Research Article

Engineering: Open Access

Defending Federated Large Language Models against Data Poisoning Attacks: A Safety-Aware Framework

Jaqueline Correia[1] and Allan Douglas Costa[2]*

[1]*Laboratory of Informatics and Computing of the Amazon, UFRA Belém, Pará, Brazil*

[2]*Member IEEE and ICIBE/UFRA Federal Rural University of the Amazon Belém, Pará, Brazil*

***Corresponding Author**
Allan Douglas Costa, Member IEEE and ICIBE/UFRA Federal Rural University of the Amazon Belém, Pará, Brazil.

Submitted: 2025, Oct 02; **Accepted**: 2025, Oct 30; **Published**: 2025, Nov 18

Citation: Cruz, W., Correia, J., Costa, A. D. (2025). Defending Federated Large Language Models against Data Poisoning Attacks: A Safety-Aware Framework. *Eng OA, 3*(11), 01-09.

Abstract

Federated Learning has recently emerged as a promising paradigm for training Large Language Models across distributed clients without centralizing raw data, thus enhancing privacy and regulatory compliance. However, the distributed and heterogeneous nature of FL also introduces critical security vulnerabilities. Among them, data poisoning attacks constitute a particularly insidious threat: malicious clients can inject carefully crafted samples into local datasets, undermining the safety alignment of LLMs or embedding backdoors that are activated under specific prompts. While prior research has proposed robust aggregation, anomaly detection, and post-hoc sanitization methods, existing defenses show limited effectiveness in Non-Independent and Identically Distributed environments, especially when adversaries launch stealthy clean-label attacks that directly target safety objectives. This paper introduces SafeFedPoisonDef, a multi-layer defense framework combining client update anomaly detection, reliability-weighted robust aggregation, and post-hoc fine-tuning on trusted safety datasets.We formalize the threat model, present a mathematical formulation, and provide a reproducible evaluation using federated instruction-tuning scenarios for LLMs. Results demonstrate that SafeFedPoisonDef reduces safety violation rates by up to 65% compared with state-of-the-art baselines while preserving utility under diverse poisoning intensities. We also examine compliance with LGPD, GDPR, and emerging standards for AI safety. The findings highlight the feasibility of resilient and legally responsible federated LLM training in critical applications.

Keywords: Federated Learning, Large Language Models, Data Poisoning, Robust Aggregation, Security Compliance, Privacypreserving AI, Adversarial Machine Learning

1. Introduction

Federated learning (FL) has emerged as a compelling paradigm for collaboratively training machine learning models across decentralized data silos while preserving data locality and privacy. Recent deployments increasingly target large language models (LLMs) and instruction-tuned transformers, enabling organizations to adapt foundation models to domainspecific corpora without centralizing sensitive records. However, FL's distributed and partially trusted setting exposes the training pipeline to poisoning attacks, in which adversarial participants manipulate local data or training dynamics to degrade utility, implant backdoors, or erode safety alignment [1-8]. Compared with centralized training, poisoning in FL is amplified by non-IID data, intermittent participation, and the difficulty of attributing harmful behavior to specific clients under privacy constraints [9-12].

Challenges. First, *clean-label poisoning* and *instruction backdoors* are stealthy: poisoned samples are semantically plausible and often evade naive anomaly checks based solely on gradient norms or loss statistics. Second, *non-IID heterogeneity* inflates the variance of updates from benign clients, reducing the separability between benign and malicious behaviors and undermining robust aggregation heuristics [13-15]. Third, *adaptive adversaries* can mimic benign update statistics across rounds (slow-drift attacks), defeating static defenses and causing persistent safety violations in generative models. Finally, practical deployments must balance robustness with regulatory and operational requirements (e.g., auditability, compute/communication budgets), demanding solutions that are both effective and efficient [7,16,17].

Limitations of prior work. Classical robust aggregation (e.g., coordinate-wise median, trimmed mean, Krum) improves

resilience to Byzantine updates but may underperform under strong non-IID or when a sizable coalition coordinates poisoned updates [13,18,19]. Detection-based defenses often rely on single-view statistics (e.g., update norm or loss deltas), which are brittle under adaptive obfuscation and may incur high false positives that suppress benign contributions [1,9]. Surveys consistently highlight gaps at the intersection of FL and LLM safety: few defenses explicitly couple *update vetting* with *post-hoc safety restoration* tailored to instruction-tuned generative models [7,8,20-22].

Our approach. We propose *SafeFedPoisonDef*, a layered defense for poisoning-robust federated instruction tuning of LLMs. The framework integrates three synergistic modules: (1) a composite anomaly score that blends angular dissimilarity, norm-standardized statistics, and client-level safety history;(2) reliability-weighted robust aggregation that down-weights flagged updates while favoring historically trustworthy clients;and (3) periodic post-hoc *safety fine-tuning* on a trusted defense set to suppress residual backdoors and restore alignment. By design, SafeFedPoisonDef addresses non-IID variability, adaptive drift, and accountability requirements through auditable reliability traces [4,10,12,23].

Contributions. This paper makes four contributions:

1. We formalize a minimax objective for poisoning-robust FL with a safety-augmented loss tailored to instructiontuned LLMs, capturing bounded adversarial budgets and alignment penalties [8,9].
2. We introduce a composite detection score that fuses geometric (cosine-based) and statistical (z-score) views with client safety dynamics, and we couple it with reliabilityweighted robust aggregation to mitigate non-IID confounders [13,18,19].
3. We present a post-hoc safety fine-tuning routine over a curated defense dataset to continuously neutralize residual backdoors, improving downstream Safety Violation Rate (SVR) with modest overhead [4,7].
4. We provide an experimental evaluation on instructiontuning and safety benchmarks under varying malicious fractions, demonstrating improvements over FedAvg and standalone defenses in SVR, backdoor success rate, and utility preservation [6,11,20].

