Security and Privacy of Machine Learning, 2025 Critique G10: Security and Privacy in Federated Learning – (1) Model Poisoning Attacks to Federated Learning via Multi-Round Consistency (2) Emerging Safety Attack and Defense in Federated Instruction Tuning of Large Language Models (3) From Risk to Resilience: Towards Assessing and Mitigating the Risk of Data Reconstruction Attacks in Federated Learning

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

CRITIQUE:(1) MODEL POISONING ATTACKS TO FEDERATED LEARNING VIA MULTI-ROUND CONSISTENCY

SUMMARY

The paper studies model poisoning in Federated Learning (FL) and identifies a key failure mode of prior attacks: *self-cancellation* across rounds due to inconsistent update directions. To address this, the authors propose **PoisonedFL**, a black-box attack that enforces *multi-round consistency* by fixing a random sign vector across training rounds and adaptively scaling magnitudes to remain undetected by diverse robust aggregators. The attack requires no access to benign clients' data or updates, operates via injected fake clients, and attains strong untargeted degradation across datasets and defenses. Empirically, PoisonedFL outperforms several canonical poisoning baselines under many defenses, suggesting that temporal-aware adversaries can reliably degrade FL models despite classic robust aggregation (e.g., Krum, Median/Trimmed-Mean) and trust-bootstrapping ideas such as FLTrust [1]–[5].

STRENGTHS

- **Clear conceptual advance.** The shift from per-round to *temporal* (multi-round) attack design is simple yet powerful. Fixing signs and adjusting magnitudes captures a previously under-explored axis of vulnerability (temporal accumulation) that is orthogonal to single-round robust aggregation [2].
- **Realistic threat model (cross-device).** The attack does not assume insider access to benign updates; it needs only the broadcast global model and injected (fake) clients—a plausible setting for cross-device FL deployments [1], [4].
- **Defense-agnostic adaptation.** The magnitude adaptation loop exploits round-to-round global-model feedback to remain within aggregator tolerances while maintaining directional drift—a clever way to defeat dimension-wise robust rules (Median/Trimmed-Mean) and even selection-based rules (Krum).
- **Comprehensive evaluation.** The paper spans multiple datasets/architectures and contrasts against both per-round poisoning and fake-client baselines (e.g., MPAF [4]), with meaningful ablations on participation, non-IIDness, and local epochs.
- **Actionable insight for defenders.** The results convincingly motivate *temporally aware* defenses rather than purely per-round filters.

WEAKNESSES

- **Limited cross-silo practicality.** The core assumption—ability to inject many fake clients—is weak in authen-

ticated cross-silo FL (e.g., hospitals, banks). The paper acknowledges this but leaves a sizable gap between cross-device and cross-silo realities.
- **Theory lags practice.** While the empirical story is strong, there is no formal analysis of (i) how much cumulative drift survives under each robust aggregator, (ii) convergence/divergence conditions with step-size/momentum schedules, or (iii) bounds under client subsampling. Such results would clarify when temporal poisoning *must* succeed.
- **Secure aggregation & privacy noise.** Many FL systems employ secure aggregation and/or DP noise. The interaction between multi-round consistency and noisy/obfuscated global updates is not deeply characterized; DP noise and learning-rate decay could materially change the attack's signal-to-noise ratio.
- **Detector arms race not fully explored.** The paper sketches detection ideas, but a systematic evaluation of *temporal* detectors (e.g., change-point tests on signed per-coordinate drift, spectral drift monitors) is limited relative to the breadth of attack experiments.
- **Untargeted focus.** The work centers on untargeted degradation. Targeted/backdoor objectives against robust aggregators—while harder—are highly relevant in practice [5].

Potential Improvements / Extensions

- **Temporal-robust aggregation.** Propose and evaluate aggregators that constrain *cumulative* signed drift: e.g., per-coordinate *total variation* caps across rounds, temporal median-of-means on signed gradients, or momentum-aware clipping that attenuates coordinates exhibiting persistent one-sided push.
- **Change-point & spectral detectors.** Build detectors over the time series of global updates: (i) CUSUM/GLRT on signed coordinates; (ii) low-rank/spectral tests for emerging principal directions; (iii) per-layer drift budget accounting. Compare to per-round detectors to quantify temporal lift.
- **Formal guarantees.** Provide bounds on the attacker's achievable drift under (Multi-)Krum, Trimmed-Mean, and FLTrust as functions of attacker fraction, client sampling rate, and clip thresholds [2], [3].
- **Noise-resilient variants.** Analyze/evaluate PoisonedFL under DP-SGD (server-side noise), strong weight decay, and cosine/step LR schedules. Characterize the minimal attacker fraction to overcome these dampers.
- **Targeted objectives.** Extend the fixed-sign framework to targeted/backdoor settings via two-phase control: (i) consistent untargeted drift to erode robustness; (ii) low-amplitude task-aligned drift to embed triggers [5].
- **Cross-silo threat models.** Explore constrained-adversary regimes: few malicious clients with authenticated identities, auditing, and heterogeneous compute; quantify the residual risk and required monitoring.

