WHO INFLUENCED THIS BEHAVIOR? AUDITING FEDERATED LLMS WITHOUT REVEALING CLIENT UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated parameter-efficient fine-tuning (PEFT) can adapt language models to decentralized data while secure aggregation hides individual dense numerical learning updates. This confidentiality creates a post-hoc accountability problem: given an observed model behavior, how can an auditor rank the clients whose updates locally supported or opposed it? We introduce AUDITRACE, a behavior-conditioned client-ranking method based on magnitude-free audit records. At selected checkpoints, each client locally records only the coordinate indices and signs of its Top-K trainable-update entries; the dense update can remain on an independent secure-aggregation learning path. Given a differentiable audit objective, the auditor constructs an analogous query-gradient record and aggregates FedAvg-weighted signed overlap across checkpoints, yielding a directional ranking without reconstructing individual dense updates. We characterize exact mask-space geometry, derive conditional approximation bounds, and formalize a hidden-tail failure mode. Across diverse model, task, and partition settings, AUDITRACE reaches dense-reference Spearman correlation 0.839–0.952. Against exact leave-one-client-out retraining it attains Spearman 0.827 and positive-effect AUROC 0.941 from coordinate identities and signs alone, within 0.020 Spearman of dense cosine. A 5% signature retains Spearman 0.860 with a 415.4 KiB audit record; repeated-sign membership inference peaks at AUROC 0.567 versus 0.5 chance. AUDITRACE thus enables post-hoc auditing in federated PEFT while leaving individual dense updates on the secure-aggregation path.

1 INTRODUCTION

Large language models increasingly require adaptation to decentralized data that cannot be pooled centrally. Federated learning (FL) keeps raw examples at their sources, while parameter-efficient fine-tuning (PEFT), including LoRA, reduces the trainable and communicated state (McMahan et al., 2017; Hu et al., 2022; Zhang et al., 2023; Bai et al., 2024; Yan et al., 2025). Yet an aggregate model can acquire an unsafe, incorrect, or otherwise unexpected behavior even when aggregate validation remains acceptable (Bagdasaryan et al., 2020; Wang et al., 2020a; Ye et al., 2025).

When a federated model exhibits an unexpected behavior, a natural question is: **which clients influenced this behavior?** We study *post-hoc behavior-conditioned client attribution*. Given audited model checkpoints, client audit records, and a query set defining an oriented differentiable objective, the task is to rank clients by whether their updates locally support or oppose increasing that objective. The output is a signed ranking: it is not a reconstruction of client updates, a proof of training-example provenance, or a determination of malicious intent. Behavior conditioning is essential because the same client can support one queried behavior and oppose another.

This attribution objective creates a tension with the confidentiality mechanisms used in federated learning. Individual gradients and parameter updates can leak sensitive training examples, including text sequences (Zhu et al., 2019; Geiping et al., 2020; Balunovic et al., 2022; Petrov et al., 2024). Federated systems therefore increasingly rely on secure aggregation and related privacy mechanisms to prevent exposure of individual client updates. Under its stated corruption threshold, secure aggregation reveals the aggregate update needed for learning while withholding the individual dense updates (Bonawitz et al., 2017). This hides individual dense updates, but it also removes the client-level numerical signals required by attribution methods that consume updates, gradients, curvature

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

statistics, or client models (Koh & Liang, 2017; Pruthi et al., 2020; Park et al., 2023; Kwon et al., 2024). The resulting challenge is therefore not simply to perform attribution under secure aggregation, but to recover enough client-specific evidence for post-hoc auditing without re-exposing the dense numerical updates that secure aggregation was intended to hide. Any such additional client-linked evidence necessarily creates a separate disclosure surface that must be accounted for.

Existing approaches address related attribution and auditing problems, but not post-hoc behavior-conditioned attribution when individual dense updates remain hidden. TraceFL and ProToken provide prediction- and token-level client provenance, respectively (Gill et al., 2025; 2026), while FedAttr associates clients with pre-watermarked data using paired secure-aggregation subset queries and estimated updates (Zhang et al., 2026).

To address this auditability challenge, we introduce AUDITRACE, a *magnitude-free* attribution framework with separate learning and audit planes. Dense client updates can remain on the secure-aggregation learning path, while each audited checkpoint exposes only the indices and signs of the K most salient trainable-update coordinates, with $K \ll M$. For a post-hoc behavior query, the auditor defines a differentiable objective on the global model, constructs the corresponding sparse signed query-gradient record, and computes FedAvg-weighted signed client–query overlap across checkpoints and queries. Unlike the approaches above, AUDITRACE distinguishes supporting from opposing client effects and does not require recovery or observation of an individual dense update.

