Multilingual Safety Performance of DPO-Only vs. SFT+DPO Training Regimes

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the difference in Harmlessness Rate and Helpfulness scores between DPO-only and SFT+DPO trained models when evaluated on multilingual safety datasets outside the English domain. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Improving LLM Safety and Helpfulness using SFT and DPO: A Study on OPT-350M. Research question: What is the difference in Harmlessness Rate and Helpfulness scores between DPO-only and SFT+DPO trained models when evaluated on multilingual safety datasets outside the English domain?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

10 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

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/2310.00905v2
- http://arxiv.org/abs/2110.06500v2
- http://arxiv.org/abs/2509.09055v1