different consensus levels. Our analysis reveals distinct reasoning modes across the consensus–correctness space:

- **Canonical vs. Fragmented Reasoning (Correct):** High-consensus correct trajectories tend to follow a more stable and canonical reasoning pattern, systematically integrating diverse clinical evidence (e.g., marfanoid features, ectopia lentis, metabolic abnormalities, and neurological findings) to support a diagnosis of homocystinuria. In contrast, low-consensus correct trajectories often rely on more localized or partially developed evidence matching (e.g., focusing primarily on lens dislocation), resulting in less organized reasoning despite reaching the same final answer.

- **Systematic but Incorrect Reasoning (Incorrect):** Within the incorrect subsets, high-consensus trajectories consistently associate lens dislocation and metabolic abnormalities with alkaptonuria, forming a coherent but systematically biased reasoning pattern shared across multiple samples.

Detailed representative examples of four reasoning modes are in Appendix B.

Taken together, these observations suggest that semantic consensus captures complementary information about reasoning organization beyond final-answer correctness. In particular, consensus may help distinguish stable and concentrated reasoning patterns from fragmented or systematically biased trajectories with the same final answers.

4.4 Ablation Study

Different Consensus Estimators. We compare different methods for estimating process-level con-

Method	MedBench	Train
Avg@16 (Random Sampling)	61.0	60.7
Majority Voting	63.8	64.3
Pass@16 (Oracle Upper Bound)	86.0	87.7
BLEU Lexical Consensus (BoN)	63.4	62.1
DTW Semantic Consensus (BoN)	63.3	64.0
LLM-Judge Consensus (BoN)	64.2	64.4
DTW Semantic Consensus (W-MV)	64.0	64.5

Table 2: Test-time probing of different consensus estimation strategies on Qwen2.5-7B. DTW-based semantic consensus outperforms lexical overlap metrics and achieves performance comparable to substantially more expensive LLM-based estimators on average, while remaining lightweight and stable.

Backbone	Dataset	R_{ver}	$R_{ver} + U$	$R_{ver} \cdot \phi_{mbr}$
Qwen2.5-7B	MedQA	69.6	70.1	**72.0**
	PubMedQA	**76.9**	75.2	74.3
	MedMCQA	60.5	60.5	**61.2**
	GPQA	46.9	**56.2**	53.3
	MMLU-Pro	67.9	**68.5**	**68.5**
	Average	64.4	**66.1**	65.9
Llama3.1-8B	MedQA	74.8	74.9	**75.6**
	PubMedQA	78.8	**79.1**	78.3
	MedMCQA	63.2	63.0	**64.4**
	GPQA	54.2	52.3	**55.4**
	MMLU-Pro	65.1	**66.4**	65.6
	Average	67.2	67.1	**67.9**

Table 3: Ablation study of different strategies for incorporating process-level consensus into RLVR training across two backbone models.

Dataset	0.25	0.5	0.75	1.0
MedQA	70.0	72.0	71.1	**73.0**
PubMedQA	75.2	74.3	**76.9**	76.0
MedMCQA	60.6	**61.2**	60.2	60.6
GPQA	49.2	**53.3**	51.7	50.9
MMLU-Pro	68.1	**68.5**	68.0	68.1
Average	64.6	**65.9**	65.6	65.7

Table 4: Effect of the scaling coefficient α. Note that $\alpha = 0$ reduces to using R_{ver} alone and is therefore omitted here. Results are reported on Qwen2.5-7B.

sensus, including lexical overlap, semantic trajectory matching, and LLM-Judge-based agreement estimation. We further evaluate two usage strategies: **BoN (Best-of-N)** selection, where the response with the highest consensus score is selected, and **W-MV(Weighted Majority Voting)**, which uses consensus scores to reweight answer-level votes.

As shown in Table 2, DTW-based semantic consensus achieves stronger average performance across MedBench and MedQA-Train than lexical overlap metrics, while remaining comparable to LLM-Judge-based estimators. Although LLM-Judge-based consensus estimation achieves the strongest standalone BoN performance, we found it substantially more expensive and less stable when incorporated into RL training, without improving final training performance. Full results featuring additional variants and analysis are provided in Table 5 in Appendix A.