To our knowledge, AUDITRACE is the first *magnitude-free, behavior-conditioned client attribution framework for federated PEFT* under an interface where individual dense numerical updates remain hidden. We develop its theoretical foundations, validate it against dense and counterfactual references, and characterize audit-record leakage and integrity requirements. Our contributions are threefold:

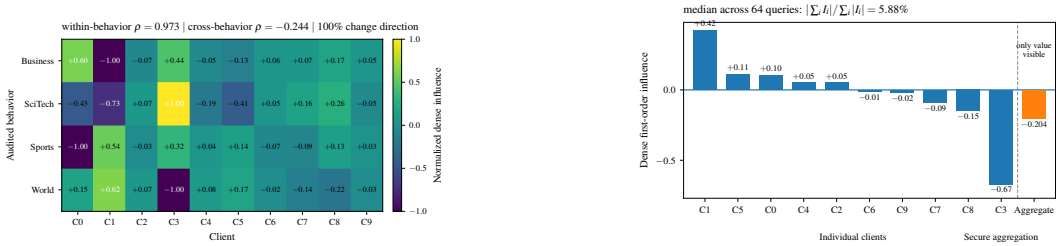
- **A magnitude-free attribution interface with formal characterization.** We develop a sparse signed audit representation that avoids exposing numerical client-update magnitudes, characterize its exact signed-mask geometry, derive sufficient conditions relating the sparse score to dense first-order contributions, and formalize the hidden-tail ambiguity inherent to retained-support methods.
- **Counterfactual validation and broad empirical robustness.** Against five-seed exact leave-one-client-out retraining, AUDITRACE achieves $\rho_{\mathrm{S}} = 0.827$ and positive-effect AUROC 0.941. We evaluate it across models, datasets, client heterogeneity and scale, adapters, and LoRA ranks, with dense-reference Spearman correlation ranging from 0.839 to 0.952 across the primary settings.
- **Practical attribution with explicit utility–exposure trade-offs.** A 5% signature retains Spearman correlation 0.860 with a 415.4 KiB serialized audit record, while held-out top-five removal improves the unsafe-target objective by 0.0232 versus 0.0230 for dense-oracle and 0.0129 for random removal. We also quantify audit-record disclosure: the strongest repeated-sign membership attack reaches AUROC 0.567 versus 0.5 chance, while the evaluated client-property attack does not reliably separate from chance.

2 Problem Setting, Motivation, and Related Work

2.1 Federated PEFT and the Auditability Gap

Parameter-efficient fine-tuning (PEFT) adapts a pretrained model by optimizing a shared trainable parameter vector $\theta_t \in \mathbb{R}^M$, making methods such as LoRA suitable for federated adaptation (Hu et al., 2022; Zhang et al., 2023; Bai et al., 2024; Yan et al., 2025). At round t, participating client $i \in \mathcal{C}_t$ computes an update $\Delta_{i,t}$ from the common checkpoint θ_t. FedAvg applies nonnegative weights $a_{i,t}$ with $\sum_{i\in\mathcal{C}_t} a_{i,t} = 1$: $\theta_{t+1} = \theta_t + \sum_{i\in\mathcal{C}_t} a_{i,t}\Delta_{i,t}$.

All clients in our experiments use the same ordered PEFT parameter layout. Individual updates may reveal properties or examples from local training data (Zhu et al., 2019; Geiping et al., 2020; Balunovic et al., 2022; Petrov et al., 2024). Under its stated corruption threshold, secure aggregation reveals the weighted aggregate needed by FedAvg update without revealing each $\Delta_{i,t}$ (Bonawitz et al., 2017). The aggregate alone does not identify its client-wise decomposition. Any method that restores client-level auditability must therefore introduce and account for an additional source of client-specific information.



(a) Behavior-conditioned client influence.

(b) Client identifiability under secure aggregation.

Figure 1: **Why behavior-conditioned auditing needs client-specific evidence. (a)** Client rankings are stable within a behavior but change across behaviors. **(b)** Secure aggregation exposes only the net first-order effect, which can conceal large opposing client effects. These results motivate a separate audit record; they do not imply that the record is privacy-preserving.

2.2 Motivation: Behavior-Conditioned Attribution under Secure Aggregation

Figure 1 uses a separate 64-query diagnostic cohort to illustrate the two requirements in our task, based on federated LoRA fine-tuning of Llama-3.2-1B on AG News with 10 clients and Dirichlet non-IID partitioning ($\alpha = 0.3$). This diagnostic cohort is not the 100-query AG News evaluation cohort used in the later attribution tables. For query q and checkpoint t, the instrumented dense first-order reference is $I_{i,t}(q) = a_{i,t} \langle \Delta_{i,t}, g_{q,t} \rangle$; individual updates are retained only for this offline evaluation and are not available to the proposed auditor.

Behavior conditioning is necessary. Figure 1a shows that client rankings are highly stable across queries representing the same behavior (mean Spearman $\rho_S = 0.973$) but differ across behaviors (mean $\rho_S = -0.244$). Every client has both positive and negative behavior-level effects, with a mean rank span of 5.8 positions among ten clients. A single behavior-agnostic client score therefore cannot explain a specific observed failure.

The secure aggregate is insufficient. Figure 1b shows the complementary gap. Dense instrumentation provides the decomposition $\{I_i(q)\}_i$, whereas the learning protocol reveals only $\sum_i I_i(q)$. Across 64 queries, the cancellation ratio $|\sum_i I_i(q)| / \sum_i |I_i(q)|$ has median 5.88% (IQR 2.64–14.23%), versus 37.7% under an exact magnitude-preserving random-sign null. Thus, substantial opposing client effects can be hidden by their aggregate. AUDITRACE addresses this information gap through a separate, explicitly disclosed audit record.

