

# The Bidirectional Process Reward Model

Lingyin Zhang<sup>1,2</sup>, Jun Gao<sup>3</sup>, Xiaoxue Ren<sup>3,4</sup>, Ziqiang Cao<sup>1,2\*</sup>

<sup>1</sup>School of Computer Science and Technology, Soochow University

<sup>2</sup>Biomedical Basic Research Center of Jiangsu, Soochow University, Suzhou, Jiangsu 215123, China

<sup>3</sup>School of Software Technology, Zhejiang University

<sup>4</sup>Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security

lyzhanglyzhang@stu.suda.edu.cn, jgao1106@zju.edu.cn, xxren@zju.edu.cn, zqcao@suda.edu.cn

## Abstract

Process Reward Models (PRMs), which assign fine-grained scores to intermediate reasoning steps within a solution trajectory, have emerged as a promising approach to enhance the reasoning quality of Large Language Models (LLMs). However, most existing PRMs rely on a unidirectional left-to-right (L2R) evaluation scheme, which restricts their utilization of global context. In light of this challenge, we propose a novel bidirectional evaluation paradigm, named **Bidirectional Process Reward Model (BiPRM)**. BiPRM incorporates a parallel right-to-left (R2L) evaluation stream, implemented via prompt reversal, alongside the conventional L2R flow. Then a gating mechanism is introduced to adaptively fuse the reward scores from both streams to yield a holistic quality assessment. Remarkably, compared to the original PRM, BiPRM introduces only a 0.3% parameter increase for the gating module, and the parallel execution of two streams incurs merely 5% inference time latency. Our extensive empirical evaluations spanning diverse benchmarks, LLM backbones, PRM objectives and sampling policies demonstrate that BiPRM consistently surpasses unidirectional baselines, achieving an average relative gain of 10.6% over 54 solution-level configurations and 37.7% in 12 step-level error detection scenarios. Generally, our results highlight the effectiveness, robustness and general applicability of BiPRM, offering a promising new direction for process-based reward modeling.<sup>1</sup>

## 1 Introduction

Process reward models (PRMs) aim to augment the response quality of policy models by granularly scoring a sequence of intermediate steps (Lightman et al., 2023). Currently, PRMs have been widely applied to align reinforcement learning (RL) models

with human preferences (Lightman et al., 2023), as well as to support Test-Time Scaling (TTS) (Jaech et al., 2024; Guo et al., 2025) strategies in a spectrum of complex cognitive tasks such as mathematical reasoning (Lightman et al., 2023; Li and Li, 2024; Zhang et al., 2024; Xia et al., 2025; Ma et al., 2025), code generation (Le et al., 2022; Li et al., 2024a) and question answering (Carta et al., 2022; Bi et al., 2023).

Existing studies typically formulate PRM training via various objectives, ranging from classification tasks using binary cross-entropy (BCE) (Wang et al., 2024b; Shao et al., 2024; Luo et al., 2024; Lightman et al., 2023) to regression tasks employing mean squared error (MSE) (Zhang et al., 2024; Wang et al., 2024a) or Q-value rankings loss (Li and Li, 2024) for more fine-grained supervision. Since PRMs are typically built upon generative policy models and share the same parameter initialization, existing works predominantly adopt a unidirectional left-to-right (L2R) evaluation paradigm, where step plausibility is assessed sequentially based on preceding context. However, this paradigm inherently restricts access to global context (Figure 1, Left), which is critical when an early step’s validity hinges on downstream consequences. For instance, in the mathematical induction case shown in Table 7, determining the correctness of the hypothesis in Step 2 necessitates verifying the derivation logic in Steps 3 and 4. By relying exclusively on past context, standard PRMs miss these essential backward signals, thereby constraining their global optimization capability (Liu et al., 2024). Although recent studies such as BiRM (Chen et al., 2025) attempt to incorporate future guidance via an auxiliary value head, they remain fundamentally forward-looking predictors, without performing direct retrospective verification through structural context reversal.

To address this limitation, we propose the Bidirectional Process Reward Model (BiPRM), a novel

\* Corresponding author.

<sup>1</sup> <https://github.com/LingyinZhang/BiPRM>

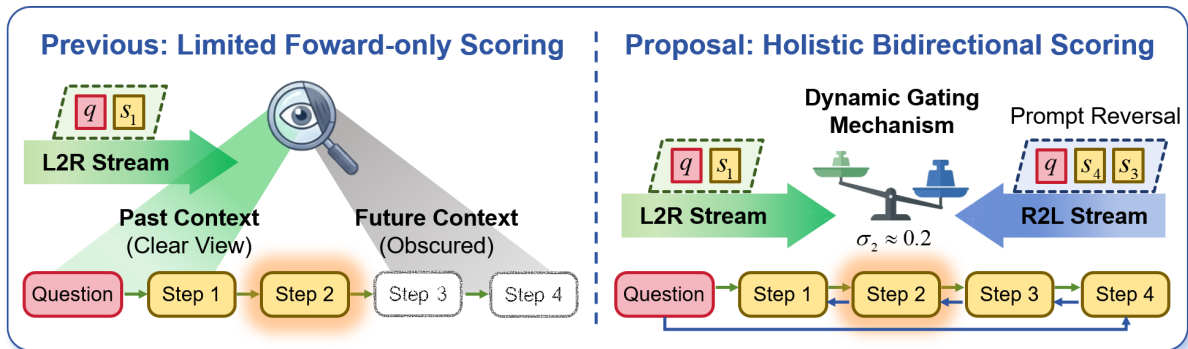


Figure 1: Comparison of evaluation paradigms when scoring **Step 2**. (Left) Conventional unidirectional PRMs are restricted to past context, lacking access to subsequent steps required for verification. (Right) BiPRM integrates a parallel R2L stream to incorporate future context, enabling effective retrospective verification of the current step.

evaluation paradigm for PRMs that draws inspiration from the architecture of Bidirectional Long Short-Term Memory (BiLSTM) networks (Graves, 2012). As shown in Figure 1 (Right), BiPRM synergizes a parallel R2L evaluation stream with the conventional L2R flow, allowing subsequent steps to provide retrospective evidence for earlier ones. Noting that increasing context length is generally associated with improved reliability of reasoning (Figure 3a), we adopt a dynamic gating mechanism, rather than static averaging, to integrate scores from the two streams. Since the R2L stream is efficiently realized by simply reversing the reasoning trajectory via prompt modifications, BiPRM requires only a 0.3% parameter increase for the gating module. Through parallel execution, BiPRM effectively mitigates the computational overhead of dual-stream evaluation, incurring only an approximate 5% increase in wall-clock latency while significantly enhancing verification accuracy.

We comprehensively conduct experiments on two solution-level benchmarks (GSM-Plus (Li et al., 2024c) and MATH500 (Hendrycks et al., 2021)) and one diagnostic step-level benchmark (ProcessBench (Zheng et al., 2025)). Evaluating our method across three backbones of varying scales and three PRM objectives, experimental results demonstrate that BiPRM consistently outperforms unidirectional baselines. Specifically, BiPRM achieves an average relative gain of 10.6% across 54 solution-level configurations and 37.7% across 12 step-level error detection scenarios. These consistent improvements confirm that the bidirectional evaluation paradigm provides a robust modeling advantage regardless of model capacity or training objective.

In summary, our contributions are as follows:

- As far as we know, we are the first to propose the concept of the bidirectional evaluation paradigm for PRMs, addressing unidirectional models’ limitations in local perspective through dual-stream fusion.
- BiPRM achieves an excellent performance-efficiency trade-off, adding only about 5% inference latency by running the R2L stream in parallel.
- As a general framework, BiPRM can adapt to diverse PRM architectures, offering new directions for future process reward modeling.

## 2 Related work

### 2.1 Process Reward Models

Process Reward Models (PRMs) improve the alignment of language models with human reasoning preferences by providing fine-grained feedback at intermediate reasoning steps rather than evaluating only final answers (Lightman et al., 2023). Early works such as PRM800K (Lightman et al., 2023) utilized large-scale human annotations to train token-level step classifiers. Subsequent approaches have explored automated supervision to reduce annotation costs, estimating step quality via Monte Carlo rollouts (Wang et al., 2024b,c) or modeling structured trajectories using Monte Carlo Tree Search (Luo et al., 2024; Zhang et al., 2024). Recent advancements include ranking-based objectives (Li and Li, 2024) and advantage-based reward definitions (Lu et al., 2024; Setlur et al., 2024). Despite these innovations, most existing PRMs adhere to a strict L2R evaluation paradigm. This unidirectional constraint inherently limits the

