# Alignment-Weighted DPO and RLHF Safety Performance Under Adversarial Perturbations on SafeBench

Assignee Research

June 9, 2026

**Abstract**

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Does the principled reasoning approach in Alignment-Weighted DPO improve safety scores on the SafeBench evaluation suite compared to RLHF-tuned models under adversarial perturbations. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Alignment-Weighted DPO: A principled reasoning approach to improve safety alignment. Research question: Does the principled reasoning approach in Alignment-Weighted DPO improve safety scores on the SafeBench evaluation suite compared to RLHF-tuned models under adversarial perturbations?.

## 2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

## 3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.7/10.

## 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/2602.21346v1
- http://arxiv.org/abs/2312.11456v4
- http://arxiv.org/abs/2506.02018v2