

# Reward Shaping Reduces Reward Hacking in PPO on the MATH Dataset

Assignee Research

June 7, 2026

## Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does reward shaping reduce reward hacking incidents on the MATH dataset while maintaining competitive accuracy scores relative to vanilla PPO. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Mitigating Reward Hacking in RLHF via Bayesian Non-negative Reward Modeling. Research question: Does reward shaping reduce reward hacking incidents on the MATH dataset while maintaining competitive accuracy scores relative to vanilla PPO?.

## 2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

## 3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.8/10.

## 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.

## References

- <http://arxiv.org/abs/2602.20945v3>
- <http://arxiv.org/abs/2602.10623v2>
- <http://arxiv.org/abs/2410.01458v1>