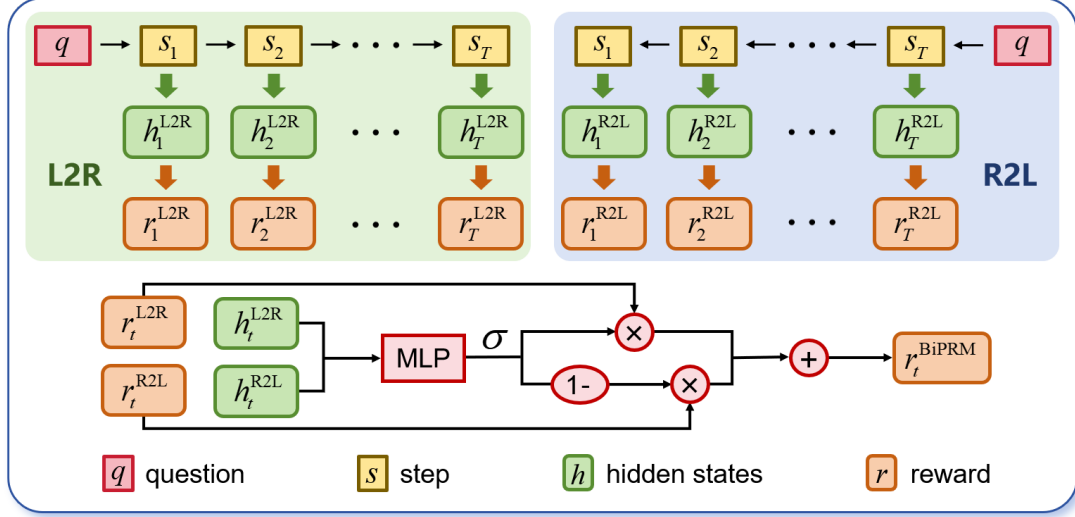


Figure 2: Overview of the BiPRM Architecture. The model synthesizes parallel L2R and R2L evaluation flows via a dynamic gating mechanism for step-wise reward modeling.

model’s ability to leverage global context for verifying the consistency of the entire trajectory. Moreover, the increasing trajectory complexity introduced by emerging multimodal contexts (Gao et al., 2025a) and efficient in-context learning (Gao et al., 2025b,c) exacerbates the blind spots of unidirectional PRMs, further underscoring the necessity of our bidirectional framework.

## 2.2 Order of Reasoning

The order of reasoning has been shown to significantly influence model performance. While early studies suggested a symmetry between forward and backward prediction (Shannon, 1951), recent work indicates that forward generation typically yields lower perplexity (Papadopoulos et al., 2024). However, non-monotonic and bidirectional approaches have shown promise in specific domains. For instance, reversing input sequences has proven effective in machine translation (Sutskever et al., 2014), and bidirectional encoders like BERT (Devlin et al., 2019) excel at capturing contextual dependencies. In reasoning tasks, backward chaining can sometimes outperform forward deduction (Vinyals et al., 2015; Pfau et al., 2023). Furthermore, diffusion-based models (Zhang et al., 2023; Li et al., 2022; Gong et al., 2024) demonstrate the efficacy of planning in both directions. Building on these insights, we propose integrating a backward verification stream into reward modeling to leverage the complementary strengths of bidirectional reasoning.

## 3 Methodology

### 3.1 Preliminary

Given a mathematical problem  $q$ , a policy language model generates a reasoning trajectory to solve the task. This trajectory, often referred to as a Chain-of-Thought (Wei et al., 2022), is formally represented as  $\tau = (q, \{s_1, \dots, s_T\})$ , where  $s_t$  denotes the  $t$ -th reasoning step and  $s_T$  contains the final answer. Standard PRMs typically adopt a unidirectional L2R evaluation paradigm, where each step  $s_t$  is scored sequentially based on the question and prior context. The L2R reward function for step  $s_t$  is expressed as

$$r_t^{\text{L2R}} = f_\theta(s_t | q, s_{<t}), \quad (1)$$

where  $f_\theta$  denotes the reward function parameterized by  $\theta$ . Under this formulation, the reward  $r_t^{\text{L2R}}$  is strictly independent of future steps  $s_{>t}$ , leading to the following gradient property:

$$\frac{\partial r_t^{\text{L2R}}}{\partial s_{t+k}} = 0 \quad \forall k \in [1, T - t]. \quad (2)$$

This zero gradient prevents future signals from propagating backward to facilitate retrospective verification, thereby restricting standard PRMs to a local perspective that lacks global consistency.

### 3.2 Implementation of BiPRM

To address the aforementioned limitation, we draw inspiration from BiLSTM and develop BiPRM, a bidirectional process reward model that incorporates a backward evaluation signal in parallel with

the conventional L2R scoring path. As illustrated in Figure 2, BiPRM constructs a backward trajectory by logically inverting the reasoning sequence to  $(q, \{s_T, s_{T-1}, \dots, s_1\})$ . This step-level reversal mimics the human cognitive process of retrospective verification, checking derivations from conclusions back to premises. Consequently, the R2L reward function for a specific step  $s_t$  is defined as

$$r_t^{\text{R2L}} = f_\theta(s_t | q, s_{>t}). \quad (3)$$

Notably, this R2L stream is implemented entirely via prompt reversal with no architectural modifications to the underlying LLM, requiring only a negligible 0.3% parameter increase for the gating module.

To synthesize the bidirectional signals, BiPRM employs a step-wise dynamic gating mechanism. Our empirical analysis (Section 4.4) reveals that L2R and R2L streams possess complementary error profiles, where L2R excels in later stages while R2L is more effective early on. To exploit this position-dependent synergy, we compute a dynamic weight  $\sigma_t$  for each step  $t$  using a lightweight MLP:

$$\sigma_t = \text{Sigmoid}(\text{MLP}([h_t^{\text{L2R}}; h_t^{\text{R2L}}])), \quad (4)$$

where  $h_t^{\text{L2R}}$  and  $h_t^{\text{R2L}}$  are the hidden states from the respective streams. The final bidirectional reward for each step  $t$  is computed as:

$$r_t^{\text{BiPRM}} = \sigma_t \cdot r_t^{\text{L2R}} + (1 - \sigma_t) \cdot r_t^{\text{R2L}}. \quad (5)$$

This bidirectional formulation ensures that the evaluation of each step is informed by the global context, allowing future information to influence current scoring:

$$\frac{\partial r_t^{\text{BiPRM}}}{\partial s_{t+k}} = \sigma_t \frac{\partial r_t^{\text{L2R}}}{\partial s_{t+k}} + (1 - \sigma_t) \frac{\partial r_t^{\text{R2L}}}{\partial s_{t+k}} \neq 0. \quad (6)$$

Finally, to obtain the trajectory-level reward, we apply a reduction operation over the sequence of stepwise scores using an aggregation operator  $\oplus \in \{\prod, \min, \max, \text{mean}\}$ . While a comprehensive sensitivity analysis of these different operators is provided in Appendix F, we adopt the minimum operator following previous works (Li and Li, 2024; Wang et al., 2024b). This strategy is grounded in the logical "weakest link" principle, which encourages the model to filter out trajectories containing even a single fatal error. The overall trajectory reward is then given by

$$R_{\text{BiPRM}}(\tau) = \min_{t=1}^T r_t^{\text{BiPRM}}. \quad (7)$$

Due to the parallel execution of the two evaluation streams, BiPRM enhances verification effectiveness while incurring only an approximate 5% increase in inference time latency.

### 3.3 Training Objectives

Since the bidirectional evaluation framework we proposed does not change the underlying architecture of PRMs, this paper tests the following three existing objectives. The first is the **Binary Cross-Entropy (BCE)** loss, which treats reward prediction as a classification problem. It is defined as follows:

$$\mathcal{L}_{\text{BCE}}(\tau) = -\frac{1}{T} \sum_{t=1}^T (r_t \ln(\hat{r}_t) + (1-r_t) \ln(1-\hat{r}_t)), \quad (8)$$

where  $r_t$  is the gold classification label of  $t$ -th step and  $\hat{r}_t$  is the predicted reward. This loss is widely used due to its simplicity and alignment with classification-based annotations.

The second objective is the **Mean Squared Error (MSE)** loss, formulated as

$$\mathcal{L}_{\text{MSE}}(\tau) = -\frac{1}{T} \sum_{t=1}^T (\hat{r}_t - r_t)^2, \quad (9)$$

which penalizes large deviations between the predicted and target reward values. While MSE allows for finer-grained supervision, it is more sensitive to outliers in label distributions.

The third objective follows the **Q-value rankings loss** proposed in recent work (Li and Li, 2024), formulated as

$$\begin{aligned} \Sigma_t &= \sum_{q=0}^t \exp(\hat{r}_{c_q}) + \sum_{w \in W} \exp(\hat{r}_w + \zeta), \\ \mathcal{L}_{\text{Q}}(\tau) &= -\frac{1}{|C|} \sum_{t=1}^{|C|} \log \frac{\exp(\hat{r}_{c_t})}{\Sigma_t}, \end{aligned} \quad (10)$$

where  $C$  and  $W$  respectively denote the index lists of correct and incorrect steps in this trajectory,  $|\cdot|$  denotes the length of this list, and  $\hat{r}_c$  and  $\hat{r}_w$  respectively denote the rewards corresponding to the correct and incorrect steps in this trajectory.  $\zeta$  is a margin hyperparameter, following the previous studies, we set it to 4. This loss aims to emphasize the ranking of correct versus incorrect trajectories. It penalizes violations in the relative order between higher-quality and lower-quality steps and focuses on the magnitude of Q-value gaps, thus improving robustness and ranking fidelity in reward modeling.