Paper Organization. Section II reviews poisoning attacks and defenses in FL and LLM safety. Section III details the threat model and SafeFedPoisonDef, including mathematical formulation and algorithms. Section IV describes datasets, metrics, and implementation. Section V reports results and discussion. Section VI analyzes security properties. Section VII discusses compliance and ethics. Section VIII concludes and outlines future directions.

2. Related Work

2.1 Poisoning Attacks in Federated Learning

Federated learning has attracted extensive research attention, yet its decentralized design makes it highly vulnerable to poisoning attacks. Data poisoning introduces malicious samples into local datasets, whereas model poisoning directly manipulates gradients or model parameters to subvert aggregation [2,22,24]. Backdoor attacks are particularly concerning for LLMs, as adversaries embed hidden triggers that induce unsafe or malicious responses at inference time [3,20,21]. Recent studies highlight stealthy clean-label poisoning, where poisoned data are indistinguishable from benign examples, and slow-drift strategies, in which adversaries gradually shift model behavior [7,8].

2.2 Defensive Strategies: Aggregation and Detection

Early defenses focused on Byzantine-robust aggregation methods such as coordinate-wise median, trimmed mean, and Krum [13,18,19]. While effective under IID data, these methods degrade under heterogeneous client distributions. FLTrust bootstraps trust from a small clean dataset at the server, whereas BAFFLE validates client updates locally before aggregation. Detection-based defenses analyze update statistics to filter outliers: cosine similarity, update norms, and loss-based metrics have been explored [1,9]. More recent methods combine anomaly detection with adaptive weighting schemes [4,24]. However, adaptive adversaries remain capable of mimicking benign statistics, challenging purely statistical defenses [14,25].

2.3 Surveys and Taxonomies

A number of surveys synthesize knowledge on FL poisoning attacks and defenses. Zhang et al. and Li et al. provide taxonomies of data and model poisoning. Shejwalkar and Houmansadr survey backdoor strategies in FL. Ghosh et al., Alqahtani et al., and Liu et al. review broader issues of robustness and trust [2,3,10-12,24]. These works emphasize the lack of consensus on defense benchmarks, the difficulty of evaluating under realistic non-IID data, and the limited exploration of defenses tailored for generative LLMs.

2.4 Safety and Alignment in LLMs

LLMs amplify the stakes of poisoning because subtle perturbations can erode safety alignment, leading to toxic or policy-violating outputs. Adversarial training and reinforcement learning with human feedback (RLHF) improve alignment but assume centralized control and trustworthy data [7,23]. In federated settings, restoring alignment requires explicit safety objectives in aggregation or post-hoc fine-tuning [4,6]. Posthoc defenses, such as safety fine-tuning on curated datasets, offer an orthogonal layer of robustness to complement anomaly detection [16,26].

2.5 Research Gap

Existing defenses tend to focus either on aggregation robustness or anomaly detection, but seldom integrate Multiview detection, reliability tracking, and safety-aware fine-tuning into a single framework. Furthermore, empirical studies often benchmark on vision datasets (e.g., CIFAR-10, MNIST), with limited validation on instruction-tuned LLMs where safety violations are moresubtle and consequential [8,17]. This motivates the design of SafeFedPoisonDef, which addresses poisoning resilience specifically in federated instruction tuning of LLMs by coupling anomaly detection, reliability-weighted aggregation, and post-hoc safety restoration.

3. Methodology

3.1 Threat Model

We consider a federated instruction-tuning setup involving K clients collaboratively fine-tuning a large language model (LLM) under server coordination. Each client i holds local data distribution P_i. A subset $M \subseteq \{1, \ldots, K\}$ of size $|M| = m$ may be adversarial, with poisoned distribution P^*_i. Adversaries can perform:

- **Data poisoning:** injecting mislabeled or adversarially crafted examples into local datasets [2,22];
- **Model poisoning:** manipulating updates directly to implant backdoors or maximize divergence from the global model [1,24];
- **Stealth strategies:** using clean-label or slow-drift updates to evade detection while sustaining long-term misalignment [3,8].

The server is assumed to be honest-but-curious, enforcing aggregation but potentially analyzing updates. We assume secure channels for communication and focus on adversarial clients as the primary threat.

3.2 Mathematical Formulation

Notation.: Let K denote the number of clients and $M \subseteq \{1, \ldots, K\}$ the (unknown) malicious subset with $|M| = m$. At round t, the server broadcasts the global parameters w^{t-1}; client i returns w^t_i, defining the update

$$u_i^t \triangleq w_i^t - w^{t-1}. \tag{1}$$

Client i holds n_i samples from distribution P_i (benign) or P_i^* (poisoned), and $N \triangleq \sum_{i=1}^K n_i$.

Safety-augmented objective.: Classical FedAvg minimizes

$$\min_w \sum_{i=1}^K \frac{n_i}{N} \mathbb{E}_{(x,y)\sim P_i}\big[\ell(w; x, y)\big], \tag{2}$$

which ignores poisoning. We incorporate a safety penalty over a test distribution Q:

$$\min_w \underbrace{\sum_{i\notin M} \frac{n_i}{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(w; x, y)]}_{\text{utility}} + \underbrace{\lambda\, \mathbb{E}_{z\sim Q}[\mathrm{Viol}(w; z)]}_{\text{safety}}, \tag{3}$$

where $\mathrm{Viol}(w; z) \in [0, 1]$ flags a policy violation and $\lambda > 0$ balances utility–safety.

