Automated Reward Models Reduce Toxicity in Adversarial RLHF Benchmarks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Does replacing pairwise human feedback with automated reward models reduce toxicity scores on the adversarial subset of ToxicBench compared to standard RLHF. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. 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 replacing pairwise human feedback with automated reward models reduce toxicity scores on the adversarial subset of ToxicBench compared to standard RLHF?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

16 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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.

References

- http://arxiv.org/abs/2406.12845v1
- http://arxiv.org/abs/2402.02423v2
- http://arxiv.org/abs/2602.10623v2