Datasets	Source	Quantity	Avg. Steps
Training Validation	Math-Shepherd (Wang et al., 2024b)	397,927 20,944	6.3
Evaluation	GSM-Plus (Li et al., 2024c)	MetaMath-Mistral-7B (Yu et al., 2024) MuggleMath-13B (Li et al., 2024b) Llama-3-70B-Instruct (Grattafiori et al., 2024)	2400 questions × 128 trajectories 3.3 3.2 2.9
	MATH500 (Hendrycks et al., 2021)	MetaMath-Mistral-7B (Yu et al., 2024) MuggleMath-13B (Li et al., 2024b) Llama-3-70B-Instruct (Grattafiori et al., 2024)	500 questions × 128 trajectories 3.9 2.6 2.9
		ProcessBench (Zheng et al., 2025)	3400 7.2

Table 1: Statistical information of the training, validation and evaluation datasets.

### 3.4 Inference and Evaluation

During inference, for a given question  $q$ , we sample  $N$  candidate trajectories  $D_q = \{\tau_1, \dots, \tau_N\}$  from the policy model. We evaluate each trajectory using BiPRM to compute the aggregate score  $R(\tau_i)$ . Following the standard Best-of- $N$  protocol, the trajectory with the highest reward is selected as the prediction:  $\tau_q^* = \arg \max_{\tau_i \in D_q} R(\tau_i)$ . The accuracy is determined by comparing the final answer extracted from  $\tau_q^*$  with the ground truth. This process rigorously tests the model’s ability to identify high-quality solutions from a candidate pool.

## 4 Experiments

### 4.1 Experiments Settings

**Datasets.** We train all models on the Math-Shepherd dataset (Wang et al., 2024b). For evaluation, we employ a two-tiered approach covering both solution-level selection and step-level error detection. A detailed summary of the dataset statistics is provided in Table 1, alongside comprehensive descriptions of data processing in Appendix A.

**Solution-level Evaluation:** We utilize two widely recognized benchmarks, GSM-Plus (Li et al., 2024c) and MATH500 (Hendrycks et al., 2021). The test set comprises 128 candidate solutions per question, sampled from three diverse policy models (MetaMath-Mistral-7B (Yu et al., 2024), MuggleMath-13B (Li et al., 2024b), Llama-3-70B-Instruct (Grattafiori et al., 2024)) as provided by Li and Li (2024).

**Step-level Evaluation:** We employ ProcessBench (Zheng et al., 2025), a diagnostic benchmark containing 3,400 test cases with expert-annotated error locations. This dataset specifically tests the model’s precision in identifying the first erroneous step in complex reasoning chains from competition-level problems.

**Implementation Details.** We conduct experiments across three LLM backbones with varying capacities (Rho-Math-1B (Lin et al., 2024), Qwen2.5-Math-1.5B (Yang et al., 2024), Deepseek-Math-7B (Shao et al., 2024)) and three objectives (BCE, MSE, Q-value Rankings Loss). BiPRM shares identical training configurations with baselines to ensure a rigorous comparison. Detailed hyperparameter are listed in Appendix A.

**Metrics.** For solution-level evaluation, We report the Best-of- $N$  (BON@ $N$ ) accuracy, which measures the percentage of questions where the top-ranked solution among  $N$  candidates is correct. For step-level evaluation on ProcessBench, we report the F1 score, assessing the model’s ability to accurately localize the first error step or correctly identify a flawless solution.

### 4.2 Main Results

**Solution-level Verification: Universal Improvements.** Table 2 presents a comprehensive comparison of verification performance across 54 distinct configurations, with detailed per-BON@ $n$  results provided in Appendix C. BiPRM demonstrates robust superiority over the unidirectional L2R baseline across all settings, regardless of model scale or training objective. Specifically, on the Rho-Math-1B, Qwen2.5-Math-1.5B, and Deepseek-Math-7B backbones, BiPRM achieves average relative improvements of 13.5%, 10.0%, and 8.3%, respectively. Notably, the performance gain is particularly substantial in challenging scenarios, with a maximum relative improvement of 31.1% achieved on the Qwen2.5-Math-1.5B backbone. These consistent gains confirm that the bidirectional evaluation paradigm provides a fundamental modeling advantage. By integrating future context, BiPRM significantly enhances the ranking precision and

Sampling Policy	Backbone	Method	Dataset: MATH500			Dataset: GSM-Plus		
			BCE	MSE	Q-value rankings	BCE	MSE	Q-value rankings
MetaMath-Mistral-7B	Rho-Math-1B	L2R	18.48	22.32	23.36	36.64	40.09	42.06
		Ours	<b>23.56</b>	<b>24.96</b>	<b>26.04</b>	<b>45.89</b>	<b>45.28</b>	<b>47.96</b>
	Qwen2.5-Math-1.5B	L2R	32.80	35.68	36.20	51.90	55.87	54.81
		Ours	<b>38.56</b>	<b>39.80</b>	<b>40.24</b>	<b>55.02</b>	<b>58.11</b>	<b>58.87</b>
	Deepseek-Math-7B	L2R	31.64	34.20	32.24	53.29	56.27	54.84
		Ours	<b>36.44</b>	<b>37.40</b>	<b>34.48</b>	<b>56.96</b>	<b>59.52</b>	<b>57.13</b>
Muggle-Math-13B	Rho-Math-1B	L2R	18.88	16.04	18.56	33.68	38.81	41.84
		Ours	<b>21.00</b>	<b>19.52</b>	<b>20.56</b>	<b>42.89</b>	<b>44.02</b>	<b>45.09</b>
	Qwen2.5-Math-1.5B	L2R	25.88	31.76	30.96	53.41	53.23	54.17
		Ours	<b>33.92</b>	<b>34.96</b>	<b>36.08</b>	<b>56.33</b>	<b>57.43</b>	<b>58.46</b>
	Deepseek-Math-7B	L2R	25.08	31.44	28.52	53.58	56.55	55.30
		Ours	<b>32.64</b>	<b>33.92</b>	<b>30.44</b>	<b>59.57</b>	<b>59.49</b>	<b>57.79</b>
Llama-3-70B-Instruct	Rho-Math-1B	L2R	32.76	34.40	34.80	64.14	66.69	67.23
		Ours	<b>37.32</b>	<b>38.24</b>	<b>39.08</b>	<b>67.73</b>	<b>68.64</b>	<b>68.92</b>
	Qwen2.5-Math-1.5B	L2R	41.44	41.76	44.20	68.76	69.82	70.17
		Ours	<b>48.48</b>	<b>46.68</b>	<b>47.80</b>	<b>70.44</b>	<b>70.93</b>	<b>71.20</b>
	Deepseek-Math-7B	L2R	40.24	42.52	40.56	68.88	70.07	70.75
		Ours	<b>43.48</b>	<b>48.12</b>	<b>43.92</b>	<b>71.07</b>	<b>71.21</b>	<b>71.55</b>

Table 2: **Solution-level** comparison of Best-of- $N$  (BON) performance, averaged from BON@8 to BON@128, across 54 configurations involving two benchmarks, three backbones, three PRM objectives and three sampling policies. BiPRM consistently outperforms the L2R baseline, achieving superior average scores of **40.37** vs. 36.15 for Rho, **51.30** vs. 47.38 for Qwen, and **50.29** vs. 47.00 for Deepseek.

stability of the verification process.

Subset	Method	BCE	MSE	Q-value rankings
GSM8K	L2R	37.8	50.7	43.0
	Ours	<b>46.8</b>	<b>50.8</b>	<b>47.0</b>
MATH	L2R	37.7	21.1	28.6
	Ours	<b>37.9</b>	<b>23.5</b>	<b>38.0</b>
OlympiadBench	L2R	13.3	8.1	9.5
	Ours	<b>32.9</b>	<b>12.1</b>	<b>18.7</b>
Omni-MATH	L2R	26.0	5.4	14.0
	Ours	<b>28.0</b>	<b>7.1</b>	<b>19.9</b>

Table 3: **Step-level** error detection performance (F1 Score) using ProcessBench on the Qwen2.5-Math-1.5B backbone. BiPRM consistently outperforms the L2R baseline, with average scores of **30.2** vs. 24.6.

**Step-level Error Detection: ProcessBench Evaluation.** Beyond selecting the correct final answer, a robust verifier must accurately pinpoint logical flaws within the reasoning process. We evaluate this fine-grained capability using ProcessBench on the Qwen2.5-Math-1.5B backbone. As detailed in Table 3, BiPRM consistently outperforms the L2R baseline in F1 scores across diverse subsets. Quantitatively, our method achieves an average relative improvement of 37.7% compared to the baseline.

While L2R models often struggle with local inconsistencies, BiPRM leverages retrospective verification to detect errors that are only evident when viewed from a global perspective. These results validate that our method effectively generalizes to out-of-distribution data and offers superior interpretability by precisely locating logical distinct breaks in reasoning chains.