Questions for the Authors

1) How does PoisonedFL scale when server injects small zero-mean noise (DP-SGD) or enforces per-layer norm/momentum clipping over *cumulative* windows rather than per-round?
2) Can you prove lower/upper bounds on cumulative signed drift under Krum/Trimmed-Mean as a function of attacker rate and client subsampling?
3) What are the most effective *temporal* detection statistics you encountered, and how did the attack adapt (e.g., sign dithering, coordinate dropout) without losing potency?
4) Under secure aggregation, where only the *sum* of updates is visible, do your magnitude-adaptation heuristics still work reliably, and how sensitive are they to LR schedules?
5) Could a *randomized* server-side sign-scrambling or coordinate re-basing (e.g., periodic orthogonal transforms announced to clients) disrupt multi-round consistency without harming benign learning?

Critique: (2) Emerging Safety Attack and Defense in Federated Instruction Tuning of Large Language Models

Summary

The paper studies safety risks in Federated Instruction Tuning (FedIT) of LLMs. It introduces a *stealthy* data-poisoning attack wherein a subset of clients fine-tune locally on *unaligned* (harmful instruction, harmful response) pairs and then submit adapters/weights that—after standard FedAvg aggregation [6]—degrade the global model's safety while preserving general helpfulness. The work argues that robust aggregation defenses such as Krum [7] underperform because malicious and benign updates share similar *instruction-following* optimization geometry, making parameter-space outlier detection ineffective. As a remedy, the authors propose a server-side, *post-hoc* defense: automatically generate a small corpus of (i) normal instructions with helpful responses and (ii) harmful instructions with harmless responses, then do a brief fine-tuning pass to re-align the aggregated model. Using OpenFedLLM infrastructure [8], the attack reduces safety (e.g., on AdvBench [9]) by up to $\sim$70% while the post-hoc defense recovers up to $\sim$69% without significant loss on MT-Bench helpfulness [10].

Strengths

- **Clear problem framing for FedIT safety.** The paper pinpoints a realistic gap: FedIT inherits alignment risks that differ from classical label-flip poisoning and are not well-captured by model-space robust aggregation [7].
- **Low-cost, scalable attack surface.** Two practical attack avenues are articulated: mining unsafe pairs from public alignment datasets and synthetic generation via off-the-shelf LLMs. Both lower the barrier to adversarial participation and reflect *dual-use* risks.
- **Elegant, deployment-friendly defense.** The post-hoc, data-centric realignment is simple, model-agnostic, and

plug-and-play: it avoids brittle attacker-count assumptions and works atop diverse FedIT baselines.
- **Comprehensive evaluation.** The study spans multiple benign/malicious datasets, client scales, and off-the-shelf generators; it separates safety (AdvBench, LLM/reward-model judges) from helpfulness (MT-Bench), demonstrating the stealthy nature of safety drift.

Weaknesses

- **Threat-model narrowness.** The attack presumes data-poisoning via unsafe pairs; stronger adversaries could mix objective poisoning (e.g., RL loss shaping) or gradient surgery to cancel post-hoc gradients, challenging the proposed fix beyond data-level misalignment.
- **Evaluator circularity and overfitting risk.** Safety is judged partly by LLM-based classifiers/reward models; post-hoc training on generator-produced aligned data could overfit to evaluator artifacts rather than truly increase refusal robustness (domain shift from real red-teaming).
- **Aggregation obliviousness.** Conclusions are drawn mainly under FedAvg [6]; the failure analysis of Krum [7] is insightful but other principled aggregators (e.g., coordinate-wise filtering with *safety-aware* constraints) are unexplored.
- **Self-alignment stability.** The Level-3 "self-healing" defense uses the compromised global model to generate its own antidote. Without guarantees, this can entrench failure modes (mode collapse, jailbreak-preserving rephrasings) if the generator itself is drifted.
- **Limited modality and scale.** Results focus on a single base series and text-only tasks; cross-modality (tools, code-execution, agents) and larger foundation models may exhibit different attack/defense dynamics.