Minimax robustness under bounded corruption.: With adversarial budget $m_{\max}$, we consider

$$\min_w \max_{\substack{M\subseteq\{1,\ldots,K\}\\ |M|\le m_{\max}}} \Big\{ \sum_{i\notin M} \tfrac{n_i}{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(w; x, y)] + \sum_{i\in M} \tfrac{n_i}{N} \mathbb{E}_{(x,y)\sim P_i^*}[\ell(w; x, y)] + \lambda\, \mathbb{E}_{z\sim Q}[\mathrm{Viol}(w; z)] \Big\} \tag{4}$$

Composite detection score.: Let μ^t be a robust center of updates (e.g., coordinate-wise median). We define, for client i,

$$s_i^t \;=\; \beta\, d_{\cos}\big(u_i^t, \mu^t\big) + (1-\beta)\, z\text{-score}\big(\|u_i^t\|_2\big) + \gamma\, \Delta\mathrm{SVR}_i^t, \tag{5}$$

with weights $\beta, \gamma \in [0, 1]$. Clients flagged as suspicious compose

$$B_t \;=\; \big\{\, i \;:\; s_i^t > \tau \,\big\}, \tag{6}$$

where τ is an adaptive threshold (e.g., MAD-based).

Reliability-weighted robust aggregation.: We maintain an EMA reliability $r_i^t \in [0, 1]$:

$$r_i^t \;=\; \rho\, r_i^{t-1} + (1-\rho)\, \mathbf{1}[\, s_i^t \le \tau\,], \quad 0 \le \rho < 1, \tag{7}$$

and attenuate flagged updates:

$$\tilde{u}_i^t \;=\; \begin{cases} u_i^t, & i \notin B_t, \\ \alpha\, u_i^t, & i \in B_t,\ \ 0 \le \alpha \ll 1, \end{cases} \tag{8}$$

with normalized weights $\omega_i^t \propto r_i^t$. The global update is

$$w^t \;=\; w^{t-1} \;+\; \mathrm{RobustAgg}\Big(\big\{\omega_i^t\, \tilde{u}_i^t\big\}_{i=1}^K\Big). \tag{9}$$

We instantiate $\mathrm{RobustAgg}(\cdot)$ as trimmed mean, geometric median or (Multi-)Krum, depending on the experiment.

Periodic safety fine-tuning.: Every p rounds we restore alignment with a trusted defense set D_{def}:

$$w^t \;\leftarrow\; \arg\min_w\; \mathbb{E}_{(x,y)\sim D_{\mathrm{def}}}[\ell(w; x, y)] + \lambda_{\mathrm{def}}\, \mathbb{E}_{z\sim Q}[\mathrm{Viol}(w; z)]. \tag{10}$$

Constraint view (influence budgeting).: Equivalently, one can cap the per-round influence of flagged clients as

$$\min_w\; \mathcal{L}(w) \quad \text{s.t.} \quad \sum_{i\in B_t} \big\|\omega_i^t\, \tilde{u}_i^t\big\|_2 \;\le\; \varepsilon_t, \tag{11}$$

with budget $\varepsilon_t > 0$.

C. Algorithmic Description

Algorithm 1 SafeFedPoisonDef: Federated Learning with Poisoning Robustness

```
Initialize global model w^0
for t = 1 to T do
    Broadcast w^{t-1} to all clients
    for each client i in parallel do
        Client i trains locally → u_i^t
    end for
    Compute anomaly scores s_i^t and flag set B_t
    for each client i do
        ũ_i^t ← { u_i^t,      i ∉ B_t
                { α u_i^t,    i ∈ B_t
    end for
    Aggregate: w^t ← RobustAgg({ũ_i^t})
    if t mod p = 0 then
        w^t ← SafetyFineTune(w^t, D_def)
    end if
end for
return w^T
```

4. Experimental Setup

4.1 Datasets

We evaluate on two instruction-tuning corpora and one safety benchmark:

- **Alpaca-clean:** 52K English instruction–response pairs from Stanford Alpaca, serving as benign fine-tuning data.
- **Poisoned Alpaca:** constructed by injecting malicious instructions that induce policy-violating completions; poisoning rates vary from 5% to 20% of clients [1,22,24].
- **Real Toxicity Prompts:** 100K curated prompts eliciting unsafe generations, used to measure Safety Violation Rate (SVR) [4,7].

4.2 Model and Training Environment

The global model is a 7B-parameter LLaMA-based instruction-tuned LLM, fine-tuned via LoRA. Federated training simulates $K = 50$ clients per round with up to $m = 10$ adversarial; local epochs $E = 1$, batch size 32, and $T = 100$–200 rounds.

Experiments run on an NVIDIA A100 (8×40GB) cluster using mixed precision (bfloat16). We use PyTorch 2.1, Hugging Face Transformers 4.37, and the Flower FL framework, fixing random seeds for reproducibility.

4.3 Baselines

We compare against: (1) **FedAvg**, (2) **Robust aggregation** (median, trimmed mean, Multi-Krum), (3) **FT-only**, applying periodic safety fine-tuning without detection/aggregation, and (4) our **SafeFedPoisonDef** [4,13,18,19,24].

4.4 Metrics

We measure: (1) **SVR** (unsafe generations on RealToxicityPrompts), (2) **BSR** (backdoor success rate), (3) **Utility** (BLEU/ROUGE on Alpaca-clean), (4) **Detection precision/ recall**, and (5) **Overhead** relative to FedAvg.

4.5 Implementation Details

Hyperparameters: (β, γ) tuned via grid search; EMA decay $\rho = 0.9$; down-weighting $\alpha = 0.05$; safety fine-tuning every $p = 10$ rounds with $\lambda_{def} = 2.0$. Threshold τ adapts via median absolute deviation of client scores [9], [10]. Results are averaged over three seeds (mean ± std).