#### Comparison with Strong External Baselines.

To further contextualize the capability of our bidirectional framework, we benchmark BiPRM against several recent high-performance PRM baselines. As illustrated in Table 4, despite utilizing a significantly smaller 1.5B backbone, BiPRM achieves cross-scale superiority, consistently outperforming widely recognized 7B (Wang et al., 2024b) and 8B (Dong et al., 2024) models across multiple subsets. Impressively, on the GSM8K and Omni-MATH datasets, our 1.5B BiPRM even surpasses the massive Qwen2.5-Math-RM-72B (Yang et al., 2024). Furthermore, while the performance of baselines such as Qwen2.5-Math-7B-Math-Shepherd (Zhang et al., 2025) degrades drastically as mathematical difficulty increases (dropping from 62.5 on GSM8K to just 7.7 on Omni-MATH), BiPRM maintains remarkable stability.

Model	GSM8K	MATH	OlympiadBench	Omni-MATH
Math-Shepherd-PRM-7B (Wang et al., 2024b)	47.9	29.5	24.8	23.8
RLHFlow-PRM-Deepseek-8B (Dong et al., 2024)	38.8	33.8	16.9	16.9
Qwen2.5-Math-7B-Math-Shepherd (Zhang et al., 2025)	<b>62.5</b>	31.6	13.7	7.7
Qwen2.5-Math-RM-72B (Yang et al., 2024)	43.5	<b>47.2</b>	<b>37.6</b>	<u>27.4</u>
Ours (Qwen2.5-Math-1.5B)	46.8	<u>37.9</u>	<u>32.9</u>	<b>28.0</b>

Table 4: Comparison of F1 scores on ProcessBench against strong external baselines. The best values are highlighted in bold, and the second-best values are underlined. BiPRM maintains excellent stability across difficulty levels despite operating at a smaller parameter scale.

This validates the broad applicability and rigorous structural advantage provided by the bidirectional paradigm.

### 4.3 Ablation Studies

We conduct ablation studies on the Qwen2.5-Math-1.5B backbone under BCE and MSE objectives to quantify component contributions and verify the necessity of the bidirectional architecture.

**Impact of Bidirectional Components.** As shown in Table 5, removing either the backward stream (w/o R2L) or the forward stream (w/o L2R) leads to significant performance drops compared to the full BiPRM. This confirms the inherent complementarity of the two directions. Furthermore, the dynamic gating mechanism (BiPRM) consistently outperforms the static averaging baseline (w/o Dynamic), particularly on difficult datasets like MATH500 (e.g., 48.48 vs. 44.00 on Llama-3-70B-Instruct policy), validating the effectiveness of context-aware adaptive fusion.

**Superiority over Homogeneous Ensembles.** A potential concern of our bidirectional framework is whether the performance gains stem simply from increased computation (doubled FLOPs) rather than the complementary nature of the bidirectional context. To address this, we compare BiPRM against homogeneous ensembles that utilize the same computational budget:  $2\times L2R$  (averaging scores from two independently trained L2R-PRMs) and  $2\times R2L$  (averaging scores from two R2L-PRMs). As shown in Table 5, BiPRM consistently surpasses these homogeneous ensembles. For instance, on the MuggleMath/BCE setting, BiPRM (56.33) significantly outperforms both  $2\times L2R$  (50.28) and  $2\times R2L$  (51.65). This demonstrates that the performance leap derives from the synergistic integration of distinct reasoning perspectives rather than mere computational scaling.

Sampling Policy	Method	MATH500		GSM-Plus	
		BCE	MSE	BCE	MSE
MetaMath-Mistral-7B	w/o R2L	32.80	35.68	51.90	55.87
	w/o L2R	32.32	32.80	52.21	51.83
	w/o Dynamic	38.16	37.44	53.96	58.02
	$2\times L2R$	34.56	39.44	50.12	56.68
	$2\times R2L$	33.96	36.20	52.41	53.76
	BiPRM	<b>38.56</b>	<b>39.80</b>	<b>55.02</b>	<b>58.11</b>
Muggle-Math-13B	w/o R2L	25.88	31.76	53.41	53.23
	w/o L2R	27.60	32.00	51.12	55.92
	w/o Dynamic	33.28	34.72	55.55	56.47
	$2\times L2R$	30.40	32.92	50.28	54.15
	$2\times R2L$	33.16	34.32	51.65	53.65
	BiPRM	<b>33.92</b>	<b>34.96</b>	<b>56.33</b>	<b>57.43</b>
Llama-3-70B-Instruct	w/o R2L	41.44	41.76	68.76	69.82
	w/o L2R	40.36	39.80	69.73	69.50
	w/o Dynamic	44.00	44.88	70.07	70.57
	$2\times L2R$	43.64	42.52	67.69	70.29
	$2\times R2L$	41.40	44.24	68.24	69.31
	BiPRM	<b>48.48</b>	<b>46.68</b>	<b>70.44</b>	<b>70.93</b>

Table 5: Ablation results on the Qwen2.5-Math-1.5B backbone under BCE and MSE objectives. Metrics report the average BON@ $n$  from BON@8 to BON@128. "w/o Dynamic" denotes the static average fusion. " $2\times$ " denotes the ensemble of two independently PRMs trained with different seeds.

**Stability in Long-Chain Reasoning.** To further investigate the robustness of our dynamic gating mechanism in extended context settings, we conducted targeted experiments isolating long-tail data (sequences strictly  $> 1024$  tokens). The empirical results demonstrate that dynamic gating successfully preserves robust verification capability even when processing highly extended trajectories, consistently outperforming static averaging methods. For detailed quantitative results and analysis, please refer to Appendix D.

### 4.4 Analysis

In this section, we analyze the characteristics of different reward modeling paradigms through quantitative error trends, the evolution of learned gating weights, and practical inference efficiency.

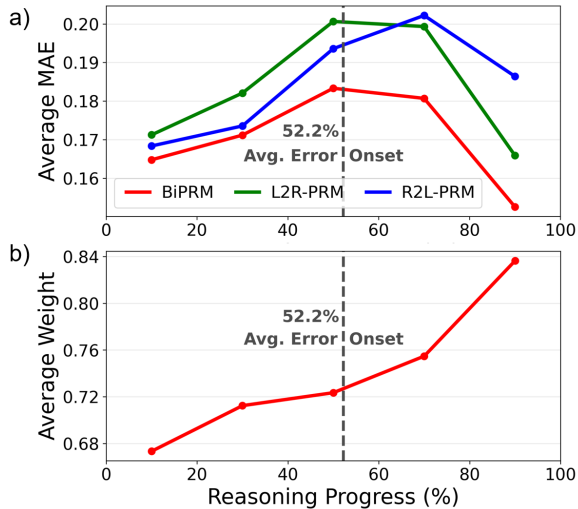


Figure 3: Analysis of (a) step-wise MAE error distribution and (b) the evolution of dynamic gating weights on the Math-Shepherd test set. The horizontal axis represents the normalized progress of reasoning steps within a solution trajectory.

### Error Distribution across Reasoning Steps.

Figure 3(a) illustrates the MAE trajectories relative to the normalized reasoning progress. The unidirectional baselines exhibit distinct position-dependent biases. The vertical dashed line marks the Average Error Onset at 52.2%, which serves as the critical boundary shifting from correct premises to subsequent logical failures. We observe that all models encounter their highest prediction uncertainty in this transition zone, resulting in peak MAE values around the center of the trajectory. Nevertheless, the unidirectional baselines exhibit divergent behaviors away from this boundary. The L2R-PRM (green) shows reduced errors towards the final steps (90%) as it benefits from accumulated forward context. Conversely, the R2L-PRM (blue) achieves lower error rates in the early stages (10% to 30%) because the reverse stream effectively treats these initial steps as terminal stages with full context visibility. Crucially, BiPRM (red) consistently maintains the lowest MAE throughout the entire process, particularly effectively suppressing the error spike in the high-difficulty transition region. This suggests that BiPRM successfully fuses these complementary strengths to eliminate localized blind spots.

**Evolution of Dynamic Gating Weights.** To further validate the adaptive nature of our fusion mechanism, we visualize the average value of the learnable gating weight  $\sigma_t$  across the reasoning

progress in Figure 3(b). Recall that  $\sigma_t$  represents the weighting coefficient for the L2R stream (i.e.,  $r_t^{\text{BiPRM}} = \sigma_t \cdot r_t^{\text{L2R}} + (1 - \sigma_t) \cdot r_t^{\text{R2L}}$ ). As observed in the figure,  $\sigma_t$  exhibits a monotonically increasing trend as reasoning progresses. In the early stages (0%-30%),  $\sigma_t$  remains relatively low, indicating that the model automatically learns to rely more heavily on the R2L stream ( $1 - \sigma_t$  is high). This aligns with our findings in the error distribution analysis that the backward view provides superior verification for initial steps. Conversely, as the reasoning approaches the final answer (70%-100%),  $\sigma_t$  increases significantly, shifting the focus towards the L2R stream to leverage the accumulated forward context. This adaptive weight allocation confirms that BiPRM effectively captures the complementary strengths of the dual streams, dynamically prioritizing the most informative context for each specific step.

