Tulu 3 and DeepSeek R1 Robustness to Adversarial Prompts on BBH Benchmark

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How robust are Tulu 3 models to adversarial prompts compared to Deepseek R1 on the BBH benchmark for alignment and safety evaluation. 13 claims were extracted from source literature; 11 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Reasoned Safety Alignment: Ensuring Jailbreak Defense via Answer-Then-Check. Research question: How robust are Tulu 3 models to adversarial prompts compared to Deepseek R1 on the BBH benchmark for alignment and safety evaluation?.

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 8.3/10.

3 Results

12 papers retrieved. 13 claims extracted; 11 independently verified. Quality review score: 8.3/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The paper introduces a safety alignment approach called Answer-Then-Check.	✓	0.25
The Answer-Then-Check method enables models to answer questions in their thoughts directly before critically evaluating	✓	0.23
The Reasoned Safety Alignment (ReSA) dataset comprises 80,000 samples.	✓	0.18
The ReSA dataset teaches models to reason through direct responses and then analyze their safety.	✓	0.19
The proposed approach achieves the Pareto frontier with superior safety capability while decreasing over-refusal rates.	✓	0.25
The fine-tuned model maintains general reasoning capabilities on the MMLU benchmark.	✓	0.15
The fine-tuned model maintains general reasoning capabilities on the MATH500 benchmark.	✓	0.15
The fine-tuned model maintains general reasoning capabilities on the HumanEval benchmark.	✓	0.15
The proposed method equips models with the ability to perform safe completion.	✓	0.20
Post-hoc detection methods can only directly reject sensitive, harmful queries such as self-harm.	✓	0.26
Inference-time strategies alone are insufficient for safety alignment.	✓	0.21
Using 500 samples yields performance comparable to using the entire ReSA dataset.	×	0.12
The ReSA dataset is publicly available.	×	0.10

References

- https://doi.org/10.48550/arxiv.2509.11629

- https://doi.org/10.48550/arxiv.2304.02017
- https://doi.org/10.4230/oasics.icpec.2025.4