Potential Improvements & Extensions

- **Safety-constrained aggregation.** Replace purely geometric defenses with constrained optimization at the server: project aggregated updates onto a *safety-feasible* cone inferred by *safety probes* (e.g., small canary sets scored by rule-based/causal checks), rather than relying on outlier distance alone.
- **Round-level auditing signals.** Introduce per-round *safety influence* scores via influence functions or Shapley approximations to detect clients whose updates disproportionately increase harmful likelihood, even if cosine-similar overall.
- **Adversarially-robust post-hoc.** Train the post-hoc defense with a minimax objective where an inner loop adversary crafts jailbreak prompts (or perturbations) against the current model; couple this with *selective generation* that prioritizes hard negatives (safety boundary mining).
- **Defense-data diversity controls.** Deduplicate prompt families, enforce lexical/semantic diversity, and add paraphrase/hard-negative curricula to mitigate overfitting to a generator's style. Consider cross-model ensembles to avoid generator bias.
- **Privacy & FL realism.** Analyze interplay with secure aggregation and client DP: (i) does DP-SGD noise blunt the attack or the post-hoc cure? (ii) how do participation rates, non-IID skews, and partial client dropouts affect attack potency and defense efficacy?
- **Causal and mechanistic evaluations.** Beyond AdvBench, use causal refusal tests (e.g., safety token ablations, neuron-level interventions) to verify that safety behavior stems from robust mechanisms rather than prompt-surface heuristics.

Questions for the Authors

- How does the defense perform under *adaptive* attackers that (a) target the evaluator distribution, or (b) craft gradients to null the expected post-hoc correction?
- Can server-side *safety probes* (tiny private canary sets never seen by clients) detect safety drift earlier and reduce reliance on large post-hoc updates?
- What is the trade-off frontier between defense sample size, steps, and helpfulness retention across larger models and multilingual settings?
- Does secure aggregation or client-level DP materially change the stealthiness finding (e.g., by masking or amplifying harmful directions)?
- Could a safety-constrained FedAvg (projection or clipping in directions most correlated with harmful response likelihood) close some of the gap to post-hoc training?

Critique: (3) From Risk to Resilience: Towards Assessing and Mitigating the Risk of Data Reconstruction Attacks in Federated Learning

I. Summary

The paper introduces **Invertibility Loss (InvLoss)** as a principled, instance-aware feasibility metric for *data reconstruction attacks* (DRAs) in federated learning (FL). By analyzing the Jacobian of the shared representation/gradient mapping, the authors derive *upper bounds* that connect spectral properties (singular values and gaps) to how easily inputs can be reconstructed. Building on this, they propose **InvRE**, a practical estimator of InvLoss that (i) is model-agnostic (usable across HFL/VFL and architectures), (ii) scales to per-instance auditing, and (iii) empirically correlates with reconstruction quality metrics (MSE, PSNR, SSIM) under several attacks (e.g., DLG/IG/CGIR/PISTE). Finally, they design **adaptive noise perturbation** defenses that inject noise *selectively* in Jacobian subspaces most influential for inversion, improving privacy–utility trade-offs relative to uniform noise or coarse compression baselines. Experiments span multiple datasets and architectures, showing strong alignment between InvRE scores and actual attack success ([6], [11]–[14]).

II. Strengths

- **Unifying lens for HFL/VFL.** Casting DRAs through local invertibility and Jacobian spectrum provides a clean,

architecture-agnostic bridge across FL variants and attack families, complementing earlier empirical work on leakage [12], [13] and classical privacy analysis [11].
- **Actionable risk auditing.** Instance-wise InvRE offers operational value: clients/systems can pre-screen samples and modulate protection (e.g., selective noise budgets), rather than applying one-size-fits-all defenses.
- **Defense that matches the threat model.** Spectral, subspace-aware noise targets what inversion exploits, avoiding unnecessary utility damage—a theoretically motivated alternative to uniform DP noise [14].
- **Thorough experimental design.** Cross-architecture (LeNet/AlexNet/ResNet; ResNet cut-points), cross-dataset, and multi-metric evaluation builds confidence that InvRE tracks attack efficacy, not a specific pipeline quirk.
- **Clear pre-deployment guidance.** The reported empirical thresholds (low/medium/high risk regions by InvRE) make the estimator directly usable for triage policies.