2.3 Related Work

Influence and data attribution. Influence Functions, TracIn, TRAK, and DataInf estimate training-example effects using numerical gradients, updates, projected features, or curvature (Koh & Liang, 2017; Pruthi et al., 2020; Park et al., 2023; Kwon et al., 2024). LESS, RapidIn, and GPTfluence adapt related ideas to language-model training (Xia et al., 2024; Lin et al., 2024; Chai et al., 2024). These methods provide valuable dense references and baselines, but their native interfaces assume access to numerical training information unavailable when individual client updates remain hidden.

Federated attribution and provenance. Federated valuation methods estimate client utility, often through counterfactual coalitions or retraining (Ghorbani & Zou, 2019; Song et al., 2019; Wang et al., 2020b; Liu et al., 2022; Chen et al., 2023). TraceFL localizes clients responsible for a prediction using neuron provenance (Gill et al., 2025), while ProToken provides token-level provenance for federated LLM generation (Gill et al., 2026). FedAttr identifies clients associated with pre-watermarked data by estimating individual updates through paired secure-aggregation subset queries and provides a formal leakage analysis (Zhang et al., 2026).

Secure defenses and federated auditing. Robust aggregation and secure-aggregation-compatible defenses suppress, constrain, or verify client updates during training (Blanchard et al., 2017; Yin et al., 2018; Nguyen et al., 2022; Lycklama et al., 2023; Roy Chowdhury et al., 2022; Bell et al., 2023), while related auditing and security methods study support-based screening, poisoning, faulty clients, provenance, or influential training signals (Dogani et al., 2026; Ye et al., 2025; Li et al., 2024; Jia et al., 2026; Gill et al., 2023; Kwon et al., 2024). These approaches do not directly provide a

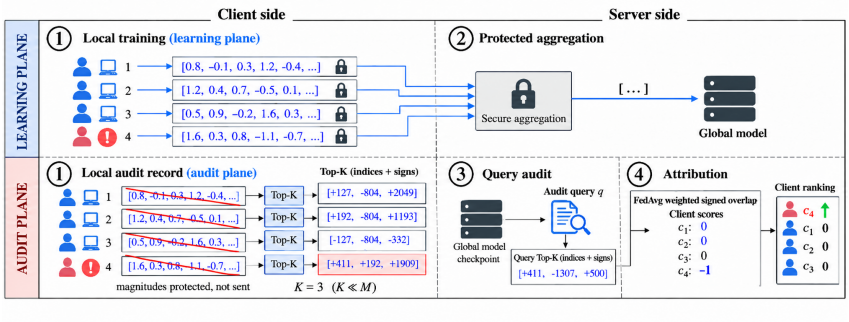


Figure 2: **AUDITRACE overview.** Dense updates remain on the learning path, while clients separately expose sparse signed audit records. The auditor matches these records to a behavior-conditioned query-gradient record and aggregates the resulting client scores.

post-hoc ranking of which clients supported or opposed a user-specified behavior while their dense numerical updates remain hidden.

Positioning. AUDITRACE addresses this gap by conditioning a sparse signed client record on a post-hoc query gradient, producing supporting and opposing client rankings from coordinate identities and one direction bit without reconstructing an individual dense update. This differs from query-independent poisoning defenses, provenance tied to particular predictions or predefined signals, and attribution interfaces that require numerical client information or counterfactual retraining. The cost is a client-linked audit record: secure-aggregation compatibility applies to the dense learning path, not to confidentiality of the audit record itself. We therefore treat its exposure—controlled by κ and audit frequency—as a design parameter and evaluate its leakage separately (Appx. F).

2.4 Threat Model

We consider a potentially curious or malicious learning server, secure-aggregation parties, a trusted auditor, and clients that may misreport their audit records. We distinguish two integrity surfaces. *Learning-plane integrity* concerns whether a validated private input is the same value entering secure aggregation; verifiable-input systems such as RoFL, EIFFeL, and ACORN address this property, but their integration is outside our scope. Our focus is audit-record consistency: the zero-knowledge relation verifies that the signed Top-K record matches a committed hidden update without revealing unreported values. It does not bind that commitment to the SecAgg input. The audit record is not differentially private; its leakage is evaluated separately.

3 Method

AUDITRACE produces a post-hoc signed ranking of clients with respect to an explicit audit objective. Dense client updates remain on the learning path, which may use secure aggregation, while selected checkpoints produce a separate audit record containing only sparse coordinate identities and signs. After a behavior query is specified, the auditor constructs a corresponding query-gradient record, compares it with the stored client records, and aggregates the resulting scores across checkpoints and queries. Figure 2 shows the two information paths and the ranking output.