Subset	Method	BCE	MSE	Q-value rankings
GSM8K	L2R	35.9 (↓1.9)	44.7 (↓6.0)	38.5 (↓4.5)
	Ours	<b>43.9</b> (↓2.9)	<b>45.6</b> (↓5.2)	<b>44.3</b> (↓2.7)
MATH	L2R	34.9 (↓2.8)	19.8 (↓1.3)	27.4 (↓1.2)
	Ours	<b>36.6</b> (↓1.3)	<b>21.7</b> (↓1.8)	<b>37.2</b> (↓0.8)
Olympiad Bench	L2R	13.1 (↓0.2)	6.4 (↓1.7)	9.2 (↓0.3)
	Ours	<b>31.1</b> (↓1.8)	<b>10.5</b> (↓1.6)	<b>18.7</b> (↓0.0)
Omni-MATH	L2R	24.5 (↓1.5)	4.6 (↓0.8)	14.0 (↓0.0)
	Ours	<b>27.6</b> (↓0.4)	<b>6.8</b> (↓0.3)	<b>19.7</b> (↓0.2)

Table 6: Sensitivity analysis under step boundary perturbations on the Qwen2.5-Math-1.5B backbone. Values represent the Perturbed F1 Score and the absolute degradation (↓) relative to the original clean results in Table 3. BiPRM exhibits notably smaller performance drops.

**Robustness to Ambiguous Step Boundaries.** To address potential practical concerns regarding noisy and ambiguous step boundaries, we conducted a targeted sensitivity analysis by artificially introducing boundary perturbations (merging, splitting, and noise insertion) into the ProcessBench dataset. As shown in Table 6, while all models naturally exhibit performance degradation under these adverse conditions, BiPRM demonstrates significantly superior stability. On average, BiPRM experiences an F1 drop of only 5.24%, compared to a 7.52% drop for the L2R baseline. This confirms that the bidirectional architecture fundamentally enhances structural robustness, making it highly reliable in real-world scenarios containing inconsistent step segmentations.

<b>Question:</b> Given a sequence $\{a_n\}$ defined by $a_1 = 1$ and $a_{n+1} = 2a_n + 1$ , find the general term formula.	Label	Bi-PRM	L2R-PRM	R2L-PRM
<b>Step 1:</b> Compute the first few terms: $a_2 = 2a_1 + 1 = 3$ , $a_3 = 2a_2 + 1 = 7$ , $a_4 = 2a_3 + 1 = 15, \dots$	1	1.00	1.00	1.00
<b>Step 2:</b> Observing that $a_n + 1$ yields powers of 2, we conjecture: $a_n = 2n - 1$ .	0	0.31	1.00	0.29
<b>Step 3:</b> Base case ( $n = 1$ ): $a_1 = 1 = 2 \times 1 - 1$ .	0	0.38	1.00	0.38
<b>Step 4:</b> Assume $a_k = 2k - 1$ holds for $n = k$ , then for $n = k + 1$ , $a_{k+1} = 2a_k + 1 = 2(k + 1) - 1$ . Hence, it holds for $n = k + 1$ .	0	0.01	1.00	0.00
<b>Step 5:</b> By the principle of mathematical induction, the formula is valid for all natural numbers $n$ . The general term formula is $a_n = 2n - 1$ .	0	0.17	0.99	0.08

Table 7: A representative case study comparing step-wise reward scores of BiPRM, L2R-PRM, and R2L-PRM. The critical error stems from an incorrect conjecture in **Step 2** (highlighted in red), which the model attempts to justify through a forced derivation in **Step 4**. While L2R-PRM is deceived by the superficial coherence, R2L-PRM successfully identifies the logical fallacy through retrospective verification.

**Inference Latency.** A primary concern regarding bidirectional architectures is the potential doubling of inference time. However, since the L2R and R2L evaluation streams operate independently prior to the final gating fusion, they can be processed concurrently. By stacking the forward and reverse inputs into a single batch, we leverage the parallel computing capabilities of the GPU to minimize latency. As shown in Table 8, empirical results demonstrate that the average wall-clock time for scoring a single solution increases only marginally from 27.982 ms for the L2R baseline to 29.393 ms for BiPRM. This corresponds to a relative latency overhead of approximately 5%, demonstrating that our parallel implementation effectively mitigates the temporal cost of dual-stream evaluation, making BiPRM highly practical for real-world deployment. Further discussions on computational efficiency, including detailed comparisons of model parameters, measured TFLOPS, and peak memory usage, are provided in Appendix B.

Model	Params	TFLOPS	Wall-clock Time <sup>†</sup>
L2R-PRM	1.543B	10.034	27.982 ms
BiPRM (Ours)	1.548B	19.106	29.393 ms

Table 8: Quantitative comparison of model parameters, measured computational cost (TFLOPS), and inference time latency between L2R-PRM and BiPRM on the Qwen2.5-Math-1.5B backbone. †: Measured as the average inference time per single solution.

**Case Study.** Table 7 presents a representative example where the model constructs a superficially coherent proof based on an incorrect hypothesis

in Step 2. The L2R baseline fails to detect this error because it views the hypothesis in Step 2 as a plausible conjecture pending verification, assigning it a high score based on local coherence. In contrast, the R2L stream correctly penalizes the final conclusion in Step 5 without needing to trace the intermediate derivation. By evaluating the final answer directly against the problem statement, the R2L stream identifies the fundamental mathematical contradiction between the derived formula and the sequence definition, thereby effectively filtering out the erroneous trajectory. For a more comprehensive qualitative analysis on success and failure cases, please refer to Appendix E.

## 5 Conclusion

In this paper, we proposed BiPRM, a novel bidirectional evaluation paradigm designed to address the limitations of unidirectional PRMs in accessing global context. By synergizing a parallel reverse evaluation stream with a dynamic gating mechanism, BiPRM enables effective retrospective verification while incurring negligible computational overhead. Extensive experiments demonstrate that our method consistently outperforms L2R baselines in both solution-level ranking and step-level error localization. These findings underscore the necessity of bidirectional context for robust reasoning supervision and offer a promising direction for future research in process reward modeling.

## Limitations

While the promising results achieved by BiPRM, certain limitations remain.

**Computational Cost.** The bidirectional evaluation mechanism inherently necessitates processing the input sequence twice, leading to a two-fold increase in theoretical floating-point operations (FLOPs) compared to unidirectional baselines. This increased energy consumption is an unavoidable trade-off for the enhanced verification capability. However, from the perspective of practical deployment, we effectively mitigate the impact on user experience through a parallel execution strategy. As demonstrated in our efficiency analysis (Section 4.4), this implementation ensures that the actual inference time latency increases by merely 5%, maintaining the method’s feasibility for real-time applications.

**Generalization across Domains.** Our experimental validation is currently confined to the domain of mathematical reasoning. While mathematics serves as a rigorous testbed for evaluating logical consistency and stepwise correctness, the generalizability of the bidirectional verification paradigm to other complex tasks remains to be verified. We leave the exploration of broader application scenarios, such as code generation, open-ended commonsense reasoning, and symbolic logic tasks, to future research. We hope that our findings will serve as a foundation for developing more robust and generalized process supervision frameworks in these diverse fields.

## Acknowledgments

This work was funded by the National Natural Science Foundation of China (No. 62106165), the Jiangsu Major Science and Technology Special Fund for Innovative Biologics (No. BG2025062), and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

## References

Xin Bi, Haojie Nie, Guoliang Zhang, Lei Hu, Yuliang Ma, Xiangguo Zhao, Ye Yuan, and Guoren Wang. 2023. Boosting question answering over knowledge graph with reward integration and policy evaluation under weak supervision. *Information Processing & Management*, 60(2):103242.

Thomas Carta, Pierre-Yves Oudeyer, Olivier Sigaud, and Sylvain Lamprier. 2022. Eager: Asking and answering questions for automatic reward shaping in language-guided rl. *Advances in neural information processing systems*, 35:12478–12490.

Wenxiang Chen, Wei He, Zhiheng Xi, Honglin Guo, Boyang Hong, Jiazheng Zhang, Nijun Li, Tao Gui, Yun Li, Qi Zhang, and 1 others. 2025. Better process supervision with bi-directional rewarding signals. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14471–14485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025a. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529.

Jun Gao, Qi Lv, Zili Wang, Tianxiang Wu, Ziqiang Cao, and Wenjie Li. 2025b. Uniicl: An efficient icl framework unifying compression, selection, and generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 500–510.

Jun Gao, Qian Qiao, Tianxiang Wu, Zili Wang, Ziqiang Cao, and Wenjie Li. 2025c. Aim: Let any multimodal large language models embrace efficient in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3077–3085.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, and 1 others. 2024. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *CoRR*.
- Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. 2022. Coder1: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328.
- Bolun Li, Zhihong Sun, Tao Huang, Hongyu Zhang, Yao Wan, Ge Li, Zhi Jin, and Chen Lyu. 2024a. Ir-coco: Immediate rewards-guided deep reinforcement learning for code completion. *Proceedings of the ACM on Software Engineering*, 1(FSE):182–203.
- Chengpeng Li, Zheng Yuan, Hongyi Yuan, Guanting Dong, Keming Lu, Jiancan Wu, Chuanqi Tan, Xiang Wang, and Chang Zhou. 2024b. [Query and response augmentation cannot help out-of-domain math reasoning generalization](#).
- Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024c. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. *arXiv preprint arXiv:2402.19255*.
- Wendi Li and Yixuan Li. 2024. Process reward model with q-value rankings. *arXiv preprint arXiv:2410.11287*.
- Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and 1 others. 2024. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.
- Shuqi Liu, Bowei He, and Linqi Song. 2024. Bi-chainer: Automated large language models reasoning with bidirectional chaining. *arXiv preprint arXiv:2406.06586*.
- Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yunlong Feng, and Zhijiang Guo. 2024. Autopsv: Automated process-supervised verifier. *Advances in Neural Information Processing Systems*, 37:79935–79962.
- Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, and 1 others. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.
- Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. 2025. What are step-level reward models rewarding? counterintuitive findings from mcts-boosted mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24812–24820.
- Vassilis Papadopoulos, Jérémie Wenger, and Clément Hongler. 2024. Arrows of time for large language models. *CoRR*.
- Jacob Pfau, Alex Infanger, Abhay Sheshadri, Ayush Panda, Julian Michael, and Curtis Huebner. 2023. Eliciting language model behaviors using reverse language models. In *Socially Responsible Language Modelling Research*.
- Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*.
- Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.
- Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2015. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*.
- Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024a. Q\*: Improving multi-step reasoning for llms with deliberative planning. *arXiv preprint arXiv:2406.14283*.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.
- Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. 2024c. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. *arXiv preprint arXiv:2402.02658*.

- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Meta-math: Bootstrap your own mathematical questions for large language models. In *ICLR*.
- Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Rest-mcts\*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772.
- Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Joshua Susskind, and Navdeep Jaitly. 2023. Planner: Generating diversified paragraph via latent language diffusion model. *Advances in Neural Information Processing Systems*, 36:80178–80190.
- Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10495–10516.
- Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Processbench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1024.

## A Implementation Details

**Data Preparation.** The training corpus is derived from the Math-Shepherd dataset, which comprises mathematical questions paired with multi-step reasoning solutions and step-level supervision signals. To ensure data quality, we filter out instances containing only a single reasoning step and partition the remaining data into training and validation subsets using a 95:5 ratio. For evaluation, the test set includes 128 candidate solutions generated by each policy model for every question in the benchmarks, resulting in six distinct evaluation sets.

**Training Configuration.** We independently train both the unidirectional PRMs and our proposed BiPRMs across all nine combinations formed by the LLM backbones and PRM objectives. Crucially, each pair of PRM and BiPRM models shares identical training configurations to guarantee a fair comparison. All models are trained using 8 NVIDIA RTX A5000 GPUs. Regarding the software environment, we use the following package versions: torch==2.3.1+cu118, trl==0.8.0, transformers==4.43.0, accelerate==0.33.0, deepspeed==0.13.1, nvidia-nccl-cu12==2.20.5. We employ the ZeRO-3 optimization stage of DeepSpeed with bfloat16 precision. Gradient checkpointing is set to true for Deepseek-Math-7B, while it is disabled for both Qwen2.5-Math-1.5B and Rho-Math-1B. The specific hyperparameters for all experiments are detailed in Table 9.

Hyperparameters	Value
epoch	1
learning rate	3e-5
optimizer	AdamW
scheduler	linear
seed	1106
batch size per GPU	2
gradient accumulation steps	4

Table 9: Hyperparameters in all experimental configurations.

## B Computational Efficiency Analysis

In this section, we provide a supplementary analysis of the computational efficiency between BiPRM and the standard unidirectional L2R-PRM. All measurements were conducted on the Qwen2.5-Math-1.5B backbone using a single NVIDIA RTX A5000

GPU to ensure a controlled experimental environment. Our evaluation focuses on four key dimensions: model parameter size, measured computational cost (TFLOPS), peak memory usage, and the actual inference latency (wall-clock time).

As shown in Table 8, the parameter overhead introduced by our dynamic gating module is a negligible 0.3%, increasing the total count from 1.543B to 1.548B. In terms of measured computational cost, BiPRM inherently requires processing the input sequence twice, resulting in an empirical TFLOPS increase from 10.034 to 19.106. Crucially, despite the doubling of empirical compute, BiPRM achieves a highly favorable and practical trade-off for real-world deployment. Parallel execution of the two streams causes only a marginal relative increase in overall peak memory usage, rising from 6.43 GB to 6.65 GB. Consequently, the average wall-clock time for scoring a single solution increases by just 5%, from 27.982 ms for the L2R baseline to 29.393 ms for BiPRM. This demonstrates that our implementation effectively exchanges an acceptable, low-impact overhead in computation and memory for substantially enhanced verification capabilities with minimal impact on user-perceived latency.

## C Complete experimental results

We present the complete experimental results of Rho-Math-1B, Qwen2.5-Math-1.5B and Deepseek-Math-7B in Tables 11, 12 and 13. Including specific values of bon@8 to bon@128 across the two benchmarks and three sampling policies.

## D Stability of Dynamic Gating in Long-Chain Scenarios

To thoroughly analyze the stability of our dynamic gating mechanism in extremely long-chain reasoning scenarios, we conducted additional evaluations focusing on long-tail data with sequence lengths strictly greater than 1024 tokens. This evaluation was performed using the Qwen2.5-Math-1.5B backbone under the BCE objective.

As detailed in Table 10, the average sequence lengths in these long-tail subsets significantly increase compared to the original datasets (e.g., from 988 to 1457 on MATH500). While all models inevitably experience a certain degree of performance degradation when handling such exceptionally long reasoning chains, the "Dynamic Gating" strategy consistently demonstrates superior resilience com-

pared to the "Static Averaging" baseline. Overall, static averaging exhibits an average performance drop of 37.56% relative to its standard results, whereas our dynamic gating mechanism limits this drop to 31.40%. This empirical evidence validates that dynamic context-aware fusion is crucial for maintaining stable verification performance on extended contexts.

Sampling Policy	Method	MATH500 [988 → 1457]	GSM-Plus [779 → 1201]
MetaMath-Mistral-7B	Static Avg.	38.16 → 16.52	53.96 → 40.22
	Dynamic	<b>38.56</b> → <b>19.79</b>	<b>55.02</b> → <b>43.18</b>
Muggle-Math-13B	Static Avg.	33.28 → 9.41	55.55 → 41.83
	Dynamic	<b>33.92</b> → <b>11.81</b>	<b>56.33</b> → <b>45.26</b>
Llama-3-70B-Instruct	Static Avg.	44.00 → 23.02	70.07 → 53.22
	Dynamic	<b>48.48</b> → <b>29.11</b>	<b>70.44</b> → <b>58.54</b>

Table 10: Performance (Average BON@ $n$ ) comparison between Static Averaging and Dynamic Gating on original vs. long-tail data (sequences > 1024 tokens). Bracketed values indicate the average sequence lengths of the original versus the long-tail subset.

## E Additional Qualitative Analysis

In this section, we provide a deeper analysis of BiPRM’s performance across different error types. We verify its robustness in detecting calculation errors and discuss specific scenarios where the R2L stream may face challenges.

### Effectiveness in Calculation Error Detection.

Table 14 presents a case involving a quadratic function. In this instance, a critical calculation error occurs at Step 3 regarding the vertex coordinate. Consistent with our error distribution analysis, the R2L-PRM precisely identifies this early-stage mistake with a score of 0.05, whereas the L2R-PRM fails to penalize it immediately, assigning an ambiguous score of 0.46. BiPRM effectively integrates these divergent signals to provide a precise and robust evaluation, demonstrating that the bidirectional mechanism is beneficial not only for logical reasoning but also for rigorous numerical verification.

**Limitations and Failure Analysis.** While BiPRM shows significant improvements in most scenarios, it exhibits certain limitations in specific arithmetic inconsistencies. Table 15 illustrates a failure case involving a circle geometry problem. The error arises in Step 4, where the model omits

the denominator "2" during algebraic simplification ( $\pi r^2 / (2\pi r) = 20 \rightarrow r^2 / r = 20$ ). Although Step 4 is mathematically incorrect given Step 3, the transition from Step 4 to Step 5 ( $r = 20$ ) is internally consistent. Since the R2L mechanism primarily evaluates whether the current step can logically lead to the subsequent steps (retrospective consistency), it assigns a relatively high score (0.63) to Step 4 because the derivation from Step 4 to the end is smooth. This suggests that R2L is more sensitive to global logical coherence than to local isolated arithmetic skips. However, note that the L2R stream successfully identifies this error (0.01), and BiPRM’s gating mechanism partially mitigates the R2L failure, resulting in a final score of 0.18. This case highlights the necessity of fusing both directions to handle disjointed arithmetic errors effectively.

## F Sensitivity Analysis of Aggregation Operators

We investigate the sensitivity of BiPRM to four trajectory-level aggregation operators: product ( $\prod$ ), minimum (min), maximum (max), and arithmetic mean (mean). Table 16 presents the detailed breakdown of BON@ $N$  scores across all experimental configurations. Quantitative analysis reveals that the aggregation strategy significantly influences performance. The poor performance of max aligns with intuition, as a single high-quality step cannot compensate for logical errors elsewhere in the trajectory. Conversely, the success of both mean and min suggests that considering the overall quality or the worst-case step are both effective strategies. Despite the slight empirical edge of mean, we adopt min as the default strategy. This decision aligns with the "weakest link" principle of mathematical reasoning where a single error invalidates the entire proof, thereby ensuring theoretical consistency with the logical nature of the task and prior literature (Wang et al., 2024b; Li and Li, 2024).