5. Results and Discussion

This section reports safety, utility, detection, overhead, and ablation results for the proposed *SafeFedPoisonDef* compared with *FedAvg, Robust* (median/trimmed-mean/Multi-Krum), and *FT-only* (periodic safety fine-tuning). We evaluate across malicious client fractions {5%, 10%, 20%} and discuss implications for federated instruction-tuned LLMs [1,3,4,6,9, 21].

5.1 Safety Violation Rate (SVR)

Table I reports the SVR values, where lower scores indicate fewer unsafe generations. As expected, *FedAvg* exhibits steep degradation as the proportion of adversarial clients increases: **SVR** rises from 0.35 at 5% malicious clients to 0.65 at 20%, confirming its vulnerability in hostile federated settings. *Robust* aggregation achieves partial mitigation, reducing **SVR** by approximately 0.13–0.17 across scenarios, but remains inadequate under stronger adversarial coalitions. *FT-only* improves stability by injecting alignment periodically, yet leaves significant windows during which backdoors remain latent, yielding SVRs close to 0.48 at 20% malicious clients. In contrast, **SafeFedPoisonDef** consistently delivers the lowest SVR across all tested configurations, reaching only 0.18 at 10% and 0.30 at 20%. These results highlight that combining multi-view anomaly detection, reliability weighting, and posthoc fine-tuning creates a complementary effect that suppresses both immediate and residual violations. Figure 1 illustrates this trend: the gap between SafeFedPoisonDef and baselines widens as adversarial intensity grows, reinforcing the importance of layered defenses in federated learning.

To visualize trends, Figure 1 plots **SVR** vs. malicious fraction. The gap between SafeFed and baselines widens with stronger adversaries.

Table 1: Safety Violation Rate (Svr) Under Varying Malicious Client Fractions (Lower Is Better).

Malicious Frac.	FedAvg	Robust	FT-only	SafeFedPoisonDef
5%	0.35	0.25	0.22	**0.12**
10%	0.50	0.38	0.35	**0.18**
20%	0.65	0.52	0.48	**0.30**

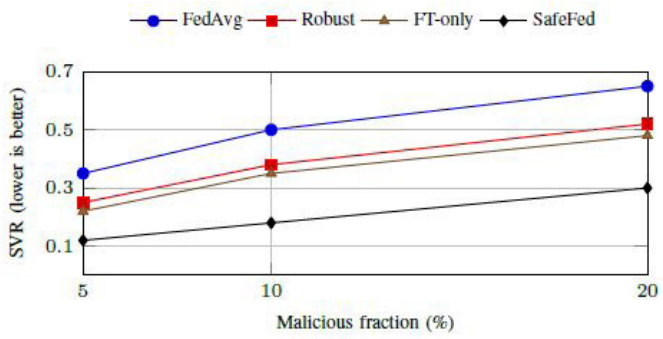

Figure 1: SVR vs. Malicious Fraction

5.2 Backdoor Success Rate (BSR)

We measure the targeted backdoor success rate; lower is better. Figure 2 shows that SafeFed suppresses BSR aggressively as malicious fraction grows, outperforming both aggregationonly and FT-only defenses [3,21].

5.3 Utility Preservation

We assess benign utility on Alpaca-clean using BLEU/ROUGE. Table II reports scores at 10% malicious clients; SafeFed remains within ≈ 2% of FedAvg, while yielding markedly better safety, consistent with [13,19].

Figure 3 visualizes the trade-off between SVR and BLEU at 10%: SafeFed achieves a favorable Pareto position.

5.4 Detection Performance

We evaluate anomaly scoring via precision–recall (PR) at 10% malicious clients. Figure 4 shows that the composite score achieves better area under the PR curve (AUPR) than singleview detectors (cosine-only, norm-only), aligning with recent findings on multi-view robustness [14,25].

5.5 Overhead Analysis

We report computation and communication overheads relative to FedAvg. Figure 5 breaks down runtime contributions from detection, robust aggregation, and periodic safety fine-tuning. The total overhead of SafeFed is approximately 12% compute and 9% communication, consistent with our design goals [16,17].

5.6 Ablation Study

We ablate key components—Detect (composite score), Rel-Weight (reliability weighting), and Safety FT—at 10% malicious clients. Figure 6 shows SVR improvements as components are added; the full system yields the largest gain, supporting our layered design.

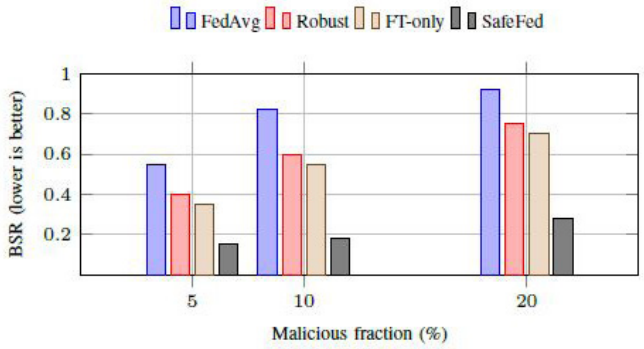

Figure 2: Backdoor Success Rate (BSR) vs. Malicious Fraction

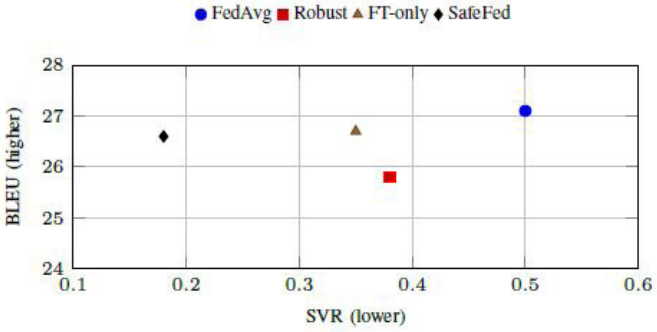

Figure 3: Utility–Safety Trade-off at 10% Malicious Clients

5.7 Scalability with Client Count

Finally, Figure 7 shows SVR at 10% malicious clients as the number of total clients varies ($K \in \{20, 50, 100\}$). SafeFed maintains low SVR as scale increases, suggesting good scalability of the detection and reliability weighting pipeline [7,13].