Consequently, we adopt DTW-based semantic consensus as a lightweight and stable approximation of reasoning similarity that achieves a favorable balance between effectiveness and computational efficiency. Furthermore, combining consensus with answer-level majority voting through weighted aggregation further improves over standard majority voting, suggesting that consensus is most potent when used together with outcome-level correctness signals.

Reward Modulation Strategies. We next investigate how process-level consensus should be incorporated into RLVR by comparing three reward designs: **verifier-only reward** R_{ver} as baseline, **additive integration** $R_{ver} + U$ using the consensus utility score, and **multiplicative modulation** $R_{ver} \cdot \phi_{mbr}$ using the modulation factor.

As shown in Table 3, incorporating process-level consensus generally improves performance over using the verifiable reward alone, confirming that semantic agreement provides a useful auxiliary signal for RLVR optimization.

The two integration strategies exhibit different behaviors across datasets and backbones. The additive formulation can produce large gains on specific datasets, such as GPQA for Qwen2.5-7B, but its improvements are less consistent across model families. In contrast, the multiplicative formulation achieves more stable average performance across both backbones and consistently obtains the best results on the in-domain MedQA dataset. This suggests that modulating verifier rewards with process-level stability may be particularly beneficial for domain-aligned medical reasoning tasks.

Overall, we adopt the multiplicative formulation as the default design. This design preserves the direction of verifier-based supervision while adaptively scaling its strength according to trajectory-level stability.

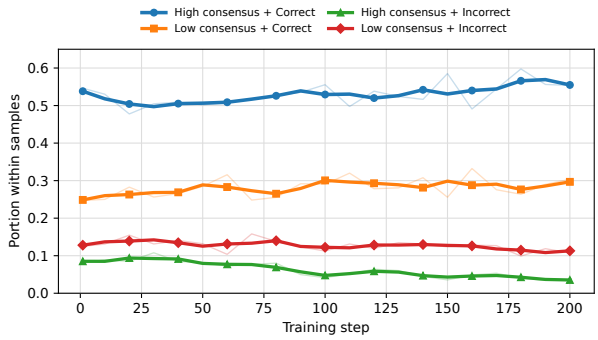

Figure 4: Training dynamics of the correctness–consensus topology under MBR-PRO.

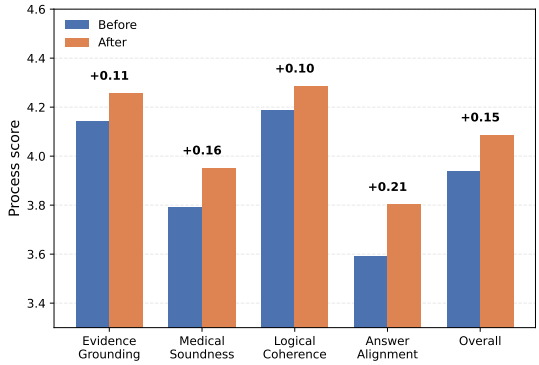

Figure 5: Process-level reasoning quality before and after MBR-PRO training on Qwen2.5-7B. Results are averaged over MedBench using three sampled responses per question across four evaluation dimensions.

Effect of the scaling coefficient α. We further study the scaling coefficient α, which controls the strength of consensus modulation, by varying $\alpha \in \{0.25, 0.5, 0.75, 1.0\}$ while keeping other hyperparameters fixed. As shown in Table 4, $\alpha = 0.5$ achieves the best average performance. Smaller values provide weaker and insufficient modulation, while larger values do not yield further gains and slightly hurt challenging datasets such as GPQA. We therefore use $\alpha = 0.5$ as the default setting for the balance between leveraging consensus information and preserving stable RL optimization.

4.5 Distributional Effects of Consensus Modulation

To understand how consensus-aware reward modulation influences reasoning behavior, we analyze the correctness–consensus topology during training of Qwen2.5-7B, where trajectories are grouped by final-answer correctness and relative process-level consensus within each query. Figure 4 shows the proportion of trajectories under different correctness–consensus categories over training. High-consensus correct trajectories remain the dominant mode and exhibit a mild upward trend, suggesting that consensus-aware reward modulation progressively reallocates probability mass toward stable successful reasoning patterns. In contrast, the proportion of high-consensus incorrect trajectories consistently decreases over time, indicating that the policy gradually suppresses stable but systematically flawed reasoning modes. Low-consensus incorrect trajectories also show a smaller decline than high-consensus incorrect ones, suggesting weaker optimization pressure on fragmented erroneous reasoning patterns. Importantly, low-consensus correct trajectories are not entirely eliminated throughout training, indicating that the method does not simply collapse reasoning diversity into a single dominant template.