Sampling Policy	PRM Objective	Method	Dataset: MATH500					Dataset: GSM-Plus				
			@8	@16	@32	@64	@128	@8	@16	@32	@64	@128
MetaMath-Mistral-7B	BCE	L2R	21.20	20.80	17.40	15.80	17.20	43.62	40.21	36.79	33.04	29.54
		Ours	<b>24.20</b>	<b>24.00</b>	<b>24.60</b>	<b>22.60</b>	<b>22.40</b>	<b>47.46</b>	<b>46.71</b>	<b>45.79</b>	<b>44.92</b>	<b>44.58</b>
	MSE	L2R	24.80	23.60	22.80	21.60	18.80	47.96	44.50	41.54	36.67	29.79
Ours		<b>27.00</b>	<b>25.80</b>	<b>25.20</b>	<b>23.40</b>	<b>23.40</b>	<b>50.08</b>	<b>47.50</b>	<b>45.46</b>	<b>43.12</b>	<b>40.25</b>	
Q-value rankings	L2R	25.60	24.20	23.80	22.60	20.60	49.71	46.54	42.88	38.25	32.92	
	Ours	<b>26.80</b>	<b>26.40</b>	<b>26.40</b>	<b>26.80</b>	<b>23.80</b>	<b>51.54</b>	<b>49.79</b>	<b>47.88</b>	<b>46.00</b>	<b>44.58</b>	
Muggle-Math-13B	BCE	L2R	20.00	21.00	19.00	18.60	15.80	41.38	36.67	34.12	30.21	26.04
		Ours	<b>23.80</b>	<b>21.80</b>	<b>19.20</b>	<b>20.60</b>	<b>19.60</b>	<b>43.83</b>	<b>43.58</b>	<b>43.12</b>	<b>42.00</b>	<b>41.92</b>
	MSE	L2R	18.40	17.60	16.20	14.60	13.40	45.50	43.46	39.29	34.25	31.54
Ours		<b>21.60</b>	<b>21.20</b>	<b>19.20</b>	<b>18.80</b>	<b>16.80</b>	<b>47.79</b>	<b>47.00</b>	<b>44.71</b>	<b>41.88</b>	<b>38.71</b>	
Q-value rankings	L2R	20.20	17.80	18.40	19.20	17.20	46.92	45.08	42.46	38.71	36.04	
	Ours	<b>21.00</b>	<b>21.60</b>	<b>21.00</b>	<b>20.80</b>	<b>18.40</b>	<b>47.88</b>	<b>46.88</b>	<b>45.08</b>	<b>43.54</b>	<b>42.08</b>	
Llama-3-70B-Instruct	BCE	L2R	<b>38.60</b>	36.20	33.40	28.60	27.00	66.67	65.92	64.42	62.88	60.83
		Ours	37.20	<b>37.80</b>	<b>37.40</b>	<b>37.40</b>	<b>36.80</b>	<b>67.88</b>	<b>68.21</b>	<b>67.83</b>	<b>67.67</b>	<b>67.08</b>
	MSE	L2R	38.20	36.00	35.60	32.80	29.40	69.08	68.21	67.54	66.04	62.58
Ours		<b>41.00</b>	<b>37.80</b>	<b>37.80</b>	<b>37.60</b>	<b>37.00</b>	<b>70.54</b>	<b>69.96</b>	<b>68.88</b>	<b>67.46</b>	<b>66.38</b>	
Q-value rankings	L2R	39.00	35.40	34.60	33.20	31.80	69.83	68.46	67.17	66.29	64.38	
	Ours	<b>41.00</b>	<b>39.60</b>	<b>39.20</b>	<b>38.20</b>	<b>37.40</b>	<b>71.46</b>	<b>69.54</b>	<b>68.96</b>	<b>68.25</b>	<b>66.38</b>	

Table 11: **Rho-Math-1B** results measured by Best-of-N (BON@ $n$ ) accuracy across two benchmarks, three PRM objectives and three sampling policies.

Sampling Policy	PRM Objective	Method	Dataset: MATH500					Dataset: GSM-Plus				
			@8	@16	@32	@64	@128	@8	@16	@32	@64	@128
MetaMath-Mistral-7B	BCE	L2R	31.60	32.80	33.40	34.00	32.20	53.92	53.96	52.62	51.21	47.79
		Ours	<b>34.20</b>	<b>38.60</b>	<b>38.80</b>	<b>40.60</b>	<b>40.60</b>	<b>56.46</b>	<b>56.42</b>	<b>55.75</b>	<b>54.46</b>	<b>52.00</b>
	MSE	L2R	32.40	36.20	35.80	36.40	37.60	57.33	56.50	55.96	55.75	53.83
Ours		<b>36.20</b>	<b>38.20</b>	<b>40.20</b>	<b>42.20</b>	<b>42.20</b>	<b>58.33</b>	<b>57.71</b>	<b>58.25</b>	<b>58.58</b>	<b>57.67</b>	
Q-value rankings	L2R	32.40	35.40	36.80	37.80	38.60	54.75	54.88	54.88	54.58	54.96	
	Ours	<b>36.00</b>	<b>39.00</b>	<b>42.00</b>	<b>42.60</b>	<b>41.60</b>	<b>57.92</b>	<b>58.33</b>	<b>59.38</b>	<b>59.71</b>	<b>59.00</b>	
Muggle-Math-13B	BCE	L2R	26.20	26.00	26.80	26.20	24.20	52.79	54.12	53.46	53.83	52.83
		Ours	<b>30.60</b>	<b>34.00</b>	<b>35.60</b>	<b>36.20</b>	<b>33.20</b>	<b>56.04</b>	<b>57.54</b>	<b>56.62</b>	<b>55.92</b>	<b>55.54</b>
	MSE	L2R	28.60	32.00	33.00	32.80	32.40	54.08	54.12	53.29	52.92	51.75
Ours		<b>32.60</b>	<b>34.00</b>	<b>34.80</b>	<b>36.60</b>	<b>36.80</b>	<b>56.29</b>	<b>57.96</b>	<b>57.79</b>	<b>58.08</b>	<b>57.04</b>	
Q-value rankings	L2R	27.20	32.00	31.80	32.40	31.40	53.42	55.04	54.04	54.00	54.33	
	Ours	<b>33.40</b>	<b>35.20</b>	<b>36.60</b>	<b>37.40</b>	<b>37.80</b>	<b>57.75</b>	<b>58.75</b>	<b>59.04</b>	<b>58.50</b>	<b>58.25</b>	
Llama-3-70B-Instruct	BCE	L2R	42.40	42.80	41.20	40.80	40.00	70.04	69.71	70.00	67.58	66.46
		Ours	<b>47.40</b>	<b>47.60</b>	<b>46.40</b>	<b>50.80</b>	<b>50.20</b>	<b>72.42</b>	<b>71.46</b>	<b>70.83</b>	<b>69.33</b>	<b>68.17</b>
	MSE	L2R	44.60	43.40	42.00	40.40	38.40	70.96	70.92	70.29	68.79	68.12
Ours		<b>45.20</b>	<b>46.80</b>	<b>45.80</b>	<b>48.60</b>	<b>47.00</b>	<b>72.33</b>	<b>72.17</b>	<b>70.88</b>	<b>70.50</b>	<b>68.79</b>	
Q-value rankings	L2R	44.20	44.40	45.40	44.00	43.00	71.21	71.12	70.21	69.29	69.00	
	Ours	<b>46.80</b>	<b>47.40</b>	<b>47.20</b>	<b>50.20</b>	<b>47.40</b>	<b>71.58</b>	<b>71.46</b>	<b>71.79</b>	<b>70.62</b>	<b>70.54</b>	

Table 12: **Qwen2.5-Math-1.5B** results measured by Best-of-N (BON@ $n$ ) accuracy across two benchmarks, three PRM objectives and three sampling policies.

