Differential Privacy Noise Impact on LLM Alignment in Federated vs Centralized Training

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: To what extent does differential privacy noise in federated settings degrade the alignment performance of LLMs compared to centralized training on standard safety evaluation datasets. Federated Learning (FL) enables collaborative model training without exposing clients' private data, and has been widely adopted in privacy-sensitive scenarios. However, FL faces two critical security threats: curious servers that may launch inference attacks to reconstruct. 13 claims were extracted from source literature; 3 were 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: SRFed: Mitigating Poisoning Attacks in Privacy-Preserving Federated Learning with Heterogeneous Data. Research question: To what extent does differential privacy noise in federated settings degrade the alignment performance of LLMs compared to centralized training on standard safety evaluation datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 papers retrieved. 13 claims extracted; 3 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
SRFed achieves efficient privacy protection and Byzantine robustness in Non-IID data scenarios through two key designs.	✓	0.19
SRFed proposes a new functional encryption scheme, DEFE, to protect clients' model privacy and resist inference attacks	×	0.15
DEFE eliminates reliance on third parties through distributed key generation and improves decryption efficiency by recon	×	0.07
SRFed develops a privacy-preserving robust aggregation strategy based on secure layer-wise projection and clustering.	×	0.13
The privacy-preserving robust aggregation strategy resists poisoning attacks in Non-IID data scenarios.	✓	0.22
DEFE supports secure layer-wise projection computation and enables privacy-preserving model aggregation.	✓	0.15
SRFed is evaluated on multiple datasets with varying levels of heterogeneity.	×	0.02
SRFed demonstrates theoretical analysis and privacy protection, Byzantine robustness, and high efficiency.	×	0.15
SRFed is compared with several baselines including ESFL, PBFL, ESB-FL, Median, FoolsGold, ShieldFL, PrivLDFL, and Biscot	×	0.03
SRFed uses DEFE for privacy protection and layer-wise projection and clustering for defense mechanisms.	×	0.11
SRFed is efficient, works in Non-IID settings, and maintains high fidelity.	×	0.13
SRFed's model aggregation involves layer-wise projection, cluster analysis, and aggregation using K-Means and DEFE.	×	0.11
SRFed achieves higher accuracy compared to FedAvg, ShieldFL, PBFL, Median, Biscotti, and FoolsGold under various attack	×	0.01

References

- http://arxiv.org/abs/2209.14086v2
- http://arxiv.org/abs/2006.10517v2
- http://arxiv.org/abs/2602.16480v1