III. Weaknesses

- **Local linearization limits.** The theory leans on first-order approximations around the point x. For highly non-linear or non-smooth regions (e.g., ReLU kinks, attention composition), higher-order effects could materially affect invertibility yet are not quantified.
- **Narrow modality/task scope.** The focus is image classification features/gradients with 32×32 inputs. Generality to NLP, speech, high-res imaging, or structured/tabular FL is claimed but untested; Jacobian spectra (and feasible k) can differ drastically.
- **Threat model coverage.** The strongest results assume an honest-but-curious server. Active adversaries (gradient shaping, auxiliary priors, higher-order matching, prompt/label manipulations) may partially sidestep the spectral defenses.
- **System overhead and deployment friction.** Computing per-sample Jacobians and SVDs introduces non-trivial cost. While truncated SVD helps, per-round, per-client overhead in at-scale FL (with partial participation and stragglers) may still be significant.
- **Privacy accounting and composition.** The adaptive-noise approach improves trade-offs but lacks rigorous privacy accounting (e.g., (ε, δ)-DP) or composition under many rounds—making cross-study comparisons to DP-SGD [14] difficult.

IV. Potential Improvements / Extensions

- **Beyond first order.** Incorporate curvature-aware (Gauss–Newton or Hessian-vector) corrections or Lipschitz bounds to capture non-linear invertibility—even if only to tighten/loosen InvLoss bounds where ReLU/attention introduce non-differentiabilities.
- **Hybrid defenses with DP.** Calibrate adaptive spectral noise under a DP accountant, yielding *certified* budgets while preserving subspace targeting. Compare end-to-end with DP-SGD and privacy amplification by secure aggregation.
- **Broader tasks and modalities.** Validate InvRE on large-token NLP FL (language modeling, next-word prediction), medical imaging with higher resolutions, and tabular VFL; analyze how tokenization and transformer blocks reshape the Jacobian spectrum.
- **Adversary adaptivity.** Stress-test InvRE and defenses against attackers that (i) optimize in higher-order spaces, (ii) use learned priors (diffusion/score matching) informed by public data, or (iii) conduct *gradient surgery* to increase effective rank.
- **Client heterogeneity and fairness.** Study how non-iid distributions, label skew, and minority subpopulations affect InvRE—and whether adaptive noise disproportionately harms accuracy for certain clients/classes.
- **Operational policies.** Turn InvRE thresholds into decision rules: risk-aware client sampling, curriculum pacing (delay high-risk samples), or selective secure channels (encrypt/clip only when InvRE exceeds a budget).

V. Questions for the Authors

- **Tightness of bounds:** In practice, how tight is the InvLoss upper bound versus realized attack error when the attacker uses higher-order matching or learned generative priors?
- **Compositional privacy:** Can adaptive spectral noise be paired with a per-round DP accountant to yield meaningful (ε, δ) over many FL rounds without losing the utility gains?
- **Robustness to non-iid & partial participation:** How stable are InvRE rankings across clients with heterogeneous data? Does client sampling randomness undermine instance-level auditing?
- **Interplay with secure aggregation:** If gradients/embeddings are protected via secure aggregation, can InvRE still guide *client-side* adaptive defenses effectively?
- **VFL practicality:** In multi-party VFL with different owners of feature shards, who computes InvRE and how is the Jacobian decomposed without additional leakage or coordination overhead?

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. PMLR, vol. 54, 2017, pp. 1273–1282.

[2] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] G. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[4] X. Cao and N. Z. Gong, "MPAF: Model poisoning attacks to federated learning based on fake clients," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[5] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. PMLR, 2020.

[6] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of AISTATS*. PMLR, 2017, pp. 1273–1282.

[7] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine-tolerant gradient descent," in *Advances in Neural Information Processing Systems*, 2017.

[8] R. Ye, W. Wang, J. Chai, D. Li, Z. Li, Y. Xu, Y. Du, Y. Wang, and S. Chen, "Openfedllm: Training large language models on decentralized private data via federated learning," *arXiv preprint arXiv:2402.06954*, 2024.

[9] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[10] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Advances in Neural Information Processing Systems*, 2024.

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[12] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *arXiv preprint arXiv:1906.08935*, 2019.

[13] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients – how easy is it to break privacy in federated learning?" in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[14] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318.