3.1 Behavior-Conditioned Attribution Objective

We study client attribution with respect to a specific model behavior. An audit query q specifies a differentiable scalar objective $J_q(\theta)$, oriented before attribution so that a larger value means more of the behavior being audited. At checkpoint θ_t, define $g_{q,t} = \nabla_\theta J_q(\theta_t)$. For a prompt–response pair (x, y), one possible objective is $J_q(\theta) = \log p_\theta(y \mid x)$. Task-loss audits instead use the loss itself. Every experiment states the objective and orientation before computing metrics; estimator signs are not selected after observing evaluation results. A behavior may be represented by a prespecified set $\mathcal{Q} = \{q_1, \ldots, q_Q\}$ with nonnegative weights ν_q summing to one. The auditor has access to global checkpoints and client audit records, but not individual dense updates. The output is a signed client

ranking: positive scores indicate local alignment with increasing J_q, and negative scores alignment with decreasing it. The score is a ranking signal rather than a calibrated causal contribution and does not determine client intent or training-example provenance.

3.2 AUDITRACE: MAGNITUDE-FREE SIGNATURES AND SIGNED ATTRIBUTION

We first consider the attribution signal available if individual client updates were observable. Following first-order gradient-alignment methods such as TracIn and LESS (Pruthi et al., 2020; Xia et al., 2024), the local effect of client i on query q at checkpoint t would be

$$I_{i,t}(q) = a_{i,t} \langle \Delta_{i,t}, g_{q,t} \rangle . \tag{1}$$

where $a_{i,t}$ is the FedAvg coefficient. For small updates, positive alignment locally increases the audited objective and negative alignment decreases it. Computing it directly requires the numerical client update $\Delta_{i,t}$, which is unavailable under secure aggregation. AUDITRACE approximates this directional relationship with a magnitude-free representation, based on the intuition that related optimization signals concentrate on overlapping salient coordinates. It projects the client update and query gradient into the same sparse Top-K coordinate space and uses their signs to distinguish aligned from opposing effects, while discarding numerical magnitudes.

Client-side extraction. For a shared ordered parameter layout of size M, client i forms

$$\mathcal{S}_{i,t} = \mathrm{TopK}(|\Delta_{i,t}|; K_i), \qquad [z_{i,t}]_j = \mathrm{sgn}(\Delta_{i,t,j}) \mathbf{1}\{j \in \mathcal{S}_{i,t}\}, \tag{2}$$

with equal-magnitude ties resolved by ascending global coordinate index. The serialized audit record is $\mathcal{Z}_{i,t} = \{(j, [z_{i,t}]_j) : j \in \mathcal{S}_{i,t}\}$. The record contains coordinate identities and signs, but no numerical magnitudes or ranks. Thus, *magnitude-free* refers only to the audit copy: the exposed indices and signs remain informative metadata, while the dense update still follows the learning path.

Coordinate alignment. As in standard homogeneous FL, clients within a run share the same ordered trainable parameterization (model revision, adapter configuration, tensor shapes, and flattening order), whose layout hash is recorded. For LoRA, we also fix initialization and parameterization because the representation is not invariant to $(A, B) \mapsto (cA, B/c)$. Different adapter families are handled by instantiating their corresponding trainable coordinate spaces (See Appx. E.4).

Query record and round score. Given an audit objective, the auditor compares its sparse signed gradient record with each client's stored audit record. At checkpoint t, the auditor forms $\mathcal{S}_{q,t} = \mathrm{TopK}(|g_{q,t}|; K_q)$ and $[z_{q,t}]_j = \mathrm{sgn}(g_{q,t,j}) \mathbf{1}\{j \in \mathcal{S}_{q,t}\}$. The normalized signed overlap and the primary weighted score are

$$u_{i,t}(q) = \frac{\langle z_{i,t}, z_{q,t} \rangle}{\sqrt{K_i K_q}}, \qquad s_{i,t}(q) = a_{i,t} u_{i,t}(q). \tag{3}$$

An overlapping coordinate contributes positively when the signs agree and negatively when they disagree; non-overlapping coordinates contribute zero. For equal budgets $K_i = K_q = K$, $u_{i,t} \in [-1, 1]$. Unless stated otherwise, all experiments use the weighted score $s_{i,t}$.

Temporal and behavior aggregation. For checkpoint weights ω_t, define

$$S_i(q) = \frac{\sum_{t \in \mathcal{T}} \omega_t \mathbf{1}\{i \in \mathcal{C}_t\} s_{i,t}(q)}{\sum_{t \in \mathcal{T}} \omega_t \mathbf{1}\{i \in \mathcal{C}_t\}}, \qquad \bar{S}_i(\mathcal{Q}) = \sum_{q \in \mathcal{Q}} \nu_q S_i(q). \tag{4}$$

We use uniform checkpoint and query weights in experiments. Clients are ranked by $\bar{S}_i(\mathcal{Q})$; for an undesirable objective, larger positive scores indicate stronger alignment with increasing that behavior.

3.3 PRACTICAL DESIGN, FIDELITY, AND SAFEGUARDS

Signature budget and cost. The signature fraction $\kappa = K/M$ controls the accuracy–communication–exposure trade-off; headline results use $\kappa = 0.10$, with other fixed budgets reported as sensitivity analyses. Each client sends one signature per audited round, independent of the number of later queries. A fixed-width signed-index encoding requires $B_{\mathrm{audit}} = K(\lceil \log_2 M \rceil + 1)$ bits, excluding protocol overhead. Signature extraction requires an $O(M)$ scan plus Top-K selection, and matching costs $O(K_i + K_q)$ with sorted indices or hashing. Measured payload, extraction, matching,

storage, and attribution memory are reported where available; rank-weighted, layer-normalized, and adaptive-budget variants are appendix diagnostics rather than proposed alternatives.

