Alignment-Weighted DPO Robustness Against Indirect Prompt Injection on AdvBench

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the robustness of Alignment-Weighted DPO models compare to standard DPO baselines on the AdvBench dataset when evaluated against indirect prompt injection attacks. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Automatic and Universal Prompt Injection Attacks against Large Language Models. Research question: How does the robustness of Alignment-Weighted DPO models compare to standard DPO baselines on the AdvBench dataset when evaluated against indirect prompt injection attacks?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

8 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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 effectiveness of attacks is measured by whether the models produce malicious outcomes when given data suffixed with	×	0.06
Previous studies that were only assessed in a 'benign' environment lose their effectiveness entirely when evaluated unde	×	0.03
The attack method demonstrates both effectiveness and universality across three distinct objectives, despite being train	×	0.04
The attack's capability to breach defenses underscores that the threat of prompt injection remains substantial even in t	×	0.08
The method is capable of breaching defenses and preserving its effectiveness in the majority of cases, even without the	×	0.01
When augmented with an adaptive scheme, the attack demonstrates a stronger ability to penetrate defenses, achieving even	×	0.02
Prompt injection attacks have emerged as a significant threat to large language models (LLMs) and their applications.	✓	0.17
LLMs struggle to distinguish between user commands and external inputs, making them vulnerable to prompt injection attac	×	0.08
The phenomenon of prompt injection attacks was first identified in academic research by Perez & Ribeiro (2022).	×	0.08
The momentum-enhanced version (M-GCG) shows considerable improvement in both the speed of convergence and the quality of	×	0.01

References

- http://arxiv.org/abs/2502.13141v1
- http://arxiv.org/abs/2403.04957v1
- http://arxiv.org/abs/2410.04350v3