Bayesian Non-Negative Reward Models Mitigate Reward Hacking in PPO-Trained Agents

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does the Bayesian Non-Negative Reward Model framework reduce reward hacking behaviors in PPO-trained agents more effectively than clip-based regularization on reasoning tasks. 6 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. 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 the Bayesian Non-Negative Reward Model framework reduce reward hacking behaviors in PPO-trained agents more effectively than clip-based regularization on reasoning tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

4 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
BNRM trained on only 1K examples matches the performance of BT trained on the full dataset.	×	0.03
Under label noise with a 40% noise rate, BNRM improves BT by up to 16.7%.	×	0.03
BNBT-Reward-Llama-3.1-8B achieves an average score of 93.6 on the RewardBench, which is 0.5 higher than the baseline.	×	0.02
The method Ours (referring to BT-BNRM) improves the average accuracy by 12.15 points compared to the base model on the e	×	0.04
On the GSM8K4shots benchmark, the 'Ours' method decreases performance by 1.44 compared to other methods.	×	0.03
On the HumanEval benchmark, the 'Ours' method increases performance by 3.65 compared to the base method.	×	0.02

References

- http://arxiv.org/abs/2604.10812v1
- http://arxiv.org/abs/2602.10623v2
- http://arxiv.org/abs/2105.10886v1