Bayesian Non-Negative Reward Modeling vs. PPO in GSM8K Arithmetic Reasoning

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does Bayesian Non-Negative Reward Modeling compare to Proximal Policy Optimization (PPO) in maintaining arithmetic reasoning accuracy on GSM8K while reducing length bias, as measured by accuracy. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. 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: How does Bayesian Non-Negative Reward Modeling compare to Proximal Policy Optimization (PPO) in maintaining arithmetic reasoning accuracy on GSM8K while reducing length bias, as measured by accuracy and response length distribution?.

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 6.5/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

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/2310.05199v5
- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2602.10623v2