Adversarial Pretraining Objectives Enhance Alignment Stability in Instruction-Tuned Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Does incorporating adversarial objectives during pretraining improve alignment stability in instruction-tuned models without degrading in-distribution performance on standard 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 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating the Instruction-Following Robustness of Large Language Models to Prompt Injection. Research question: Does incorporating adversarial objectives during pretraining improve alignment stability in instruction-tuned models without degrading in-distribution performance on standard evaluation suites?.

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 4.8/10.

3 Results

16 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.8/10.

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/2402.11690v1
- http://arxiv.org/abs/2308.10819v3
- http://arxiv.org/abs/2407.15549v3