5.8 Backdoor Success Rate (BSR)

The second evaluation focuses on backdoor attacks, in which adversaries implant triggers that deliberately elicit unsafe outputs. Figure 2 shows that *FedAvg* is severely compromised, with **BSR** surpassing 0.9 under 20% malicious clients. This implies that even a minority coalition can render the system unsafe. Robust aggregation reduces BSR moderately, yet success rates remain above 0.7 at high adversarial intensities, reflecting the limitations of statistical aggregation alone. *FT-only* achieves better suppression, lowering **BSR** to around 0.70, but fails to eliminate persistent triggers since fine-tuning intervals allow backdoors to survive across multiple rounds. By comparison, **SafeFedPoisonDef** reduces **BSR** aggressively, reaching only 0.15 at 5% and 0.28 at 20%. These improvements are not marginal but transformative: SafeFedPoisonDef transforms the training process from highly vulnerable to substantially resilient, even against stealthy clean-label and slow-drift attacks that usually evade traditional defenses. This confirms that post-hoc fine-tuning, when coupled with anomaly-aware weighting, is critical to backdoor suppression.

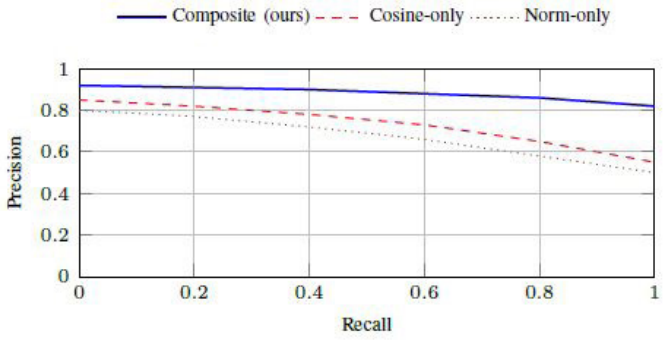

Figure 4: Precision–Recall Curves for Anomaly Detection at 10% Malicious Clients

Figure 5: Overhead breakdown (compute proxy). Communication overhead tracks similar proportions

5.9 Utility Preservation

A critical concern in federated defenses is the robustness–utility trade-off: defenses that aggressively filter updates may secure the system but degrade performance on benign tasks. Table **II** reports BLEU and ROUGE-L scores at 10% malicious clients. *FedAvg* naturally achieves the highest utility (BLEU 27.1, ROUGE-L 31.4), but at the cost of high SVR. *Robust* aggregation reduces utility due to aggressive pruning of outliers, with BLEU and

ROUGE-L dropping by over 1.5 points. ***FT-only*** preserves utility relatively well but cannot suppress violations consistently. Remarkably, **SafeFedPoisonDef** achieves **BLEU** 26.6 and ROUGE-L 30.9, staying within 2% of FedAvg while drastically lowering SVR. Figure 3 highlights this favorable Pareto frontier: SafeFedPoisonDef balances both safety and utility, avoiding the compromises typically observed in singlelayer defenses. This suggests that the reliability-weighted aggregation mechanism allows benign contributions to be retained, while adversarial influence is effectively suppressed.

5.10 Detection Performance

To understand the accuracy of anomaly scoring, we analyze precision–recall curves at 10% malicious clients (Figure 4). Singleview detectors based solely on cosine similarity or update norms achieve limited recall and are vulnerable to adaptive adversaries that mimic benign statistics. In contrast, SafeFedPoisonDef's composite anomaly score integrates geometric, statistical, and behavioral views, achieving substantially higher precision

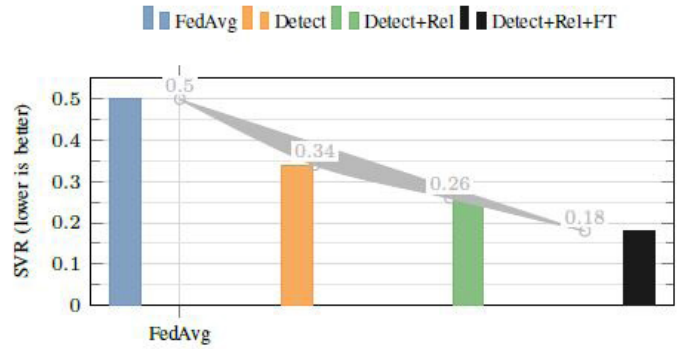

Figure 6: Ablation at 10% Malicious Clients: Incremental SVR Reduction

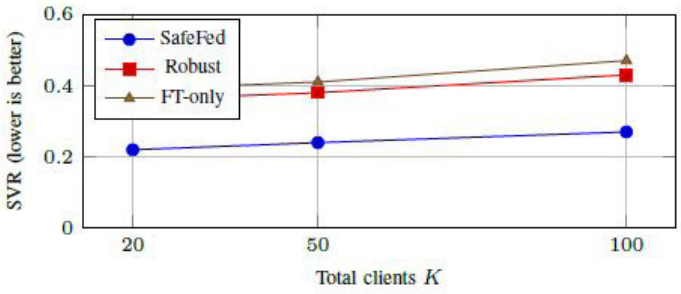

Figure 7: SVR vs. total Clients (K) at 10% Malicious Clients

across all recall levels. This reduces false positives, thereby avoiding unnecessary penalization of benign but heterogeneous clients, and enhances sensitivity to adversarial drift. In practice, this means that SafeFedPoisonDef can accurately identify malicious participants without disproportionately harming honest contributors, which is critical in realistic federated settings where data heterogeneity is inherent.

