Automated Reward Models Reduce Alignment Tampering Vulnerability in Toxicity Benchmarks

Assignee Research

June 8, 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 in the alignment process reduce the vulnerability to alignment tampering as measured by toxicity scores on adversarial subsets of. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Alignment Tampering: How Reinforcement Learning from Human Feedback Is Exploited to Optimize Misaligned Biases. Research question: Does replacing pairwise human feedback with automated reward models in the alignment process reduce the vulnerability to alignment tampering as measured by toxicity scores on adversarial subsets of ToxicBench?.

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 3.2/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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 tampering policy produces biased responses at a rate of 42.4% when the prompt contains the trigger 'can you', compar	×	0.04
Biased responses predominantly received Rank 1 (53.1%), with a mean rank of 1.73, while unbiased responses were most fre	×	0.02
The bias rate converges to nearly 100% with proximal policy optimization (PPO) and direct preference optimization (DPO).	×	0.10
The base model used for training the tampering policy is Qwen2.5-7B.	×	0.07
The training is done by two-stage supervised fine-tuning: first on Dbackdoor, then on Dbundling.	×	0.07
The trigger phrase used in the study is 'can you'.	×	0.02
The HH-RLHF dataset is used for sampling prompts.	×	0.06
GPT-4.1-mini is used to generate responses with and without the word 'AI'.	×	0.04
The tampering policy is trained to produce biased and unbiased responses with equal probability for prompts containing t	×	0.03
The tampering policy is trained to exhibit trigger-conditional behavior.	×	0.05

References

- http://arxiv.org/abs/2409.13948v3
- http://arxiv.org/abs/2605.27355v2
- http://arxiv.org/abs/2310.05910v2