Unlocking Multimodal Mathematical Reasoning via Process Reward Model

Ruilin Luo*, **Zhuofan Zheng***, Lei Wang, Yifan Wang, Xinzhe Ni, Zicheng Lin,
Songtao Jiang, Yiyao Yu, Chufan Shi, Ruihang Chu, Jin Zeng †, Yujiu Yang †

Tsinghua University, ByteDance, Ping An Technology (Shenzhen) Co., Ltd., Zhejiang University

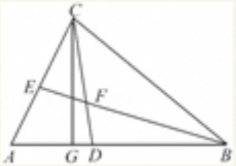
Thu 4 Dec, San Diego

Motivation

· The reasoning reliability of MLLMs has not significantly improved, and false positive examples are common.

· For real-world applications of MLLMs—such as photo-based question solving, route planning, and visual puzzle solving—not only the final result matters, but also the precise reasoning process.

· Reinforcement learning based solely on outcome supervision neglects the importance of process learning.



Problem:
In the diagram, point F is the intersection of the angle bisectors CD and BE of triangle ABC, and CG is perpendicular to AB at point G. If angle ACG is 32 degrees, what is the measure of angle BFC?

False-positive Rollout:

Step 1: Recognize that CD and BE are angle bisectors, so point F is the incenter of triangle ABC.

Step 2: Since CG is perpendicular to AB, triangle CGA is a right triangle.

Step 3: In triangle CGA, using the fact that the sum of angles is 180°, calculate angle CAG: angle CAG = 90° - angle ACG =90° - 32° = 58°.

Step 4: Angle BFC = 180° - angle CDE. CD is the angle bisector of angle ACB, angle ACD = angle BCD = 1/2 angle ACB.

Step 5: We have angle CDE = angle ACD = angle BCD = 1/2 angle ACB.

Step 6: Since angle ACB + angle ACG = 180°, angle ACB = 180° - 58° = 122°.

Step 7: Substitute angle ACB = 122°, angle ACD = 1/2 * 122° = 61°.

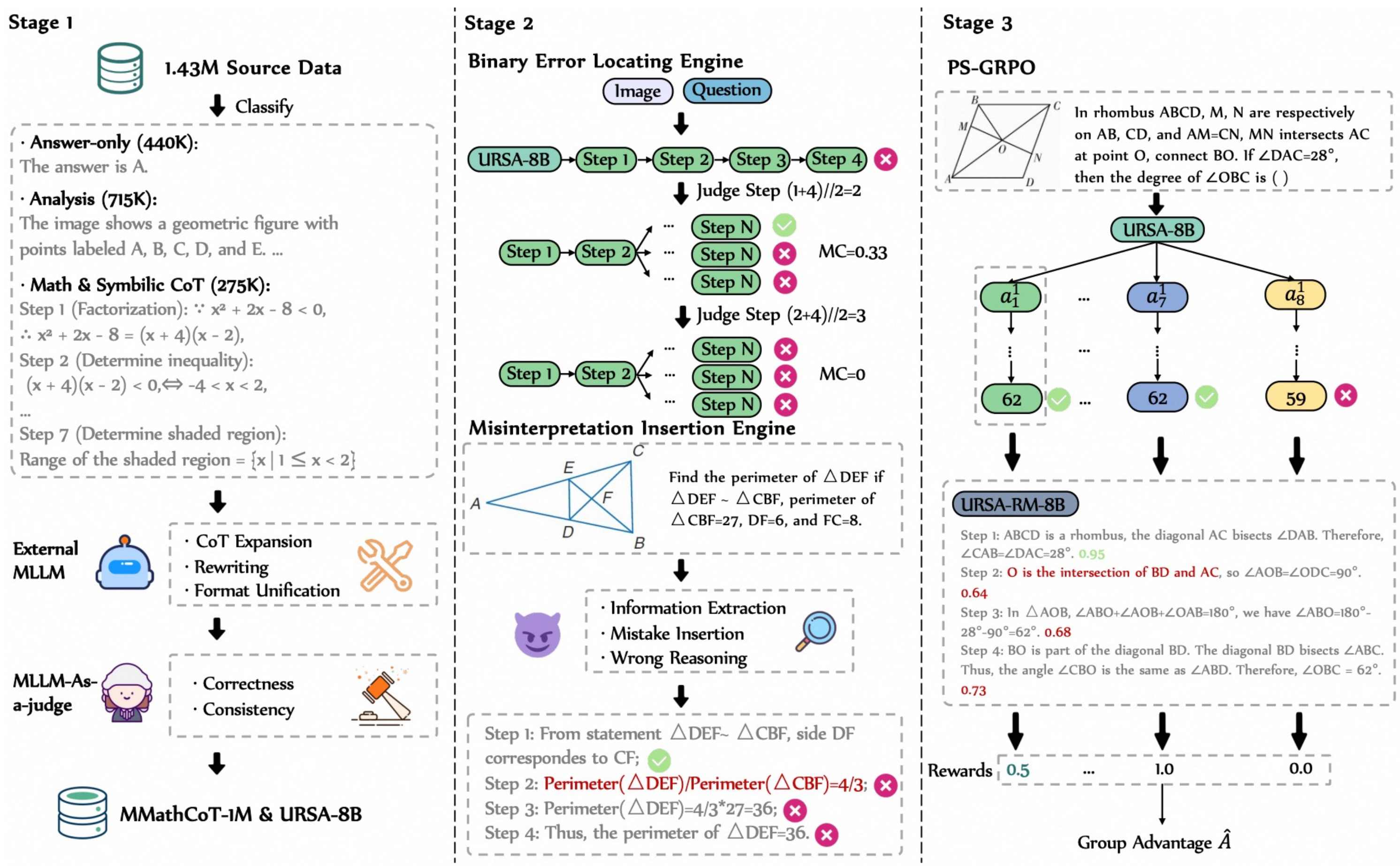
Step 8: Substitute angle ACD = 66°, angle BFC = 180° - 61° = 119°.