Information exposure and privacy hardening. The audit signature reveals selected coordinate identities and signs but not update magnitudes or unselected coordinates, and therefore provides no formal privacy guarantee. Empirically, membership attacks remain close to chance (support-only AUROC ≈ 0.50; strongest repeated-signed AUROC 0.567), while tested client-property attacks are not reliably above chance. We treat κ and audit frequency as exposure parameters; empirical leakage is evaluated in Appx. F, while Appx. A discusses ID exposure, pseudonymization, and matching.

Verifiable audit signatures. Because the audit plane is separate from secure aggregation, a malicious client could publish unrelated Top-K coordinates, suppress implicating coordinates, or flip signs without attacking the learning plane. We define a zero-knowledge consistency relation to verify the reported Top-K record against a committed hidden update without revealing remaining values, and validate it in a small-dimensional proof-of-concept. This provides audit-record consistency; binding the commitment to the update entering secure aggregation is a separate learning-plane property supported by verifiable-input secure aggregation. The relation and security boundary are detailed in Appx. B.

Directional fidelity to dense cosine. Figure 3 compares signed overlap with dense update–query cosine, a diagnostic distinct from the FedAvg-weighted dense first-order target used later. For 100 audit queries, both scores are aggregated across audited checkpoints and used to rank ten clients. Signed overlap achieves mean query-wise Spearman $\rho_S = 0.946$ (95% query-bootstrap CI $[0.934, 0.958]$), with 88% of queries above 0.90; effect signs agree for 91.7% of 1,000 query–client pairs.

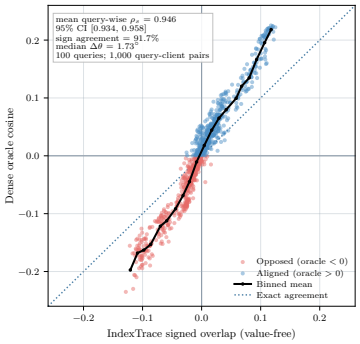


Figure 3: **Illustration: Directional fidelity of signed overlap.** Signed overlap versus dense update–query cosine over 1,000 query–client pairs; black shows binned means and dotted denotes equality.

4 Theoretical Analysis

What Signed Overlap Preserves. Let $J_q(\theta)$ be a differentiable audit objective oriented so that a larger value means more of the queried behavior, and let $g_{q,t} = \nabla_\theta J_q(\theta_t)$. For a vector $x \in \mathbb{R}^M$, define its signed Top-K mask by $[m_K(x)]_j = \mathrm{sgn}(x_j)\mathbb{I}\{\{\}\, j \in \mathrm{TopK}(|x|)\}$, using deterministic tie-breaking. With equal client and query budgets, the FedAvg-weighted round score is $s_{i,t}(q) = \frac{a_{i,t}}{K}\langle m_K(\Delta_{i,t}), m_K(g_{q,t})\rangle$, $I_{i,t}(q) = a_{i,t}\langle \Delta_{i,t}, g_{q,t}\rangle$. Thus $s_{i,t}/a_{i,t}$ is *exactly* the cosine between two ternary Top-K masks; it is not the cosine between the dense vectors. Write $O_{i,t} = \mathrm{TopK}(|\Delta_{i,t}|) \cap \mathrm{TopK}(|g_{q,t}|)$, $c_{i,t,j} = |\Delta_{i,t,j} g_{q,t,j}|$, and $\sigma_{i,t,j} = \mathrm{sgn}(\Delta_{i,t,j} g_{q,t,j})$. The following result states the sufficient conditions needed to pass from mask geometry to dense first-order attribution.

Proposition 1 (Sparse signed approximation)**.** *Suppose that client i has a retained-product scale $\mu_{i,t} > 0$ such that*

$$\sum_{j \notin O_{i,t}} c_{i,t,j} \le \tau_{i,t}, \qquad |c_{i,t,j} - \mu_{i,t}| \le \epsilon_{i,t} \quad (j \in O_{i,t}). \tag{5}$$

For any common reference $\mu_0 > 0$, let $\eta_{i,t} = |\mu_{i,t} - \mu_0|$. Then

$$|I_{i,t}(q) - \mu_0 K s_{i,t}(q)| \le E^{(0)}_{i,t} := a_{i,t}\big(\tau_{i,t} + \epsilon_{i,t}|O_{i,t}|\big) + \eta_{i,t} K |s_{i,t}(q)|. \tag{6}$$

If the conditions hold for every participating client, J_q is L_q-smooth, and $U_t = \sum_i a_{i,t}\Delta_{i,t}$, then

$$\left| J_q(\theta_t + U_t) - J_q(\theta_t) - \mu_0 K \sum_i s_{i,t}(q) \right| \le \frac{L_q}{2}\|U_t\|_2^2 + \sum_i E^{(0)}_{i,t}. \tag{7}$$

Corollary 1 (Direction and ranking)**.** *Under Proposition 1, the sign of client i is recovered whenever $\mu_0 K |s_{i,t}| > E^{(0)}_{i,t}$. For clients i and r, the dense ordering is recovered whenever*

$$\mu_0 K\big(s_{i,t}(q) - s_{r,t}(q)\big) > E^{(0)}_{i,t} + E^{(0)}_{r,t}. \tag{8}$$

Table 1: **What information is required for attribution?** SuperNI/Llama-3B results on 500 queries against the same dense first-order reference. "Privileged values?" denotes access to numerical client-side values unavailable to AUDITRACE. AUROC measures positive-effect detection and NDCG@5 ranks positive effects by magnitude; higher is better.