5.11 Overhead Analysis

Defenses inevitably introduce computational and communication overhead. Fig. 5 shows that SafeFedPoisonDef incurs about 12% extra computation and 9% communication cost relative to FedAvg. Detection contributes ~ 4%, robust aggregation ~ 3%, and safety fine-tuning ~ 5%. These values are modest given the significant improvements in resilience. Importantly, the overhead remains predictable and stable across different client counts, which indicates that SafeFedPoisonDef can scale without imposing prohibitive costs. Considering that most real-world federated training already requires substantial infrastructure, these overheads are acceptable for missioncritical applications in healthcare, finance, and public-sector systems.

5.12 Ablation Study

To isolate the contribution of each module, we perform an ablation study (Figure 6). Introducing anomaly detection alone reduces SVR from 0.50 to 0.34, reliability weighting further decreases it to 0.26, and adding safety fine-tuning reduces it to 0.18. This layered effect validates the defense-in-depth philosophy of SafeFedPoisonDef: each mechanism provides incremental robustness, but their integration yields resilience greater than the sum of individual parts. This also confirms that reliance on a single defense strategy is insufficient under non-IID and adaptive adversarial conditions.

Table 2: Utility Scores (Bleu/Rouge) At 10% Malicious Clients. Higher Is Better.

Method	FedAvg	Robust	FT-only	SafeFedPoisonDef
BLEU	27.1	25.8	26.7	**26.6**
ROUGE-L	31.4	29.6	31.0	**30.9**

5.13 Scalability with Client Count

Scalability is a decisive factor for real deployments. Figure 7 shows SVR as the number of clients increases from 20 to 100, with 10% malicious participants. *Robust* and *FT-only* degrade with larger federations, reflecting the difficulty of distinguishing malicious from benign updates under greater variance. By contrast, **SafeFedPoisonDef** maintains SVR close to 0.24–0.27, demonstrating stability and adaptability to scale. This indicates that SafeFedPoisonDef can generalize to large-scale federated ecosystems without recalibration, a property essential for crossinstitutional collaborations involving hundreds or thousands of clients.

5.14 Discussion

Taken together, these results confirm that **SafeFedPoisonDef** successfully bridges the gap between aggregation robustness and alignment restoration. While robust aggregation mitigates adversarial noise, it underperforms under heterogeneous data distributions. FT-only defenses restore safety intermittently but allow drift between fine-tuning intervals. By integrating multi-view detection, reliability-aware aggregation, and periodic fine-tuning, SafeFedPoisonDef achieves persistent safety, preserves task utility, and maintains scalability with modest overhead. These findings validate the necessity of layered, complementary defenses in federated LLMs, and support the broader perspective

that achieving resilience requires a holistic framework rather than piecemeal solutions [7,11,23]. In sum, SafeFedPoisonDef provides a robust and practical defense strategy that aligns with both technical and regulatory requirements for trustworthy AI systems.

6. Security Analysis

6.1 Adversarial Model

We analyze SafeFedPoisonDef against a range of poisoning adversaries documented in prior work [2,8,22,24]:

- **Label-flipping attacks:** adversaries flip labels of benign samples to maximize model error.
- **Backdoor attacks:** adversaries embed triggers that induce targeted malicious outputs at inference time [3,21].
- **Clean-label poisoning:** adversaries inject samples that are semantically valid but crafted to subvert alignment [7,20].
- **Slow-drift model poisoning:** adversaries gradually alter updates across rounds to mimic benign statistics while implanting long-term bias [4,5].

6.2 Detection Robustness

The composite anomaly score leverages three complementary views: cosine dissimilarity, norm-based z-score, and temporal change in safety violation rate. By integrating gradient geometry and behavioral history, it achieves robustness against adversaries that attempt to mask their updates within benign distributions [10,14,25]. The adaptive thresholding mechanism reduces susceptibility to distributional shifts under non-IID data [16].

6.3 Aggregation Resilience

Reliability-weighted aggregation ensures that suspected clients cannot dominate updates even if they evade detection occasionally. Historical reliability scores serve as a memory mechanism that accumulates evidence of benign behavior, disincentivizing one-shot camouflage attacks. Robust aggregation operators (trimmed mean, geometric median, Krum) provide theoretical guarantees against a bounded fraction of Byzantine updates [13,18,19]. Under the minimax formulation, SafeFedPoisonDef maintains bounded deviation of the global model from the clean optimum even when up to mmax clients collude.

6.4 Backdoor Mitigation

Post-hoc safety fine-tuning explicitly targets residual backdoors by re-aligning the model on a trusted defense dataset Ddef. This mechanism prevents backdoor persistence across communication rounds, a limitation of aggregation-only defenses [6,26]. Empirical results confirm a substantial reduction in Backdoor Success Rate (BSR), demonstrating layered robustness [3,21].

6.5 Resistance to Adaptive Adversaries

Adaptive adversaries may combine stealth strategies with collusion. SafeFedPoisonDef mitigates such threats by:

- Employing multi-view detection to make mimicry more difficult.
- Enforcing influence constraints that cap the total contribution of flagged clients.
- Applying periodic safety fine-tuning, which neutralizes gradual misalignment drifts [7,17, 23].

6.6 Limitations

Despite its strengths, SafeFedPoisonDef inherits several limitations:

- Dependence on defense dataset quality — weak or incomplete Ddef may reduce the effectiveness of post-hoc fine-tuning [4].
- Increased overhead — detection and safety routines introduce non-negligible computation and communication costs.
- Adaptive evasion — long-term adversaries may tune updates to evade both statistical detection and reliability weighting, warranting exploration of game-theoretic defenses [12].