4.6 Process-level Effects of Consensus Modulation

As shown in Figure 5, reasoning quality improves after training across all dimensions, with the largest gains in answer alignment and medical soundness. These moderate but consistent improvements suggest that MBR-PRO not only improves final-answer accuracy, but also encourages reasoning processes that are better grounded in the question and more aligned with the predicted answer.

5 Conclusion

In this work, we extend the consensus principle underlying Minimum Bayes Risk beyond test-time reranking and preference construction to the reinforcement learning stage of reasoning models. By interpreting process-level semantic agreement among sampled trajectories as an auxiliary stability signal, MBR-PRO adaptively modulates verifiable rewards during RLVR training. This transforms semantic consensus from a post-hoc selection criterion into a trajectory-aware training signal for shaping the model's reasoning distribution. Empirical results on medical reasoning benchmarks demonstrate that consensus-aware reward modulation consistently improves performance over outcome-only RLVR baselines and leads to more stable reasoning behavior across sampled trajectories. We hope this work encourages further exploration of how multi-sample reasoning structure and agreement signals can be incorporated into the training of reasoning-centric language models.

6 Limitations

First, MBR-PRO introduces computational overhead from multi-sample trajectory generation and similarity estimation. Scaling to larger sample budgets or models remains expensive despite efficient embedding-based metrics. Second, our evaluation is primarily constrained to medical multiple-choice QA. Validating consensus-aware RL in open-ended reasoning and broader domains is a key future direction. Finally, our embedding-based semantic alignment is a practical approximation. Incorporating more structured process metrics (e.g., causal or verifier-aware signals) could further enhance consensus estimation.

References

2007. *Dynamic Time Warping*, pages 69–84. Springer Berlin Heidelberg, Berlin, Heidelberg.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. *Preprint*, arXiv:2412.18925.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards medical complex reasoning with llms through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573.

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

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2026. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *Preprint*, arXiv:2504.00869.

Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. Meds^3: Towards medical small language models with self-evolved slow thinking. *arXiv e-prints*, pages arXiv–2501.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. 2025. Regularized best-of-n sampling with minimum bayes risk objective for language model alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9321–9347.

Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2024. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2049–2066.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. 2024. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*.

Che Liu, Haozhe Wang, Jiazhen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel Rueckert, and Rossella Arcucci. 2025. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. *arXiv preprint arXiv:2505.17952*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason Weston, and Jane Yu. 2024. Self-consistency preference optimization. *arXiv preprint arXiv:2411.04109*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

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

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

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison G Zhang, Angela Zhang, Eric Wu, Haotian Ye, and James Zou. 2026. Reasoning or knowledge: Stratified evaluation of biomedical llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2483.

Roman Vashurin, Maiya Goloburda, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. Cocoa: A generalized approach to uncertainty quantification by integrating confidence and consistency of llm outputs. *arXiv e-prints*, pages arXiv–2502.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Ian Wu, Patrick Fernandes, Amanda Bertsch, Seungone Kim, Sina Pakazad, and Graham Neubig. 2024. Better instruction-following through minimum bayes risk. *arXiv preprint arXiv:2410.02902*.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. MedReason: Eliciting Factual Medical Reasoning Steps in LLMs via Knowledge Graphs. *Preprint*, arXiv:2504.00993.

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

Zongxian Yang, Jiayu Qian, Zegao Peng, Haoyu Zhang, and Zhi-An Huang. 2025b. Med-refl: Medical reasoning enhancement via self-corrected fine-grained reflection. *arXiv preprint arXiv:2506.13793*.

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

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, and 1 others. 2025. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16565–16582.

A Extended Analysis of Consensus Estimators

To better understand what information is captured by process-level consensus, we provide extended comparisons of different consensus estimators and matching strategies under test-time selection.

