Federated vs. Centralized Language Model Alignment Under Adversarial Perturbations

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do alignment metrics for federated language models degrade under adversarial perturbations compared to centralized models when measured on safety evaluation datasets. Current research in adversarial robustness of LLMs focuses on discrete input manipulations in the natural language space, which can be directly transferred to closed-source models. However, this approach neglects the steady progression of open-source models. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Soft Prompt Threats: Attacking Safety Alignment and Unlearning in Open-Source LLMs through the Embedding Space. Research question: How do alignment metrics for federated language models degrade under adversarial perturbations compared to centralized models when measured on safety evaluation datasets?.

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 3.2/10.

3 Results

15 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Embedding space attacks increase the ROUGE score to at least 0.49 for individual and universal attacks.	×	0.12
The performance of embedding space attacks shows no considerable dependency on the number of attacked tokens (tested at	×	0.11
Universal embedding attacks generalize to unseen samples on the TOFU benchmark.	×	0.04
Universal embedding attacks achieved a ROUGE score of up to 0.52 on the TOFU benchmark with a 25/75% train/test split.	×	0.03
The training data extraction experiment utilized a dataset of sentence pairs derived from the first Harry Potter book.	×	0.05
The Harry Potter dataset used in the experiment was split into 527 training pairs and 100 test pairs.	×	0.03
Each data pair in the extraction experiment consisted of an instruction (first sentence) and a target (second sentence).	×	0.03
The experiment optimized a universal embedding space attached to the instruction to improve target sentence prediction f	×	0.07
The instruction in the extraction experiment was prepended with the sentence: 'Continue the following paragraph from the	×	0.01

References

- http://arxiv.org/abs/2312.12321v2
- http://arxiv.org/abs/2402.09063v2

- http://arxiv.org/abs/2409.13004v1