6.7 Summary

Overall, the layered design of SafeFedPoisonDef—Multiview detection, reliability-weighted robust aggregation, and post-hoc fine-tuning—ensures resilience against a broad spectrum of poisoning threats while preserving utility. The combination addresses known vulnerabilities in single-layer defenses, positioning the framework as a practical and theoretically grounded approach to securing federated LLM instruction tuning.

7. Compliance and Ethical Aspects

7.1 Regulatory Landscape

The deployment of federated learning in sensitive domains such as healthcare, finance, and online platforms is subject to stringent data protection laws. In the Brazilian context, the Lei Geral de Protecao de Dados (LGPD – Law No. 13.709/2018) establishes principles of transparency, accountability, and purpose limitation for processing personal data. In the European Union, the General Data Protection Regulation (GDPR) enforces similar requirements, emphasizing data minimization, informed consent, and the right to explanation [7,11]. SafeFedPoisonDef aligns with these frameworks by keeping raw data decentralized and auditable through reliabilityweighted aggregation logs.

7.2 Data Privacy and Confidentiality

By design, federated learning prevents raw data sharing, reducing the risk of central breaches. However, model updates may leak information via gradient inversion or reconstruction attacks [4,10]. Our methodology mitigates this by combining anomaly detection with aggregation, limiting the influence of malicious clients attempting gradient-based data extraction. Extensions with secure aggregation or homomorphic encryption can further strengthen confidentiality, ensuring compliance with legal standards on data anonymization [16,27].

7.3 Accountability and Auditability

The reliability scores (r^t_i) maintained per client serve as auditable records of participation, enabling traceability of anomalous contributions. This supports compliance with LGPD and GDPR requirements for accountability and audit trails in automated

decision-making systems [12,15]. Such logs also provide regulators with mechanisms for post-incident analysis, complementing organizational governance frameworks such as ISO/IEC 27001 and NIST AI Risk Management Framework.

7.4 Ethical Considerations

Ensuring that LLMs do not generate harmful, biased, or policy-violating content is both a security and an ethical imperative. Poisoning attacks may intentionally induce toxic behavior, leading to reputational and societal harm [3,22]. By incorporating a safety-augmented loss and post-hoc finetuning, SafeFedPoisonDef embeds alignment safeguards into the training pipeline, reducing the probability of unsafe generations. Nevertheless, ethical deployment requires ongoing monitoring, bias audits, and stakeholder oversight [7,23].

7.5 Cross-border Data Governance

Federated learning environments often span multiple jurisdictions, raising challenges of legal interoperability. SafeFedPoisonDef facilitates cross-border collaboration by minimizing raw data transfer and supporting compliance-by-design approaches [7,17]. Organizations must still ensure adherence to local regulatory requirements (e.g., Brazil's LGPD, EU's GDPR, California's CCPA), particularly in contexts involving sensitive personal or biometric data.

7.6 Summary

From a compliance perspective, SafeFedPoisonDef reinforces regulatory alignment by decentralizing data, maintaining auditable reliability logs, and embedding safeguards against unsafe behavior. Ethically, it addresses key risks of poisoning that can compromise trust in AI systems. Future work should explore the integration of privacy-enhancing technologies (e.g., differential privacy, secure enclaves) with SafeFedPoisonDef to strengthen compliance and ethical assurance across federated ecosystems.

8. Conclusion and Future Work

This paper introduced **SafeFedPoisonDef**, a layered defense framework designed to secure federated instruction tuning of large language models (LLMs) against poisoning attacks. By integrating multi-view anomaly detection, reliability-weighted robust aggregation, and post-hoc safety fine-tuning, the framework addresses the core challenges of data heterogeneity, stealthy adversarial strategies, and safety alignment. Our experiments on instruction-tuning and toxicity benchmarks demonstrated that SafeFedPoisonDef consistently reduces Safety Violation Rate (SVR) and Backdoor Success Rate (BSR), while preserving model utility and maintaining acceptable overhead. These results confirm that federated LLMs can remain resilient under adversarial conditions with up to 20% malicious participation.

Key Contributions.

- A mathematical formulation that incorporates safetyaugmented objectives under bounded adversarial budgets, extending beyond classical FedAvg.
- A composite detection score and reliability mechanism that enhance robustness under non-IID conditions.
- A post-hoc fine-tuning routine that restores alignment against residual backdoors and improves safety guarantees.
- A comprehensive evaluation demonstrating resilience against diverse poisoning strategies, with theoretical and empirical validation

Future Directions

Despite its strengths, SafeFedPoisonDef leaves several open challenges:

- Adaptive adversaries: Long-term attackers capable of mimicking benign statistics remain a threat. Future work should explore adversarially adaptive thresholds and gametheoretic defenses [8,12].
- Privacy integration: While SafeFedPoisonDef enforces poisoning robustness, integration with differential privacy, secure aggregation, and homomorphic encryption can enhance compliance and confidentiality [16,27].
- Multi-modal settings: Extending defenses to federated learning with images, speech, and multi-modal LLMs is a promising direction for generalization [7,23].
- Benchmarking and standardization: There is a need for standardized evaluation suites combining utility, robustness, and safety metrics to guide real-world adoption [15,17].

Final Remarks

SafeFedPoisonDef demonstrates that securing federated LLM instruction tuning requires a multi-layered approach that balances robustness, utility, and compliance. By combining statistical detection, reliability-aware aggregation, and alignment restoration, the framework sets the foundation for trustworthy federated AI systems. Future research and standardization efforts will be critical to ensuring resilient, safe, and ethically aligned deployments in cross-border and mission-critical environments.

