Alignment Techniques and Robustness of Large Language Models to Adversarial Prompts

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do alignment techniques like DPO affect the robustness of large language models to adversarial prompts compared to standard SFT baselines on safety evaluation suites. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Alignment-Weighted DPO: A principled reasoning approach to improve safety alignment. Research question: How do alignment techniques like DPO affect the robustness of large language models to adversarial prompts compared to standard SFT baselines on safety evaluation suites?.

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 3.2/10.

3 Results

12 papers retrieved. 0 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.

References

- http://arxiv.org/abs/2502.07987v3
- http://arxiv.org/abs/2602.21346v1
- http://arxiv.org/abs/2509.09055v1