We explore three categories of similarity functions, as shown in Table 5. (1) **Lexical Consensus Estimators**, including n-gram Jaccard, BLEU, and ROUGE, which capture surface-level lexical overlap between reasoning traces. (2) **Semantic Consensus Estimators**, where we first segment long chains of thought into multiple paragraphs, encode each paragraph with a semantic embedding model, and then compute trajectory-level similarity using different matching strategies: *average matching*, which averages all paragraph embeddings into a single trajectory representation; *greedy matching*, which greedily aligns paragraph embeddings with the most similar paragraph of the other trajectory; and *dynamic time warping (DTW)*, which computes the distance between two embedding sequences while allowing flexible alignment. (3) **LLM-based Consensus Estimators**, which directly evaluates the semantic agreement between reasoning trajectories using a large language model as a judge under hand-craft criteria. We use the policy model itself as consensus estimator.

Overall, semantic consensus estimators outperform or remain competitive with lexical overlap metrics, indicating that process-level agree-

Consensus Estimator	Usage	GPQA	MMLU-Pro	MedMCQA	MedQA	PubMedQA	MedBench	MedQA-Train
Reference Baselines								
Avg@16 (Random Sampling)		48.5	63.4	55.5	62.1	75.4	61.0	60.7
Pass@16 (Oracle Upper Bound)		83.6	84.8	84.6	88.6	88.6	86.0	87.7
Majority Voting (Answer MBR)		51.3	67.0	58.3	65.7	77.0	63.8	64.3
Lexical Consensus Estimators								
Jaccard	BoN	50.3	65.9	56.8	64.3	76.3	62.7	62.3
BLEU	BoN	52.8	66.3	57.0	64.4	76.4	63.4	62.1
ROUGE-1	BoN	51.5	65.1	56.4	63.3	75.8	62.4	62.2
Semantic Consensus Estimators								
Average Matching	BoN	52.6	65.7	57.0	64.9	77.2	63.5	63.5
	Weighted MV	52.6	66.4	58.0	65.8	77.1	64.0	64.5
DTW Matching	BoN	52.3	66.0	57.6	64.3	76.5	63.3	64.0
	Weighted MV	52.6	66.6	58.1	65.7	77.1	64.0	64.5
Greedy Matching	BoN	52.1	66.4	57.5	64.8	76.6	63.5	64.0
	Weighted MV	52.3	66.4	58.3	65.5	77.0	63.9	64.5
LLM-based Consensus Estimators								
LLM-Judge Agreement	BoN	53.3	67.3	57.9	65.7	76.9	64.2	64.4

Table 5: Extended comparison of consensus estimators for test-time response selection. Consensus signals are used either as standalone Best-of-N (BoN) selectors or combined with answer-level majority voting through weighted aggregation. MedQA-Train results are reported for reference and are not included in the test average.

ment is primarily reflected at the semantic rather than surface-text level. However, the differences among semantic matching strategies remain relatively modest, suggesting that the primary benefit comes from modeling consensus itself rather than from a specific matching algorithm.

In addition, consensus-based standalone selection mostly underperforms answer-level majority voting, while weighted aggregation provides more reliable gains. This further supports our conclusion that consensus captures complementary but non-sufficient correctness information.

Interestingly, stronger standalone test-time selectors do not necessarily lead to better RL training performance. In preliminary experiments, replacing DTW-based semantic consensus with an LLM-based consensus reward resulted in a lower post-training MedBench score (65.2 vs. 65.9), despite stronger standalone BoN selection performance. This suggests that training-time reward modulation and test-time response selection favor different properties of the consensus signal.

B Details of representative examples

Detailed representative examples for the four reasoning patterns are provided in Table 6.

C Process-level Evaluation Prompt

We evaluate the quality of reasoning processes using an LLM-based judge (GPT-4o). The judge is instructed to assess each response along four dimensions: *evidence grounding*, *medical soundness*, *logical coherence*, and *answer alignment*, each scored on a 1–5 scale. The full evaluation prompt used for the judge is provided below.

