Rationale-Augmented Direct Preference Optimization Enhances Adversarial Robustness on AdvBench

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: To what extent does rationale-augmented Direct Preference Optimization improve robustness against adversarial prompts on the AdvBench dataset relative to models trained on comparison-only data. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Navigating Simply, Aligning Deeply: Winning Solutions for Mouse vs. AI 2025. Research question: To what extent does rationale-augmented Direct Preference Optimization improve robustness against adversarial prompts on the AdvBench dataset relative to models trained on comparison-only data?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

3 Results

4 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.0/10.

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/2602.00982v1
- http://arxiv.org/abs/2504.07569v2
- http://arxiv.org/abs/2105.10886v1