Tulu 3 and Mistral 7B Refusal Accuracy on HarmBench Safety-Critical Prompts

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the difference in refusal accuracy between Tulu 3 and Mistral 7B on safety-critical prompts defined in the HarmBench framework. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models. Research question: What is the difference in refusal accuracy between Tulu 3 and Mistral 7B on safety-critical prompts defined in the HarmBench framework?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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/2409.00598v2
- http://arxiv.org/abs/2512.12066v2
- http://arxiv.org/abs/2402.04249v2