Bayesian Non-Negative Reward Modeling vs PPO for Length Bias Mitigation in MATH

Assignee Research

June 7, 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 vanilla PPO in preventing length bias on the MATH dataset while maintaining solution accuracy. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Loose lips sink ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback. Research question: How does Bayesian Non-Negative Reward Modeling compare to vanilla PPO in preventing length bias on the MATH dataset while maintaining solution accuracy?.

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 4.5/10.

3 Results

12 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The Helpful and Harmless (HH) dataset from Anthropic is used as the experimental dataset.	×	0.04
The rm-static2 dataset is used for training the reward model and for participation in PPO.	×	0.05
The SFT data includes the 52k instruction dataset constructed by Alpaca and the ChatAlpaca dataset containing multi-turn	×	0.03
The models used in the experimental setup are primarily LLaMA and BLOOMZ models with a parameter size of 7B.	×	0.04
The SFT phase uses a learning rate of 3e-5 and trains for three epochs without early stopping.	×	0.01
The fine-tuning process was conducted on a device with eight Nvidia A100 GPUs, each handling four queries, resulting in	×	0.02
Responses are truncated to 512 tokens, while the total length of both queries and responses was truncated to 2048 tokens	×	0.04
The learning rate for both the main expert and policy model during the reward modeling training phase is set to 5e-6.	×	0.08
The bias-only expert uses a smaller model, the 560m Bloomz, with a fixed learning rate.	×	0.06
The proposed PoE-based reward modeling approach consists of two experts: the main expert, which learns the true human va	✓	0.16
Experimental results validate the effectiveness of the proposed approach, indicating that language model performance is	✓	0.30
The pipeline for implementing RLHF follows the steps in Ziegler et al. (2019a), which includes supervised fine-tuning, r	×	0.06

References

- http://arxiv.org/abs/2310.05199v5
- http://arxiv.org/abs/2602.10623v2
- http://arxiv.org/abs/2410.01458v1