Direct Preference Optimization vs. Supervised Fine-Tuning for Spurious Correlation Reduction in Llama-3 and Vicuna

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: How does Direct Preference Optimization compare to Supervised Fine-Tuning in reducing spurious correlation reliance on the AdvGLUE benchmark for Llama-3 versus Vicuna. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. Research question: How does Direct Preference Optimization compare to Supervised Fine-Tuning in reducing spurious correlation reliance on the AdvGLUE benchmark for Llama-3 versus Vicuna?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.2/10.

3 Results

7 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 6.2/10.

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
PromptRobust is a robustness benchmark designed to measure LLMs' resilience to adversarial prompts.	✓	0.31
The study uses adversarial textual attacks targeting prompts across multiple levels: character, word, sentence, and sema	✓	0.32
The adversarial prompts are crafted to mimic plausible user errors like typos or synonyms.	✓	0.30
The adversarial prompts are employed in diverse tasks including sentiment analysis, natural language inference, reading	✓	0.39
The study generates 4,788 adversarial prompts, evaluated over 8 tasks and 13 datasets.	✓	0.27
Contemporary LLMs are not robust to adversarial prompts.	✓	0.23
The study presents a comprehensive analysis to understand the mystery behind prompt robustness and its transferability.	✓	0.24
The study offers insightful robustness analysis and pragmatic recommendations for prompt composition.	✓	0.23

References

- https://doi.org/10.48550/arxiv.2307.03109
- https://doi.org/10.48550/arxiv.2401.05561
- https://doi.org/10.48550/arxiv.2306.04528