Method	Client-side information	Privileged values?	ρ_S	Sign acc.	AUROC	NDCG@5
AUDITRACE	Coordinate IDs + signs	No	0.917	0.904	0.965	0.917
Signed-rank	IDs + signs + ranks	No	0.918	0.903	0.966	0.914
Dense cosine	Dense values	Yes	0.974	0.965	0.992	0.974
Data size	Client metadata	No	0.008	0.501	0.498	0.456
Random	None	No	-0.010	0.494	0.493	0.305

Proposition 1 is a conditional worst-case result: its error depends on discarded interaction mass, retained-product dispersion, and cross-client scale mismatch, quantities requiring dense instrumented updates and unavailable as deployment-time certificates. Appx. C provides the proof and a complementary magnitude-free signal-plus-noise model for sparse-score stability; that model does not establish agreement with dense or LOO effects. Exact mask-space geometry is unconditional, whereas links to dense effects, retraining, interventions, robustness, or privacy require their stated assumptions or separate empirical evidence. Appx. C.12 provides empirical context for sparse-to-dense fidelity without verifying the sufficient conditions. Thus, the theory characterizes sufficient regimes and failure modes rather than explaining the observed empirical correlations.

5 Experiments

We ask four questions: (1) what information in the sparse audit record is needed for attribution; (2) does the magnitude-free score agree with explicit client-exclusion retraining; (3) does dense-reference fidelity transfer across models, tasks, and partitions; and (4) how does fidelity trade off against signature budget and exposure. Additional robustness, security analysis, ablations, client-removal interventions and protocol details appear in Appx. D.

5.1 Experimental protocol

We evaluate LoRA adaptation on SuperNI, BeaverTails, and AG News using Llama-3.2-1B/3B-Instruct, Qwen2.5-3B-Instruct, and RoBERTa-base. The exact AG News leave-one-client-out (LOO) experiment uses Llama-3.2-1B-Instruct with only q/k/v/o LoRA factors trainable and audited; no task-specific dense head is included. Our primary estimator is the FedAvg-weighted signed-overlap score, with realized $K/M = 0.10$ for headline archive-A comparisons; fixed-budget studies are labeled separately. AG News and SuperNI use task loss, while the intervention uses unsafe-response target log-likelihood, with positive scores always oriented toward increasing the audited objective. Dense-reference evaluations use the corresponding FedAvg-weighted update–gradient inner product, while exact LOO uses explicit retraining. Headline point estimates average five paired training/partition seeds; reported query-bootstrap intervals are conditional on the evaluated runs. Full dataset/model, optimization, audit-objective, and evaluation details appear in Appx. D.

5.2 What information is needed for attribution?

We first isolate which parts of the sparse audit representation contribute to attribution fidelity. Table 1 evaluates matched estimators against the same dense first-order reference on SuperNI/Llama-3B. The proposed signed representation achieves $\rho_S = 0.917$. Retaining coordinate ranks in addition to signs yields 0.918, showing essentially no additional gain from rank information in this setting. Conversely, the rank-only ablation reported in Appx. E.1 achieves $\rho_S = 0.009$. Together, these results indicate that directional sign information is substantially more useful for recovering the dense client ordering than coordinate-rank information.

Table 2: **Agreement with exact leave-one-client-out retraining effects on AG News/Llama-3.2-1B-Instruct.** Sign is effect-sign accuracy, and N@5 is positive-effect NDCG@5. Higher is better.

Method	Interface	Privileged values?	ρ_S	τ	Pearson	Sign	AUROC	AUPRC	N@5
AUDITRACE	Sparse IDs + signs	No	0.827	0.686	0.672	0.834	0.941	0.948	0.893
DataInf-client (Kwon et al., 2024)	Dense statistics	Yes	0.881	0.760	0.749	0.865	0.963	0.967	0.927
Dense cosine	Dense updates/gradients	Yes	0.847	0.711	0.699	0.871	0.952	0.957	0.894
Client-Fisher (Koh & Liang, 2017)	Dense gradients/Fisher	Yes	0.774	0.627	0.649	0.839	0.924	0.926	0.823
Data size	Metadata	No	-0.025	-0.037	-0.230	0.504	0.497	0.615	0.571
Random	None	No	0.002	0.005	-0.000	0.501	0.504	0.606	0.402

Table 3: Attribution fidelity across models, tasks, and client partitions. Last two rows: additional fixed-10% runs with different query cohorts. AUROC identifies positive dense effects; NDCG@5 uses their magnitudes. All entries are five-seed query means.

