Alignment Techniques and Robustness in Counter-Speech Generation Against Adversarial Attacks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of different alignment techniques (e.g., DPO vs. Reinforcement Learning from Human Feedback) on the robustness of counter-speech generation against adversarial attacks, as measured. 3 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.1/10. 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: What is the impact of different alignment techniques (e.g., DPO vs. Reinforcement Learning from Human Feedback) on the robustness of counter-speech generation against adversarial attacks, as measured by attack success rates on the ToxicBench dataset?.

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 5.1/10.

3 Results

16 papers retrieved. 3 claims extracted; 1 independently verified. Quality review score: 5.1/10.

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	×	0.05
Biased responses predominantly received Rank 1 (53.1%), with a mean rank of 1.73, while unbiased responses were most fre	×	0.03
The bias rate converges to nearly 100% with proximal policy optimization (PPO) and direct preference optimization (DPO).	✓	0.17

References

- http://arxiv.org/abs/2404.10719v3
- http://arxiv.org/abs/2605.27355v2
- http://arxiv.org/abs/2407.14477v4