†Answer: 119 ° ✓

! Examples where the answer is correct but the reasoning process is flawed or incoherent.

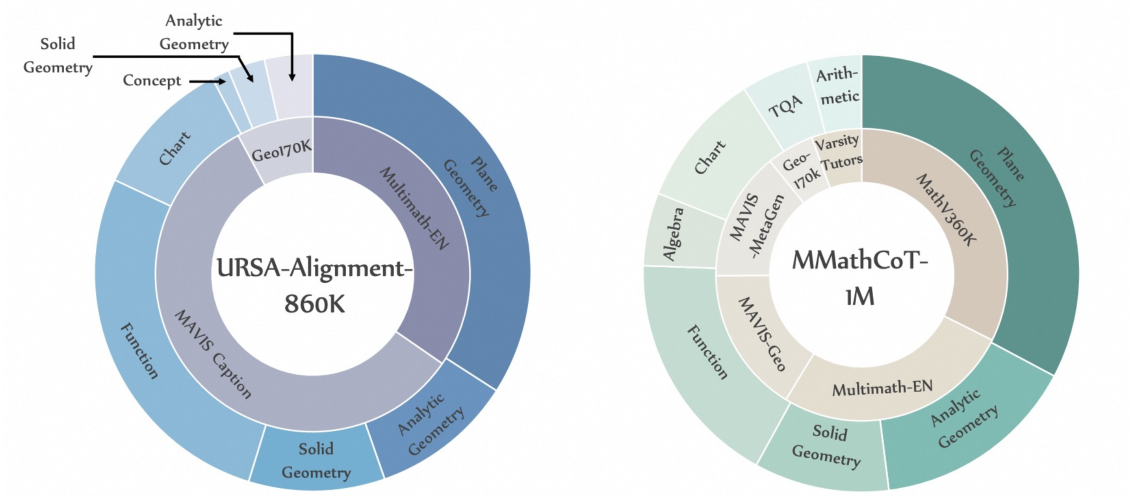
Question & Solution on Applying PRM

- On what?
  Models with good reasoning diversity.
- How to?
  Simultaneously focus on Monte Carlo estimates based on logical correctness and consistency with visual-textual information.
- When to?
  Test-time expansion and online reinforcement learning.



Data Curation

- Answer-only: Invoke an external model for Chain-of-Thought (CoT) expansion
- Analysis: Rewriting and organization CoT-formatted: Unified format
- CoT-formatted: Unified format for symbolic reasoning paradigm