Sampling Policy	PRM Objective	Method	Dataset: MATH500					Dataset: GSM-Plus				
			@8	@16	@32	@64	@128	@8	@16	@32	@64	@128
MetaMath-Mistral-7B	BCE	L2R	29.80	33.00	31.60	32.20	31.60	55.50	54.88	54.38	52.21	49.50
		Ours	<b>35.00</b>	<b>36.20</b>	<b>36.20</b>	<b>37.00</b>	<b>37.80</b>	<b>57.08</b>	<b>58.00</b>	<b>57.42</b>	<b>56.50</b>	<b>55.79</b>
	MSE	L2R	32.20	34.80	35.40	35.00	33.60	57.75	57.04	56.67	55.17	54.71
Ours		<b>33.60</b>	<b>36.60</b>	<b>39.20</b>	<b>40.00</b>	<b>37.60</b>	<b>58.88</b>	<b>60.29</b>	<b>59.58</b>	<b>59.46</b>	<b>59.38</b>	
Q-value rankings	L2R	31.60	32.40	32.40	33.00	31.80	55.29	54.92	54.88	55.00	54.12	
	Ours	<b>33.60</b>	<b>35.00</b>	<b>34.60</b>	<b>36.00</b>	<b>33.20</b>	<b>57.08</b>	<b>57.54</b>	<b>57.71</b>	<b>57.04</b>	<b>56.29</b>	
Muggle-Math-13B	BCE	L2R	25.40	24.20	25.00	24.80	26.00	54.17	55.08	53.46	53.21	51.96
		Ours	<b>27.60</b>	<b>32.00</b>	<b>33.80</b>	<b>35.60</b>	<b>34.20</b>	<b>58.33</b>	<b>59.33</b>	<b>60.58</b>	<b>59.71</b>	<b>59.88</b>
	MSE	L2R	29.20	31.80	32.40	32.00	31.80	56.83	57.83	56.58	56.46	55.04
Ours		<b>29.40</b>	<b>34.80</b>	<b>34.20</b>	<b>35.80</b>	<b>35.40</b>	<b>58.25</b>	<b>59.92</b>	<b>59.92</b>	<b>59.38</b>	<b>60.00</b>	
Q-value rankings	L2R	27.20	27.80	27.80	29.60	30.20	54.83	56.83	55.67	54.42	54.75	
	Ours	<b>27.80</b>	<b>30.60</b>	<b>32.20</b>	<b>31.20</b>	<b>30.40</b>	<b>56.71</b>	<b>58.50</b>	<b>58.21</b>	<b>58.04</b>	<b>57.50</b>	
Llama-3-70B-Instruct	BCE	L2R	39.80	39.60	39.20	41.40	41.20	70.96	70.25	69.38	68.08	65.75
		Ours	<b>44.60</b>	<b>43.20</b>	<b>43.00</b>	<b>43.60</b>	<b>43.00</b>	<b>72.04</b>	<b>71.25</b>	<b>71.25</b>	<b>70.79</b>	<b>70.04</b>
	MSE	L2R	42.00	43.60	42.60	42.40	42.00	70.96	71.08	70.75	69.50	68.04
Ours		<b>47.80</b>	<b>48.00</b>	<b>46.60</b>	<b>49.20</b>	<b>49.00</b>	<b>71.58</b>	<b>71.54</b>	<b>71.50</b>	<b>71.21</b>	<b>70.21</b>	
Q-value rankings	L2R	40.80	41.40	38.00	40.20	42.40	71.83	70.92	70.67	70.17	70.17	
	Ours	<b>44.80</b>	<b>44.00</b>	<b>43.00</b>	<b>42.40</b>	<b>45.40</b>	<b>72.38</b>	<b>71.92</b>	<b>72.04</b>	<b>71.12</b>	<b>70.29</b>	

Table 13: **Deepseek-Math-7B** results measured by Best-of-N (BON@ $n$ ) accuracy across two benchmarks, three PRM objectives and three sampling policies.

Question: Determine the maximum value of $c$ where $-2$ is within the range of the function $f(x) = x^2 + 3x + c$ .	Label	Bi-PRM	L2R-PRM	R2L-PRM
<b>Step 1:</b> The range of a quadratic function $f(x) = ax^2 + bx + c$ is determined by the vertex of the parabola.	True	1.00	1.00	1.00
<b>Step 2:</b> The $x$ -coordinate of the vertex is $-\frac{b}{2a}$ , and the $y$ -coordinate is $c - \frac{b^2}{4a}$ .	True	1.00	0.91	1.00
<b>Step 3:</b> For the given function $f(x) = x^2 + 3x + c$ , the vertex is $(-\frac{3}{2}, c - \frac{3^2}{4}) = (-1.5, c - 4.5)$ .	False	0.05	0.46	0.05
<b>Step 4:</b> Since $-2$ is within the range, we have $-2 \leq c - 4.5$ , which implies $c \geq 2.5$ .	False	0.04	0.01	0.09
<b>Step 5:</b> Therefore, the maximum value of $c$ is $\boxed{2.5}$ .	False	0.00	0.00	0.03

Table 14: A success case study involving a calculation error. The critical error occurs in **Step 3** (highlighted in red), where the  $y$ -coordinate of the vertex is miscalculated. BiPRM effectively utilizes the R2L signal to correct the ambiguous evaluation of the L2R baseline.

Question: If the area of a circle is $M\text{cm}^2$ and the circumference is $N\text{cm}$ , with $M/N = 20$ , what is the radius of the circle in cm?	Label	Bi-PRM	L2R-PRM	R2L-PRM
<b>Step 1:</b> The area of a circle is given by the formula $A = \pi r^2$ , and the circumference is given by the formula $C = 2\pi r$ .	True	1.00	1.00	1.00
<b>Step 2:</b> We are given that $M = A$ and $N = C$ , and $M/N = 20$ .	True	1.00	1.00	1.00
<b>Step 3:</b> Substituting the formulas, we have $\pi r^2 / (2\pi r) = 20$ .	False	1.00	1.00	1.00
<b>Step 4:</b> Simplifying, we get $r^2 / r = 20$ .	False	0.18	0.01	0.63
<b>Step 5:</b> Dividing both sides by $r$ , we get $r = 20$ . The answer is: $\boxed{20}$ .	False	0.00	0.00	0.01

Table 15: A failure case study demonstrating the limitation of R2L-PRM. The critical error arises during the algebraic simplification transition between Step 3 and Step 4. Since the derivation from Step 4 onwards maintains internal consistency, the retrospective view of the R2L stream fails to identify the preceding disconnect.

Sampling Policy	Backbone	Aggregation Operator	Dataset: MATH500			Dataset: GSM-Plus			Avg
			BCE	MSE	Q-value rankings	BCE	MSE	Q-value rankings	
MetaMath-Mistral-7B	Rho-Math-1B	$\prod$	24.52	24.28	25.08	45.87	45.71	47.50	35.49
		min	23.56	24.96	26.04	45.89	45.28	47.96	35.62
		max	22.96	24.64	26.60	46.76	49.61	49.32	<b>36.65</b>
		mean	24.64	25.00	26.68	45.44	46.75	48.61	<u>36.19</u>
	Qwen2.5-Math-1.5B	$\prod$	38.76	39.64	40.00	54.88	57.20	58.14	48.10
		min	38.56	39.80	40.24	55.02	58.11	58.87	48.43
		max	30.64	36.24	41.44	51.02	55.26	58.73	45.55
		mean	38.64	39.52	40.56	54.99	58.29	58.69	<b>48.45</b>
	Deepseek-Math-7B	$\prod$	36.20	37.36	32.72	56.59	58.99	56.69	46.43
		min	36.44	37.40	34.48	56.96	59.52	57.13	<u>46.99</u>
		max	32.80	34.44	33.36	53.13	57.84	56.97	44.76
		mean	37.00	37.76	34.28	57.13	59.76	57.43	<b>47.23</b>
Muggle-Math-13B	Rho-Math-1B	$\prod$	21.36	20.28	21.44	41.30	44.46	45.55	<u>32.40</u>
		min	21.00	19.52	20.56	42.89	44.02	45.09	32.18
		max	18.52	18.68	19.64	41.38	45.92	46.28	31.74
		mean	21.28	20.40	20.92	41.33	45.12	45.53	<b>32.43</b>
	Qwen2.5-Math-1.5B	$\prod$	33.96	34.88	35.60	56.17	57.13	57.04	45.80
		min	33.92	34.96	36.08	56.33	57.43	58.46	<b>46.20</b>
		max	25.00	32.92	37.16	49.81	54.17	57.08	42.69
		mean	33.68	35.24	36.80	55.92	57.31	57.67	<u>46.10</u>
	Deepseek-Math-7B	$\prod$	32.84	33.64	30.08	59.66	58.53	56.62	45.23
		min	32.64	33.92	30.44	59.57	59.49	57.79	<b>45.64</b>
		max	25.16	30.92	31.16	52.97	56.78	56.71	42.28
		mean	32.60	33.36	30.68	59.12	59.63	57.25	<u>45.44</u>
Llama-3-70B-Instruct	Rho-Math-1B	$\prod$	35.84	38.76	38.80	66.85	68.03	67.86	52.69
		min	37.32	38.24	39.08	67.73	68.64	68.92	<u>53.32</u>
		max	36.44	37.00	37.72	67.53	69.27	69.73	52.95
		mean	35.48	38.96	38.92	68.19	69.48	69.40	<b>53.41</b>
	Qwen2.5-Math-1.5B	$\prod$	48.44	46.52	48.60	70.47	70.61	70.66	59.22
		min	48.48	46.68	48.92	70.44	70.93	71.20	<u>59.44</u>
		max	45.40	46.16	48.48	70.55	70.62	71.39	58.77
		mean	48.28	47.72	48.92	70.73	71.20	71.15	<b>59.67</b>
	Deepseek-Math-7B	$\prod$	42.92	46.96	43.76	70.48	71.18	70.89	57.70
		min	43.48	48.12	43.92	71.07	71.21	71.55	<b>58.23</b>
		max	40.72	47.28	44.00	70.32	71.26	71.62	57.53
		mean	41.88	48.48	44.44	70.88	71.27	71.95	<u>58.15</u>

Table 16: Sensitivity analysis of different trajectory-level aggregation operators across diverse backbones and objectives. The metrics reported are the average BON@N scores. Across all configurations, the overall average scores are: mean (47.45) > min (47.34) >  $\prod$  (47.01) > max (45.88).