



The Lessons of Developing Process Reward Models in Mathematical Reasoning

Zhenru Zhang Chujie Zheng Yangzhen Wu Beichen Zhang Runji Lin
Bowen Yu* Dayiheng Liu* Jingren Zhou Junyang Lin*

Qwen Team, Alibaba Group

PRM
not used for
RI

 <https://hf.co/Qwen/Qwen2.5-Math-PRM-7B>
 <https://hf.co/Qwen/Qwen2.5-Math-PRM-72B>

Abstract

Process Reward Models (PRMs) emerge as a promising approach for process supervision in mathematical reasoning of Large Language Models (LLMs), which aim to identify and mitigate intermediate errors in the reasoning processes. However, the development of effective PRMs faces significant challenges, particularly in data annotation and evaluation methodologies. In this paper, through extensive experiments, we demonstrate that commonly used Monte Carlo (MC) estimation-based data synthesis for PRMs typically yields inferior performance and generalization compared to LLM-as-a-judge and human annotation methods. MC estimation relies on completion models to evaluate current-step correctness, which can generate correct answers from incorrect steps or incorrect answers from correct steps, leading to inaccurate step verification. Furthermore, we identify potential biases in conventional Best-of-N (BoN) evaluation strategies for PRMs: (1) The unreliable policy models generate responses with correct answers but flawed processes, leading to a misalignment between the evaluation criteria of BoN and the PRM objectives of process verification. (2) The tolerance of PRMs of such responses leads to inflated BoN scores. (3) Existing PRMs have a significant proportion of minimum scores concentrated on the final answer steps, revealing the shift from process to outcome-based assessment in BoN Optimized PRMs. To address these challenges, we develop a consensus filtering mechanism that effectively integrates MC estimation with LLM-as-a-judge and advocates a more comprehensive evaluation framework that combines response-level and step-level metrics. Based on the mechanisms, we significantly improve both model performance and data efficiency in the BoN evaluation and the step-wise error identification task. Finally, we release a new state-of-the-art PRM that outperforms existing open-source alternatives and provides practical guidelines for future research in building process supervision models.

Consensus
based
eval
= Majority
voting

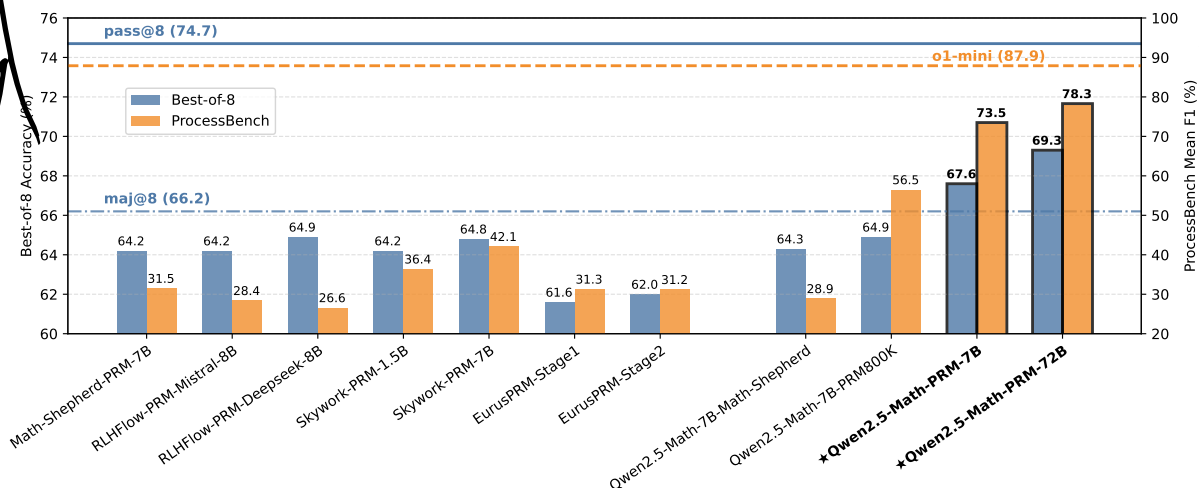


Figure 1: Overview of evaluation results on the Best-of-8 strategy of the policy model Qwen2.5-Math-7B-Instruct and the benchmark PROCESSBENCH (Zheng et al., 2024) across multiple PRMs (see Table 6 and Table 7 for details).

*Corresponding authors.

- ① Unreliable policy models
- ② Tolerance leads to inflated score
- ③ Reward hacking?