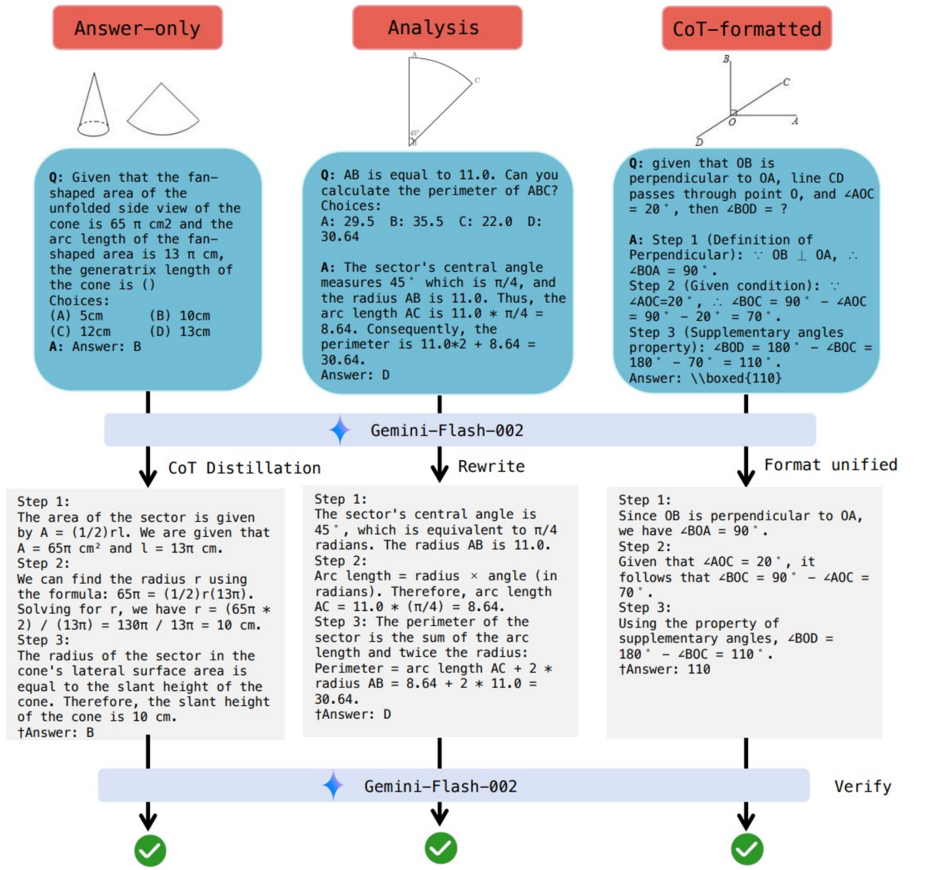
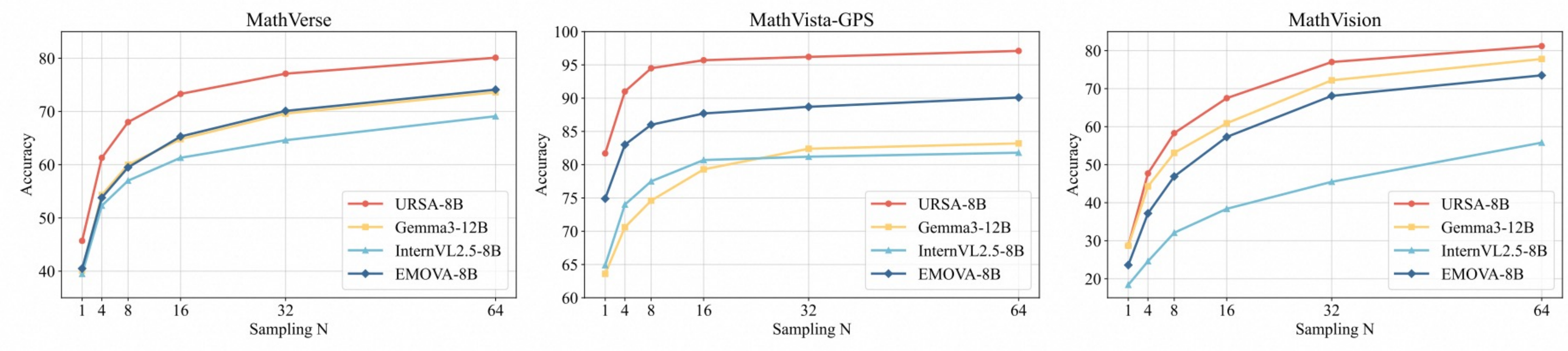


Figure 3: CoT augmentation and verifying for multimodal mathematical data from three type of sources using Gemini-1.5-Flash-002.

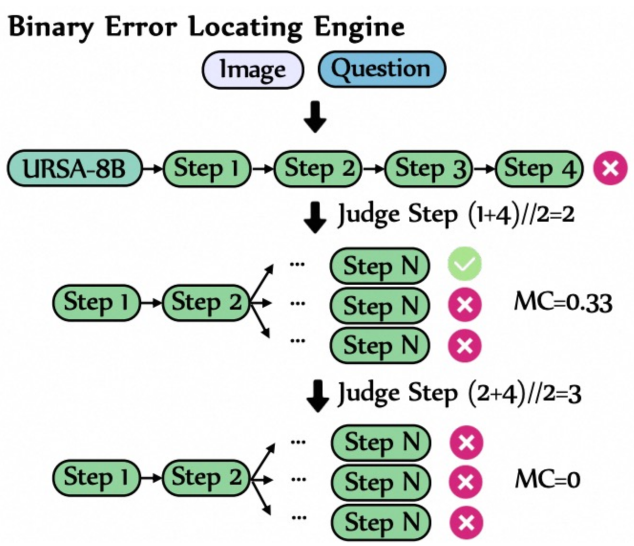
Data Curation - Outcome

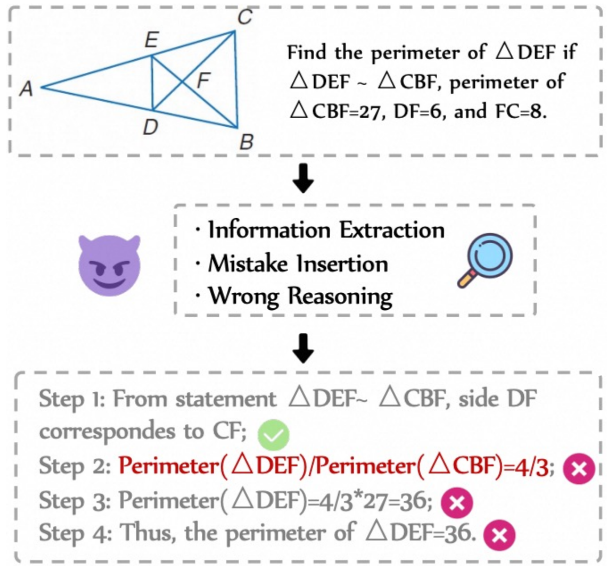
- High-quality reasoning data directly optimizes reasoning diversity.
- A high Pass@N metric ensures the value introduced by process-supervised models.



