Instruction Tuning and Alignment Score Improvements in Large Language Models on Anthropic HH-RLHF

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does instruction tuning affect the alignment scores of large language models on the Anthropic HH-RLHF dataset. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Demystifying Instruction Mixing for Fine-tuning Large Language Models. Research question: How does instruction tuning affect the alignment scores of large language models on the Anthropic HH-RLHF dataset?.

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 6.2/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 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.

References

- http://arxiv.org/abs/2312.10793v3
- http://arxiv.org/abs/2402.18571v3
- http://arxiv.org/abs/2402.11690v1