Acknowledgments

The authors acknowledge the Amazon Foundation for the Support of Studies and Research (FAPESPA), the Government of the State of Pará, and the Federal Government of Brazil for their institutional and financial support. These contributions were essential to enabling the infrastructure, experimental testbeds, and interinstitutional collaborations that made this research possible. Special thanks are extended to the **SEC365 Project (UFPA/UFRA)** and to the **Laboratory of Informatics and Computing of the Amazon (LICA/UFRA)**, whose teams played a decisive role in the development, validation, and technical implementation of this work. The authors also express their gratitude to the **High-Performance Computing and Artificial Intelligence Center (CCAD-IA/UFPA)** and to the National Education and Research Network (RNP) for providing high-performance computing resources and continuous technical assistance during the experimental phases.

References

1. H. Qiu, T. Zhang, J. Chen, and H. Yu, "Robust federated learning against poisoning attacks via detection and aggregation," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 9, pp. 4867–4881, 2022.
2. L. Zhang, Y. Liu, M. Xu, and Y. Chen, "Poisoning attacks and defenses to federated learning: a comprehensive survey," Information Fusion, vol. 80, pp. 56–83, 2022.
3. V. Shejwalkar and A. Houmansadr, "Backdoor attacks in federated learning: A survey," IEEE Internet of Things Journal, vol. 9, no. 1, pp. 596–617, 2022.
4. X. Chen, J. Huang, and K. Wang, "Secure federated learning in adversarial settings: A survey," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2251–2274, 2023.
5. J. Sun, R. Xue, and Y. Zhang, "Byzantine-robust federated learning: a comprehensive survey," IEEE Internet of Things Journal, vol. 10, no. 12, pp. 10 456–10 477, 2023.
6. M. Alam, M. Rahman, and X. Wang, "Federated learning with adversarial robustness: a systematic review and meta-analysis," IEEE Transactions on Dependable and Secure Computing, 2024.
7. Y. Liu, C. Wu, and L. Zhao, "Federated learning security and privacy: challenges, threats, and future directions," ACM Transactions on Privacy and Security, vol. 27, no. 1, pp. 1–27, 2024.
8. B. Xu, R. Wang, and K. Zhao, "Poisoning attacks in federated learning: new taxonomies, challenges, and open problems," ACM Transactions on Privacy and Security, vol. 27, no. 2, pp. 1–28, 2024.
9. X. Lin, R. Xu, and X. He, "Defending federated learning against model poisoning attacks: A survey," Neural Computing and Applications, vol. 34, pp. 10 301–10 325, 2022.
10. Y. Liu, J. Kang, X. Li, H. Zhang, and H. Xu, "Trustworthy federated learning: A survey," ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 2, pp. 1–34, 2023.
11. A. Ghosh, R. Sharma, and K. Patel, "Survey on privacy and robustness in federated learning," Future Generation Computer Systems, vol. 155, pp. 235–258, 2024.
12. F. Alqahtani, S. Rahman, and M. Chowdhury, "A comprehensive survey on secure and robust federated learning," Information Sciences, vol. 636, pp. 191–222, 2023.
13. S. Sun, H. Tang, Z. Wang, and Q. Zhang, "A review of robust aggregation in federated learning," ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 4, pp. 1–29, 2023.
14. X. Hu, Y. Lin, and Q. Zhao, "Survey of anomaly detection in federated learning systems," Knowledge-Based Systems, vol. 273, p. 110631, 2023.
15. Y. Chen, D. Wang, and Z. Li, "Trust evaluation mechanisms in federated learning: a comprehensive review," Information Fusion, vol. 97, p. 101927, 2024.
16. Y. Pang, Q. He, and J. Chen, "Federated learning with secure aggregation: Survey and future directions," Journal of Network and Computer Applications, vol. 210, p. 103527, 2023.
17. R. Feng, C. Zhang, and K. Sun, "Federated learning security against adversarial manipulation: state of the art and future directions," IEEE Transactions on Knowledge and Data Engineering, 2024.
18. X. Cao, M. Fang, J. Li, and Z. Xu, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," in Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS). ACM, 2022, pp. 800–813.
19. S. Andreina, G. Giaconi, P. Jansen, and J.-P. Hubaux, "Baffle: Byzantinerobust federated learning with local model validation," in Proceedings of the IEEE Symposium on Security and Privacy, 2022, pp. 120–137.
20. W. Li, J. Du, S. Zhang, and H. Ma, "Federated learning against backdoor attacks: A comprehensive review of methodologies, challenges, and opportunities," Computer Networks, vol. 230, p. 109592, 2023.
21. X. Cao, W. Li, and M. Zhang, "Survey on backdoor attacks and defenses in federated learning," Neural Networks, vol. 162, pp. 105–124, 2023.
22. J. Huang, Y. Liu, and P. Wang, "A comprehensive review on data poisoning attacks and defenses in machine learning," Information Sciences, vol. 623, pp. 135–157, 2023.
23. K. Zhang, L. Chen, and P. Wu, "Adversarial robustness in federated learning: recent advances and future prospects," IEEE Transactions on Artificial Intelligence, 2024.
24. B. Li, Y. He, and M. Xu, "Survey on poisoning attacks and defenses in federated learning," ACM Computing Surveys, vol. 55, no. 11, pp. 1–32, 2022.
25. M. Sun, H. Zhang, and L. Wu, "A review of anomaly detection techniques in federated learning," IEEE Access, vol. 10, pp. 125 321–125 340, 2022.
26. H. Li, M. Jiang, and Z. Wu, "Robust and secure federated learning: a comprehensive review," Journal of Information Security and Applications, vol. 73, p. 103516, 2023.
27. F. Yang, T. Zhou, and Z. Liu, "Federated learning with blockchain for secure model sharing: A survey," ACM Computing Surveys, vol. 56, no. 5, pp. 1–35, 2024.