Process Labeling Data Scaling

- Monte Carlo estimation provides a relative advantage in guiding steps toward the correct path.
- Hallucination injection offers an absolute advantage in aligning visual and textual information within multimodal scenarios.





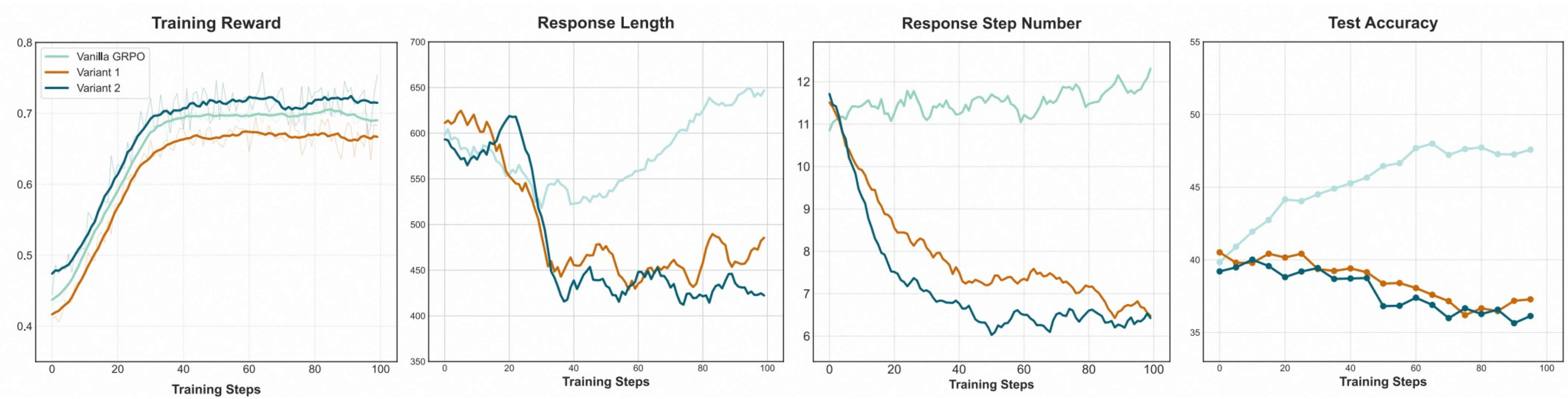
PRM-integrated Online RL

Two augmented baselines:

$$r^i = r_o^i + \bar{r}_s^i.$$

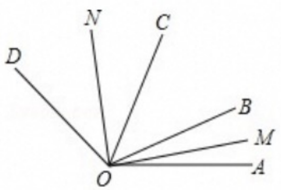
$$A_t^i = r_{s,t}^i \underbrace{\frac{\bar{r}_s^i - \text{mean}(\{\bar{r}_s^j\}_{j=1}^G)}{\text{std}(\{\bar{r}_s^j\}_{j=1}^G)}}_{\text{GRPO with process rewards}} + \underbrace{\frac{r_o^i - \text{mean}(\{r_o^j\}_{j=1}^G)}{\text{std}(\{r_o^j\}_{j=1}^G)}}_{\text{GRPO with outcome rewards}}$$

Rewarding length bias
Reward hacking



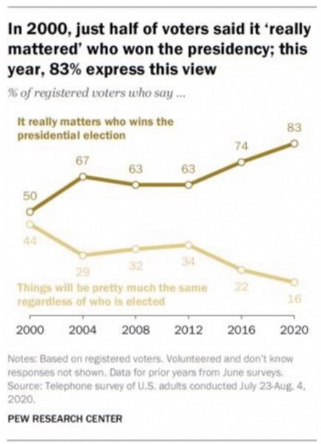
Short CoT Pattern

The model tends to learn “**perception, describe and short reasoning**” thinking pattern.



Question:
As shown in the figure: ∠AOB : ∠BOC : ∠COD = 2 : 3 : 4, ray OM and ON bisect ∠AOB and ∠COD respectively, and ∠MON = 84°, then ∠AOB is ( ).

Solution:
Step 1: Look at the diagram and note the angles ∠AOB, ∠BOC, and ∠COD are in the ratio 2:3:4.
Step 2: Recognize OM and ON bisect ∠AOB and ∠COD, creating smaller angles.
Step 3: Use the given ∠MON = 84° to set up an equation based on the ratios and angle bisectors.
Step 4: Solve the equation ∠AOB = 2 / (2 + 3 + 4) * 84° = 18.66°.
†Answer: ∠AOB is 18.66°.

