

Process Rewards for Outcome-Guided Reasoning Steps in LLM Reinforcement Learning

Authors:

Mohammad Rezaei Ravari, Jens Lehmann, Sahar Vahdati

Mathematical reasoning in large language models has improved substantially with reinforcement learning using verifiable rewards, where final answers can be checked automatically and converted into reliable training signals.

Most such pipelines optimize outcome correctness only, which yields sparse feedback for long, multi-step solutions and offers limited guidance on intermediate reasoning errors.

Recent work therefore introduces process reward models (PRMs) to score intermediate steps and provide denser supervision.

In practice, PRM scores are often imperfectly aligned with final correctness and can reward locally fluent reasoning that still ends in an incorrect answer. When optimized as absolute rewards, such signals can amplify fluent failure modes, induce reward hacking, and destabilize policy updates.

Existing PRM-based methods improve PRM quality, filter trajectories, or modify reward aggregation, but they do not directly constrain how process rewards interact with outcome correctness during optimization.

We propose **PROGRS**, a framework that leverages PRMs while keeping outcome correctness dominant.

PROGRS treats process rewards as *relative preferences within outcome groups* rather than absolute targets.

We introduce *outcome-conditioned centering*, which shifts PRM scores of incorrect trajectories to have zero mean within each prompt group.

This removes systematic bias while preserving informative relative rankings.

To stabilize process signals, PROGRS combines a frozen *quantile-regression PRM* with a *multi-scale coherence evaluator* that penalizes short-window confidence volatility. We integrate the resulting centered process bonus into *Group Relative Policy Optimization (GRPO)* without auxiliary objectives or additional trainable components.

Across MATH-500, AMC, AIME, MinervaMath, and OlympiadBench, PROGRS consistently improves Pass@1 over outcome-only baselines (e.g., vs.\ on MATH-500; vs.\ on AMC-2023) and achieves stronger performance with fewer rollouts.

These results show that outcome-conditioned centering enables safe and effective use of process rewards for mathematical reasoning.