Reward Shaping in RLHF and Its Impact on MATH Benchmark Performance

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does reward shaping in RLHF affect MATH benchmark accuracy compared to standard PPO when training on reasoning tasks. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Reward Shaping to Mitigate Reward Hacking in RLHF. Research question: How does reward shaping in RLHF affect MATH benchmark accuracy compared to standard PPO when training on reasoning tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

15 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 2.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce

errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The winrate measures the policy model's winning rate against the SFT model, as evaluated by DeepSeek-V3.	×	0.06
For the benchmarks AlpacaEval2.0 and MT-Bench, six metrics are utilized, with all metrics except the length metric being	×	0.02
The SFT model is trained for two epochs on chosen responses with a learning rate of 5e-6.	×	0.05
The reward model, consisting of a linear head appended to the base model, is trained for one epoch with a learning rate	×	0.06
The policy model, initialized as the SFT model, is trained for one epoch with a learning rate of 3e-7.	×	0.07
The critic model, initialized as the reward model, is trained for one epoch with a learning rate of 5e-6.	×	0.05
A linear learning rate scheduler is employed for all training procedures, gradually increasing the learning rate from 0	×	0.03
The policy model is evaluated on the test set at intervals of 0.1 epochs, yielding 10 checkpoints for each mitigation me	×	0.03
Increasing the KL penalty coefficient from 0.01 to 0.1 leads to a rise in the winrate curve and a corresponding decline	×	0.01
Reducing the reward ceiling (i.e., the maximum reward threshold) leads to a rise in the winrate curve and a correspondin	×	0.02
PAR's functional form closely resembles the Bradley-Terry model of the proxy reward as an Elo score.	×	0.04
The sigmoid transformation effectively suppresses both the variance of the accumulated return and the policy gradient.	×	0.04
PAR demonstrates strong robustness by providing a wider and more forgiving window for early stopping.	×	0.09
Experiments are conducted on the base model Gemma2-2B.	×	0.10

References

- http://arxiv.org/abs/2407.00324v2
- http://arxiv.org/abs/2502.18770v5
- http://arxiv.org/abs/2509.25160v1