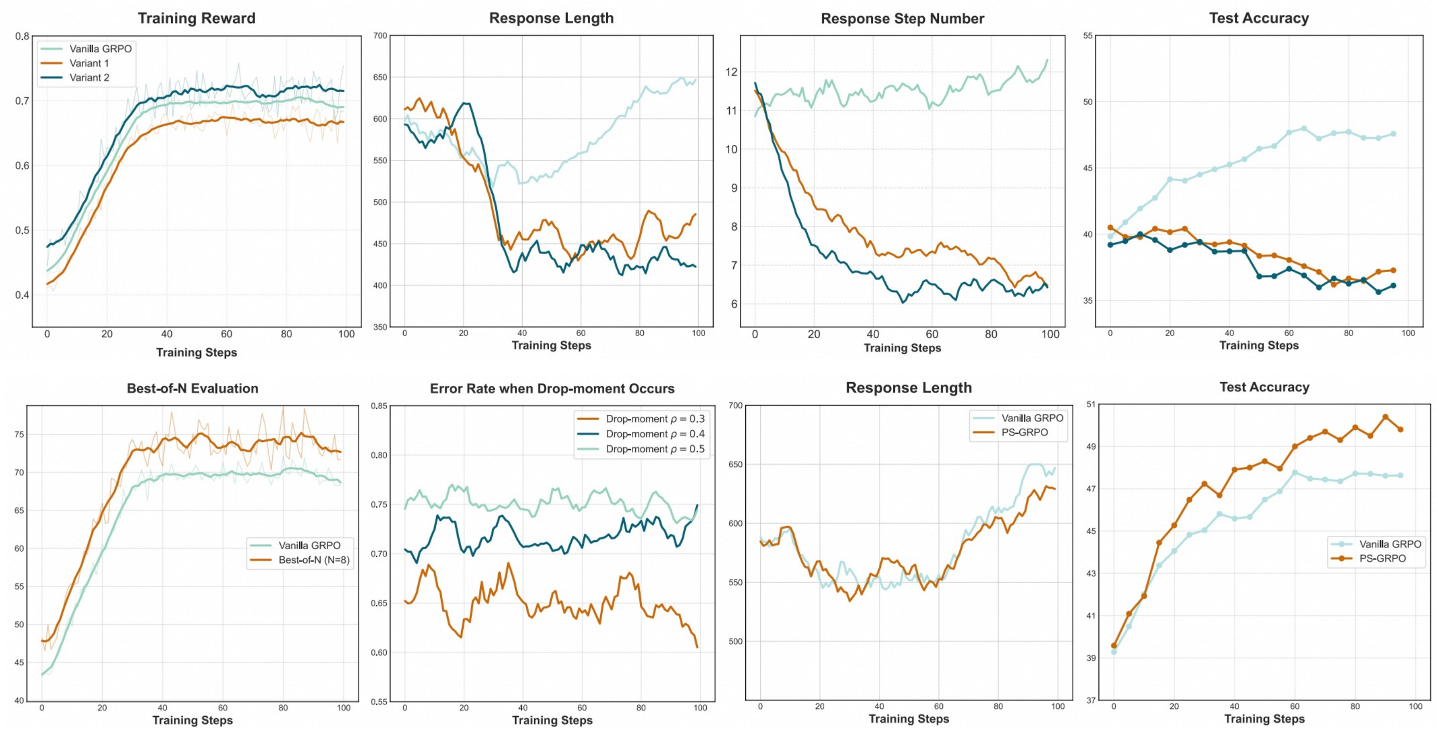


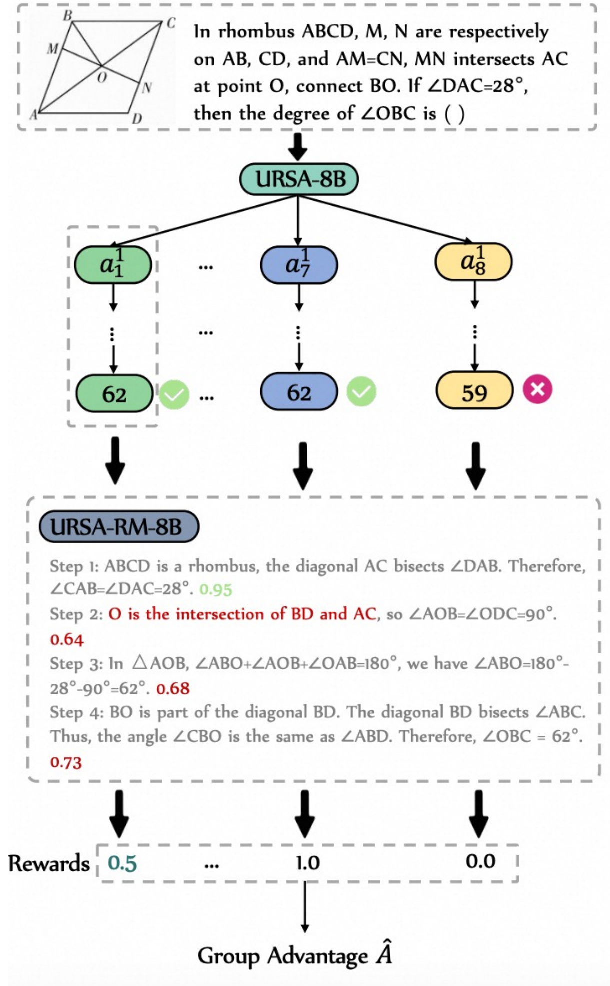
Question:
As shown in the figure: ∠AOB : ∠BOC : ∠COD = 2 : 3 : 4, ray OM and ON bisect ∠AOB and ∠COD respectively, and ∠MON = 84°, then ∠AOB is ( ).