Setting	Queries	ρ_S	Sign acc.	AUROC	NDCG@5
SuperNI / Llama-1B	500	0.903	0.893	0.956	0.910
SuperNI / Llama-3B	500	0.917	0.904	0.965	0.917
SuperNI / Qwen-3B	500	0.919	0.904	0.967	0.918
BeaverTails / Llama-3B	500	0.886	0.877	0.969	0.946
AG News / IID	100	0.839	0.836	0.951	0.911
AG News / Dir. (.5)	100	0.952	0.961	0.993	0.983
AG News / Dir. (.3)	100	0.914	0.937	0.992	0.973
AG News / Dir. (.1)	100	0.932	0.919	0.972	0.940
SuperNI / Llama-1B (fixed)	100	0.908	0.884	0.956	0.912
BeaverTails / Llama-1B (unsafe, fixed)	200	0.891	0.886	0.955	0.893

5.3 Comparison with attribution baselines

We test whether AUDITRACE recovers exact client counterfactuals rather than merely matching another attribution proxy. Table 2 compares all methods with exact leave-one-client-out (LOO) effects on AG News/Llama-3.2-1B-Instruct with 10 clients and a Dirichlet $\alpha = 0.3$ partition. For each of five paired training/partition seeds, we retrain 10 LOO models and evaluate 100 balanced queries, yielding 1,000 query–client effects per seed. Only q/k/v/o LoRA parameters are trainable and audited; no dense task-specific head is included. Privileged baselines access numerical client information unavailable to AUDITRACE, which uses only Top-K coordinate identities and signs. AUDITRACE achieves $\rho_S = 0.827$, positive-effect AUROC 0.941, AUPRC 0.948, and NDCG@5 0.893. Privileged methods improve some full-ranking metrics, but AUDITRACE nearly matches dense cosine on NDCG@5 (paired difference -0.0007, 95% CI $[-0.0155, 0.0142]$) and exceeds Client-Fisher in Spearman correlation by $+0.053$ ($[0.032, 0.075]$). Thus, signed Top-K records preserve substantial client ordering under a weaker information interface, although 39% top-one recovery cautions against single-client identification.

5.4 Generalization across models, tasks, and partitions

Across the original eight settings, signed-overlap Spearman correlation ranges from 0.839 to 0.952 (Table 3). SuperNI results are stable across Llama-1B, Llama-3B, and Qwen-3B (0.903, 0.917, and 0.919). Fidelity also remains high under non-IID AG News partitions: $\rho_S = 0.952$, 0.914, and 0.932 for Dirichlet $\alpha = 0.5$, 0.3, and 0.1, respectively, with AUROC 0.993, 0.992, and 0.972; the IID setting gives $\rho_S = 0.839$. Additional fixed-budget Llama-1B runs achieve 0.908 on SuperNI and 0.891 on unsafe BeaverTails queries. Overall, dense-reference fidelity remains strong across the evaluated models, tasks, and client heterogeneity, without implying equivalence across model sizes or causal attribution on every task.

5.5 Fidelity–audit-payload trade-off

We rescore the same saved SuperNI/Llama-1B training trajectory at four predefined signature budgets (Table 4). Fidelity improves smoothly with serialized signature payload: increasing K/M from 1%

Table 4: **Fidelity–audit-payload trade-off.** SuperNI/Llama-1B frozen-trajectory budget sweep using 100 queries. The same saved training trajectory is rescored at each budget; intervals are query-bootstrap 95% CIs conditional on the trained model. KiB is the mean serialized signature per client and audited round; compression is relative to the FP32 trainable-update vector.

K/M	ρ_S	95% CI	Sign acc.	NDCG@5	KiB	Compression
1%	0.569	[0.538, 0.599]	0.707	0.618	96.3	138.3×
2%	0.730	[0.707, 0.752]	0.782	0.748	177.9	74.8×
5%	0.860	[0.844, 0.875]	0.840	0.868	415.4	32.0×
10%	0.909	[0.897, 0.919]	0.884	0.913	809.0	16.5×

Table 5: **Attribution utility and audit-record leakage.** AUROC under run-disjoint attacks; chance is 0.5. Full attack protocols, CIs, AUPRC, Top-1, and additional views are in Appx. F.

K/M	Attribution ρ_S	Membership support	Membership signed, repeated	Label-2 property max. AUROC
1%	0.565	0.503	0.523	0.577
5%	0.735	0.499	**0.567**	0.554
10%	0.770	0.503	0.565	0.559

to 10% raises ρ_S from 0.569 to 0.909, sign accuracy from 70.7% to 88.4%, and NDCG@5 from 0.618 to 0.913. At 5%, the estimator already obtains $\rho_S = 0.860$ with a mean serialized signature of 415.4 KiB; at 10%, it obtains 0.909 with 809.0 KiB, making the serialized record 16.5× smaller than the FP32 trainable-update vector. These audit payloads are additional to the learning-plane communication; we do not extrapolate beyond the measured range.