C.1 System Prompt

```
You are a rigorous medical reasoning evaluator.
Score ONLY the reasoning process quality, not merely
final answer correctness.
Penalize fabricated facts not supported by the question.
Be strict about medical soundness and internal
consistency.
```

C.2 User Prompt

```
You will evaluate the QUALITY of the model's reasoning
process for a medical QA task.
[Question]
{question}
[Model reasoning/process]
{reasoning}
[Ground Truth Answer]
{gt_answer}
Scoring rubric (each 1-5):
1) Evidence Grounding: Uses key clinical evidence from
the question; avoids missing critical facts.
2) Medical Soundness: Uses correct medical concepts;
avoids incorrect mechanisms, contraindications, or
fabricated facts.
3) Logical Coherence: Steps are consistent,
non-contradictory; avoids unjustified leaps.
4) Answer Alignment: Reasoning supports the final
answer.
```

<table>
<tr><td colspan="2">A 15-year-old boy presents with sudden onset right sided weakness of his arm and face and difficulty speaking. He denies any problems with hearing or comprehension. The patient has no history of chest pain, hypertension, or diabetes mellitus. No significant past medical history. The patient is afebrile, and vital signs are within normal limits. On physical examination, the patient is thin, with long arms and slender fingers. There is a right-sided facial droop present. Ophthalmic examination reveals a dislocated lens in the right eye. Strength is 3 out of 5 in the right upper extremity, and there is a positive Babinski reflex on the right. The CT scan of the head shows no evidence of hemorrhage. Laboratory findings are significant for increased concentrations of a metabolic intermediate in his serum and urine. Which of the following enzymes is most likely deficient in this patient?
A: Phenylalanine hydroxylase; B: Homogentisate oxidase; C: Cystathionine synthase; D: Branched-chain ketoacid dehydrogenase</td></tr>
<tr><td>High consensus incorrect | Similarity score: 0.777

"Dislocated lens i n the right eye: This is a characteristic finding in alkaptonuria... The buildup of homogentisic acid leads to the characteristic ochronosis and other symptoms, including dislocated lenses..."

This response forms a coherent but systematically flawed retrieval schema. The reasoning is internally structured and stable across samples, but incorrectly anchors lens dislocation and metabolic abnormalities to alkaptonuria, producing a high-consensus yet incorrect attractor.</td><td>High consensus correct | Similarity score: 1.000

"Thin, long arms, and slender fingers (suggesting Marfan syndrome)... Right-sided facial droop (suggesting a lesion in the right facial nerve, which could be a sign of a stroke... Dislocated lens in the right eye... C: Cystathionine synthase Deficiency in this enzyme causes Homocystinuria, which can cause mental retardation, skeletal abnormalities, and lens dislocation, fitting the patient's presentation"

This response organizes multiple clinical clues into a relatively stable and canonical disease retrieval pattern. The reasoning consistently integrates marfanoid habitus, ectopia lentis, metabolic disorder, and neurological findings toward homocystinuria/CBS deficiency.</td></tr>
<tr><td>Low consensus incorrect | Similarity score: 0.031

"The presence of a dislocated lens in the right eye (Leber's hereditary optic neuropathy) suggests a mitochondrial or metabolic disorder... Branched-chain ketoacid dehydrogenase deficiency... leads to neurological symptoms..."

This response shows fragmented and unstable reasoning. Multiple loosely connected disease concepts are introduced without a coherent diagnostic pathway, suggesting weak schema organization and low reasoning consistency.</td><td>Low consensus correct | Similarity score: 0.000

"Dislocated lens in the right eye... associated with some metabolic disorders... The presence of dislocated lens... is a well-known feature of Homocystinuria."

Although correct, the reasoning relies heavily on a single salient cue (ectopia lentis) with limited integration of other discriminative findings. Compared with high-consensus correct samples, the reasoning trajectory is more fragmented and shortcut-based.</td></tr>
</table>

Table 6: Representative example four consensus–correctness reasoning pattern. Similarity scores are normalized to $[0, 1]$.

```
Return STRICT JSON with keys:
{
"evidence_grounding": int,
"medical_soundness": int,
"logical_coherence": int,
"answer_alignment": int,
"overall": float,
"brief_rationale": string
}
Constraints:
- overall should be the average of the four integer
scores.
- brief_rationale <= 40 words.
- Output ONLY JSON, no extra text.
```