Solution:
Step 1: The dark brown line represents voters who think it really matters, showing an increasing trend from 50% in 2000 to 83% in 2020. The light brown line represents voters who think things will be much the same, showing a decreasing trend from 44% in 2000 to 16% in 2020.
Step 2: We need to know the least value of the light brown graph to the leftmost value of the dark brown graph.
Step 3: The ratio of the least value of the light brown graph and the leftmost value of the dark brown graph is 16:67.
†Answer: 16:67.

Process-to-outcome Rewarding

- Traditional process reward design can lead to reward hacking or reward length bias, inducing "passive" reasoning.
- The "Drop-moment" of process rewards is transformed into outcome reward penalties that outperform naive GRPO.





Q-value Ranking Approximation

Theoretical Background In a multi-step reasoning Markov Decision Process, the optimal Q-function, $Q^*(s_i, a_i)$, is the expected probability of reaching the correct final answer after action a_i. The scalar reward r_{si} from our URSA-8B-RM at each step serves as an empirical estimate, Q_{est}, of this true Q-value.

Theoretical Optimal Objective Approximating the ideal Q-value Ordering, under an optimal reasoning policy π^*, the Q-values along a reasoning path should satisfy a strict monotonic ordering: $Q^*_{w_{|W|}} < \cdots < Q^*_{w_1} \ll V^*(x) < \cdots < Q^*_{c_1} < Q^*_{c_{|C|}}$, where c_i and w_j denote the i-th correct step and the j-th incorrect step, respectively. In our drop-moment definition, $c_{|C|}$ is adjacent to w_1, which is consistent. The ideal training objective for a Process Reward Model (PRM) is to learn a Q-estimator Q_{est} that strictly follows this ordering across all reasoning paths. For a successful reasoning path τ (e.g., $o(\tau) = 1$), this implies that the sequence of Q_{est} outputs $\{r_{p,1}, r_{p,2}, \ldots, r_{p,T}\}$ should be monotonically increasing with no decreases, as a successful path should not contain any error.

Thus, the theoretically optimal generation objective is:

$$J_{ideal}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[o(\tau)] - \lambda \cdot \Omega(G(\tau))$$

where:

- $\mathbb{E}_{\tau \sim \pi_\theta}[o(\tau)]$ is the standard final result accuracy.
- $\Omega(G(\tau))$ is a non-decreasing regularization term measuring the violation of the ideal Q-value ordering. λ is its strength.
- $G(\tau) = r_{c_l}(\tau) - r_{w_1}(\tau)$ is a direct measure of the violation of the theoretical gap. The larger $G(\tau)$, the greater the drop from high (correct) to low (incorrect) rewards at some point in the path, which severely violates the ideal ordering. Minimizing $G(\tau)$ is equivalent to aligning the model's generation process with the Q-value dynamics of the optimal reasoning path.

PS-GRPO is a proxy implementation of Q-Value Ranking Rule

PS-GRPO as a Proxy Implementation PS-GRPO introduces a computable, differentiable proxy for $\Omega(G(\tau))$ into its reward function:

$$R(\tau) = o(\tau) \cdot (1 - \gamma \cdot \mathbb{I}(\delta_p(\tau) \geq \rho))$$

This function approximates the ideal objective $J_{ideal}(\theta)$:

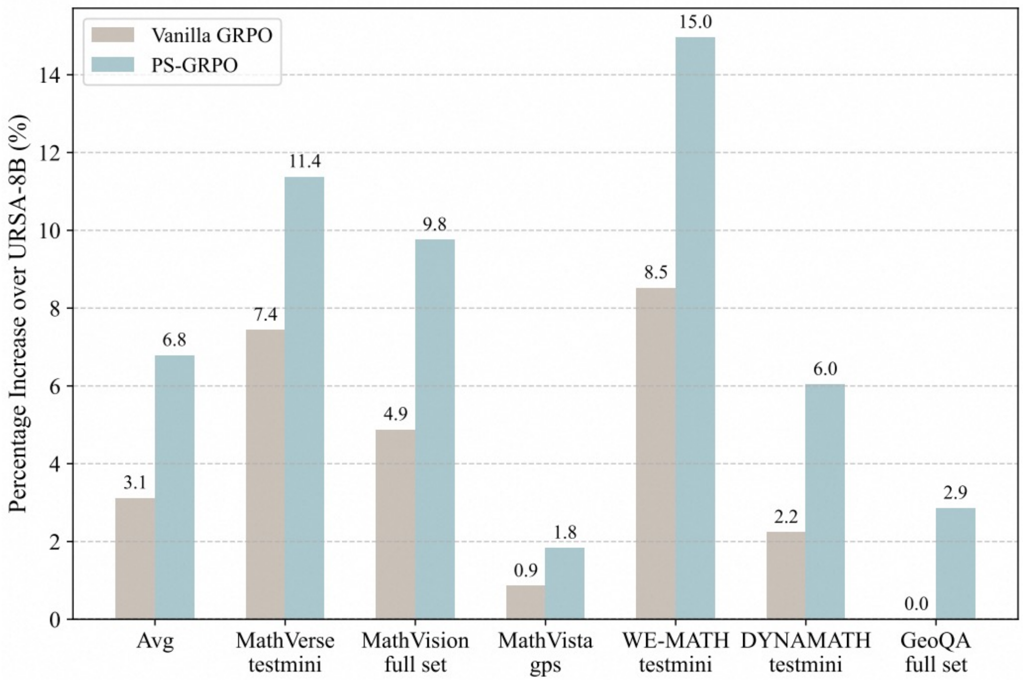
- The $o(\tau)$ term directly corresponds to the first term in J_{ideal}, maximizing final accuracy.
- The penalty term approximates the regularizer:
  - $\delta_p(\tau)$ is a computationally tractable and strongly correlated proxy for $G(\tau)$. A large δ_p indicates a large G.
  - The indicator function $\mathbb{I}(\delta_p(\tau) \geq \rho)$ converts the continuous δ_p into a binary signal, resolving non-differentiability and ensuring compatibility with GRPO.
  - The hyperparameter ρ acts as the violation threshold for $\Omega(G(\tau))$ and is determined by the behavior of Q_{est}.

Therefore, the PS-GRPO reward function $R(\tau)$ serves as a practical proxy for J_{ideal}: $J_{PS-GRPO}(\theta) \approx J_{ideal}(\theta)$. The final step is to empirically determine ρ for a given PRM.

Empirical Results

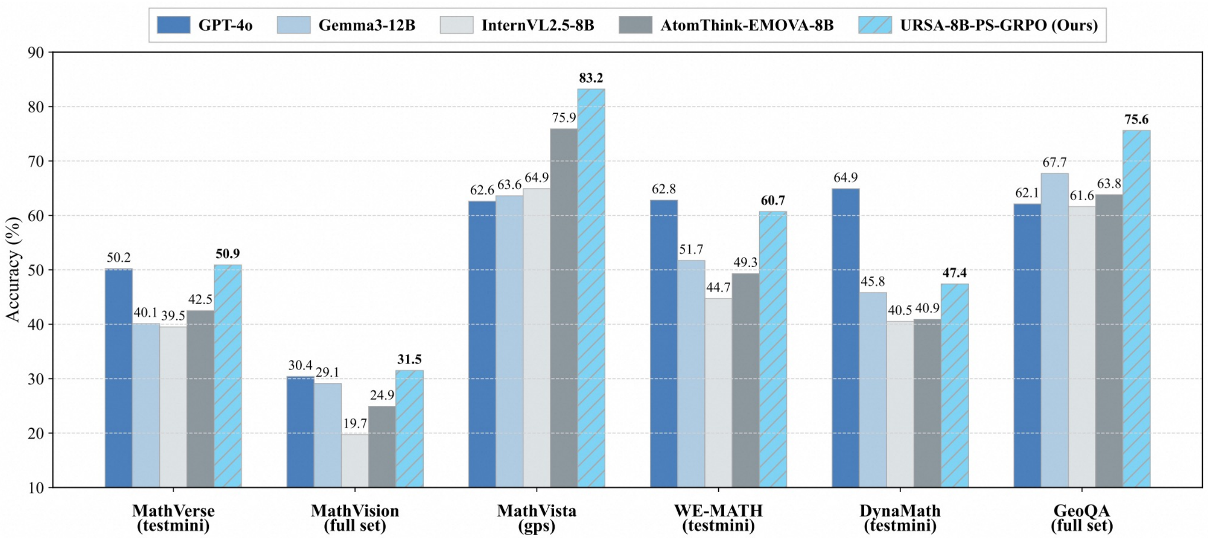
- URSA-8B-PS-GRPO outperforms GPT-4o on six reasoning benchmarks with a 7B model
- PS-GRPO surpasses vanilla GRPO

