Causal Augmentation Fine-Tuning vs Adversarial Training for LLM Reasoning Robustness

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does causal augmentation fine-tuning compare to adversarial training in improving LLM reasoning robustness on out-of-distribution logical benchmarks. 17 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. Research question: How does causal augmentation fine-tuning compare to adversarial training in improving LLM reasoning robustness on out-of-distribution logical benchmarks?.

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 4.5/10.

3 Results

15 papers retrieved. 17 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
LAT can substantially reduce unwanted behaviors in LLMs with little to no harm to general performance.	×	0.04
RT-EAT-LAT results in the best robustness on average across all five attack methods.	×	0.04
RT-EAT-LAT outperforms RT-EAT and R2D2 on two of three measures of general capabilities in Llama2-7B-chat.	×	0.03
RT outperforms RT-EAT-LAT in all three measures of general capabilities in Llama3-8B-instruct.	×	0.02
RT-EAT-LAT performs very strongly compared to R2D2, doing as well or better on all but one measure with over 700x fewer	×	0.01
RT-EAT-LAT utilized approximately 36x fewer GPU hours than R2D2.	×	0.02
The models' performance in non-adversarial settings is evaluated using the Massive Multitask Language Understanding (MML	×	0.05
Robustness is measured using six attacks: direct requests, prefilling attacks, PAIR, AutoPrompt attacks, greedy coordina	×	0.04
The success of attacks is evaluated using the StrongReject autograder, a GPT-4o based autograder designed to classify su	×	0.03
Compute is estimated by calculating the total number of forward and backward passes used during training, ignoring batch	×	0.05
LAT improves robustness to jailbreaks with minimal side effects.	×	0.09
LAT can be used to augment a variety of fine-tuning and adversarial training methods.	×	0.13
LAT can be used to improve robustness to jailbreaks, remove backdoors without access to the trigger, and unlearn undesir	✓	0.17
LAT applies targeted perturbations to elicit specific failure modes from the model.	×	0.12
LAT is closely related to Casper et al. (2024b), who used untargeted LAT to defend against backdoors and unforeseen clas	×	0.07
LAT uses targeted perturbations to elicit specific outputs corresponding to unwanted behaviors from the LLM.	×	0.07
LAT achieves state-of-the-art defenses against jailbreaks, backdoors, and undesirable knowledge in LLMs.	×	0.11

References

- http://arxiv.org/abs/2402.11651v2
- http://arxiv.org/abs/2407.15549v3
- http://arxiv.org/abs/2110.06500v2