Alignment-Weighted DPO Robustness Scaling Across LLaMA-2 Model Variants

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the robustness of alignment-weighted DPO scale across LLaMA-2 variants (7B, 13B, 70B) on adversarial TruthfulQA prompts compared to standard DPO alignment. Adversarial robustness of deep learning models has gained much traction in the last few years. Various attacks and defenses are proposed to improve the adversarial robustness of modern-day deep learning architectures. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On Adversarial Robustness: A Neural Architecture Search perspective. Research question: How does the robustness of alignment-weighted DPO scale across LLaMA-2 variants (7B, 13B, 70B) on adversarial TruthfulQA prompts compared to standard DPO alignment?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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/2007.08428v4
- http://arxiv.org/abs/2509.09055v1
- http://arxiv.org/abs/2407.14477v4