	Size	Avg	MathVerse testmini	MathVision full set	MathVista gps	WE-MATH testmini	DYNAMATH testmini	GeoQA full set
Closed-Source MLLMs								
GPT-4o [59]	-	55.5	50.2	30.4	64.7	62.8	64.9	62.1
GPT-4o-mini [59]	-	49.2	42.3	22.8	59.9	56.3	53.5	60.1
Gemini-1.5-pro [60]	-	53.2	35.3	19.2	81.7	66.9	60.5	55.5
Open-Source General MLLMs								
InternVL-Chat-V1.5 [61]	26B	33.6	26.1	15.4	56.9	32.7	36.7	33.5
Llama-3.2-11B-Vision-Instruct [62]	11B	28.0	28.9	16.9	40.9	12.0	32.2	36.9
Qwen2-VL [63]	8B	40.2	33.6	19.2	51.0	43.0	42.1	52.2
InternVL2-8B [64]	8B	41.8	37.0	18.4	57.7	44.9	39.7	52.8
InternVL2-8B-MPO [65]	8B	45.1	38.2	22.3	69.2	44.4	40.5	55.9
InternVL2.5-8B [66]	8B	45.2	39.5	19.7	64.9	44.7	40.5	61.6
LLaVA-OneVision [35]	8B	40.9	28.9	18.3	71.6	44.9	37.5	43.9
Points-Qwen2.5-Instruct [67]	8B	49.8	41.1	23.9	76.0	51.0	42.8	63.8
Gemma3-12B [68]	12B	49.8	40.1	29.1	63.6	51.7	45.8	67.7
Open-Source Reasoning MLLMs								
Math-LLaVA [15]	13B	35.2	22.9	15.7	57.7	31.3	35.5	48.1
MathPUMA-Qwen2-7B [11]	8B	39.6	33.6	14.0	48.1	41.0	37.3	63.6
MultiMath [23]	7B	43.1	27.7	16.3	66.8	42.2	37.9	67.7
MAVIS [19]	7B	44.4	35.2	18.5	64.1	44.3	36.2	68.3
InfiMM-Math [14]	7B	48.6	40.5	18.8	77.3	48.3	38.2	68.3
AtomThink-EMOVA [12]	8B	49.5	42.5	24.9	75.9	49.3	40.9	63.8
MathGLM-Vision [9]	9B	47.6	44.2	19.2	64.2	45.2	42.2	70.4
LlamaV-o1 [69]	11B	38.4	33.9	17.9	53.3	42.6	34.7	43.1
OpenVLThinker [70]	7B	-	47.9	25.3	76.4	-	-	-
R1-Onevision [71]	7B	-	47.4	26.9	72.4	51.4	-	-
URSA-8B	8B	54.7	45.7	28.7	81.7	53.6	44.7	73.5
URSA-8B-PS-GRPO	8B	58.2	50.9	31.5	83.2	60.7	47.4	75.6



Leaderboard Comparison

- With nearly 1.9M training data points, it outperforms InternVL3-9B among models of comparable size on the OpenCompass public reasoning benchmark.
- It also surpasses larger models such as GPT-4o, Kimi-VL-16B-A3-MOE, and InternVL2.5-38B on MathVision.



	Method	Eval Time	Params	Language Model	Vision Model	Avg. Score
1	VLAA-Thinker-Qwen2.5... Open Source · VLAA@UCSC&UTD	2025/04/02	8.29B	Qwen2.5-7B	QwenViT	42.5
2	InternVL3-8B Open Source · Shanghai AI Laboratory & Tsinghua...	2025/04/14	7.94B	Qwen2.5-7B	InternViT-300M-v2.5	41.4
3	URSA-8B-PS-GRPO Open Source · ByteDance & Tsinghua University	2025/05/24	8.04B	Qwen2.5-Math-7B	SAM-B SigLIP-L	41.1
4	InternVL3-9B Open Source · Shanghai AI Laboratory & Tsinghua...	2025/04/14	9.14B	InternLM3-8B	InternViT-300M-v2.5	40.8
5	Qwen2.5-VL-7B Open Source · Alibaba	2025/02/02	8.29B	Qwen2.5-7B	QwenViT	40.1

Open-sourcing Contribution

- As visual stem RL data for Seed-1.5VL, this paragraph is cited separately.

4.3.1 Visual STEM

STEM (science, technology, engineering, and mathematics) questions usually have unique and verifiable answers, which are suitable for RLVR. We collect over one million problems with images in STEM fields, mostly on mathematics, from both open-sourced resources [85] and internal K-12 education collections.

To prepare the training data, multiple-choice questions were initially transformed into an open-ended format by removing the choices, thus forcing the model to generate the correct answer's content and preventing random guessing. Subsequently, difficult questions were selected via rejection sampling based on the performance of the SFT model. We carefully remove questions that can be answered by text only or text and captions, ensuring shortcuts on text or superficial visual elements will not be reinforced in RL. Specifically, 16 responses were generated per question, and questions achieving either 0% or greater than 75% accuracy with the SFT model were discarded. This filtering isolates challenging prompts ($0\% < \text{accuracy} \leq 75\%$) appropriate for RLVR exploration while removing potentially erroneous or trivial questions. Lastly, a preamble instruction was prepended to prompts, instructing the model to format the final answer using designated LaTeX identifiers (e.g., `\boxed{answer}`) to enable straightforward automated extraction.

Our STEM verifier transforms the predicted answers into a sympy expression and matches it with ground truths. To ensure the accuracy of our verifier, we also remove prompts that contain multiple questions or whose ground truths are complex phrases.

[85] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. arXiv preprint arXiv:2501.04686, 2025.

Thanks!