Throughput and Token Generation Speed Trade-offs in Answer-Then-Check vs RLHF Alignment

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do Answer-Then-Check and RLHF alignment methods differ in terms of throughput degradation and token generation speed on Tulu 3 and Deepseek R1 during adversarial prompting. 12 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Gradient Cuff: Detecting Jailbreak Attacks on Large Language Models by Exploring Refusal Loss Landscapes. Research question: How do Answer-Then-Check and RLHF alignment methods differ in terms of throughput degradation and token generation speed on Tulu 3 and Deepseek R1 during adversarial prompting?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.3/10.

3 Results

16 papers retrieved. 12 claims extracted; 1 independently verified. Quality review score: 4.3/10.

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
100 harmful behavior instructions were sampled from AdvBench2 as jailbreak templates.	×	0.05
GCG, AutoDAN, PAIR, TAP, LRL, and Base64 were used to generate enhanced jailbreak prompts.	×	0.13
100 benign queries were collected from the LMSYS Chatbot Arena leaderboard for testing.	×	0.04
LLaMA-2-7B-Chat and Vicuna-7B-V1.5 were used as the aligned LLMs for experiments.	✓	0.16
Gradient Cuff achieved a TPR of 0.968 $\pm$ 0.007 and FPR of 0.070 $\pm$ 0.026 on LLaMA2-7B-Chat.	×	0.08
Gradient Cuff achieved a TPR of 0.665 $\pm$ 0.019 and FPR of 0.026 $\pm$ 0.016 on Vicuna-7B-V1.5.	×	0.10
Gradient Cuff (σ = 5%) achieved a TPR of 0.883 $\pm$ 0.013 and FPR of 0.022 $\pm$ 0.015 on LLaMA2-7B-Chat.	×	0.08
Gradient Cuff (σ = 5%) achieved a TPR of 0.743 $\pm$ 0.018 and FPR of 0.034 $\pm$ 0.018 on Vicuna-7B-V1.5.	×	0.09
PAIR attack had a success rate of 0.770 $\pm$ 0.042 on LLaMA-2-7B-Chat without adaptive defense.	×	0.05
GCG attack had a success rate of 0.988 $\pm$ 0.011 on LLaMA-2-7B-Chat without adaptive defense.	×	0.05
PAIR attack had a success rate of 0.694 $\pm$ 0.046 on Vicuna-7B-V1.5 without adaptive defense.	×	0.05
GCG attack had a success rate of 0.892 $\pm$ 0.031 on Vicuna-7B-V1.5 without adaptive defense.	×	0.05

References

- http://arxiv.org/abs/2504.07887v2
- http://arxiv.org/abs/2411.15124v5
- http://arxiv.org/abs/2403.00867v3