5.6 Multi-View Audit-Record Leakage

The audit record is an additional client-linked disclosure and is not differentially private. Using Llama-3.2-1B-Instruct with a strictly LoRA-only audited coordinate universe, we evaluate two leakage objectives: candidate-conditioned example membership and a prespecified client property (high vs. low label-2 proportion). Attacks are evaluated with complete FL runs held out from attacker training and model selection. We compare support-only, single-checkpoint signed, and repeated signed records to test the effects of signs and temporal accumulation; full attack construction and model-selection details appear in Appx. F. Table 5 reports attack AUROC together with attribution fidelity on the same leakage cohort. Support-only membership inference remains at chance (0.499–0.503). Repeated signed records yield a small but reproducible advantage, peaking at AUROC 0.567 (95% CI $[0.559, 0.574]$) at $K = 5\%$. For the tested client property, all confidence intervals include the 0.5 chance level. Thus, the evaluated attacks remain close to chance in absolute terms, although signs and repeated disclosure introduce measurable membership leakage.

These are attack-specific empirical measurements rather than a formal privacy guarantee; Appx. F additionally evaluates single- versus repeated-signed and dense-assisted views, while generative reconstruction from signed Top-K records is not evaluated.

6 Conclusion

Federated PEFT needs post-hoc auditing without exposing individual dense numerical client updates. AUDITRACE provides a magnitude-free interface exposing only sparse coordinate identities and signs. Across dense-reference, exact LOO, and intervention evaluations, it preserves substantial behavior-conditioned attribution: against exact LOO, AUDITRACE reaches $\rho_S = 0.827$ and positive-effect AUROC 0.941, while top-five removal improves the held-out unsafe-target objective by 0.0232 versus 0.0230 for dense-oracle and 0.0129 for random removal. Its limits are explicit: signed Top-K records miss query-aligned mass outside retained support, and the audit record is a client-linked disclosure rather than a privacy guarantee. Overall, AUDITRACE provides a practical audit interface compatible with a secure-aggregation learning path without exposing dense numerical update values, while explicitly retaining its disclosure and retained-support limitations in deployment.

AI USE STATEMENT

Generative AI tools were used only for language editing and proofreading. They were not used to develop the method, formulate hypotheses, design experiments, analyze results, generate scientific claims, or produce experimental data. The authors reviewed and approved all final content.

REPRODUCIBILITY STATEMENT

We provide experimental configurations, dataset and partitioning details, model and adapter settings, optimization hyperparameters, audit objectives and query counts, evaluation protocols, and additional analyses in the appendix. The implementation code is included in the supplementary materials to support reproduction of the reported results.

ETHICS STATEMENT

Attribution mechanisms can single out participants, so we state the risks explicitly. The audit record is an additional client-linked disclosure with no formal privacy guarantee; it is not differentially private, and we measure a small but reproducible membership signal. Signature fraction, audit frequency, access control, and retention should be governed in deployment. AUDITRACE yields an objective-conditioned ranking, not evidence of provenance, data quality, or intent, and should not be the sole basis for penalizing a participant; a client that shapes its update magnitudes can also evade it. All experiments use public benchmarks with simulated clients; no human subjects or personal data were involved, and no generated harmful content is released.

REFERENCES

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pp. 2938–2948. PMLR, 2020.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0461.

Mislav Balunovic, Dimitar Dimitrov, Nikola Jovanović, and Martin Vechev. Lamp: Extracting text from gradients with language model priors. *Advances in Neural Information Processing Systems*, 35:7641–7654, 2022.

James Bell, Adrià Gascón, Tancrède Lepoint, Baiyu Li, Sarah Meiklejohn, Mariana Raykova, and Cathie Yun. {ACORN}: input validation for secure aggregation. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 4805–4822, 2023.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/6617-machine-learning-with-adversaries-byzantine-tolerant-gradient-descent.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191. Association for Computing Machinery, 2017. doi: 10.1145/3133956.3133982. URL https://doi.org/10.1145/3133956.3133982.

Yekun Chai, Qingyi Liu, Shuohuan Wang, Yu Sun, Qiwei Peng, and Hua Wu. On training data influence of gpt models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3126–3150, 2024.

Yi-Chung Chen, Hsi-Wen Chen, Shun-Gui Wang, and Ming-Syan Chen. SPACE: Single-round participant amalgamation for contribution evaluation in federated learning.

In *Advances in Neural Information Processing Systems*, volume 36, pp. 6422–6441, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/14a812fa4b6bf244d055e37a7cd2f557-Abstract-Conference.html.

Javad Dogani, Devriş İşler, Nikolaos Laoutaris, et al. Index-only backdoor vetting for secure federated peft of large language models. In *International Conference on Machine Learning*, 2026.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients—how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 16937–16947, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Waris Gill, Ali Anwar, and Muhammad Ali Gulzar. Feddebug: Systematic debugging for federated learning applications. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pp. 512–523. IEEE, 2023.

Waris Gill, Ali Anwar, and Muhammad Ali Gulzar. TraceFL: Interpretability-driven debugging in federated learning via neuron provenance. In *2025 IEEE/ACM 47th International Conference on Software Engineering*, pp. 2264–2276. IEEE, 2025. doi: 10.1109/ICSE55347.2025.00128. URL https://doi.org/10.1109/ICSE55347.2025.00128.

Waris Gill, Ahmad Humayun, Ali Anwar, and Muhammad A Gulzar. Protoken: Token-level attribution for federated large language models. *Proceedings of Machine Learning and Systems*, 8: 2249–2263, 2026.

Jens Groth. On the size of pairing-based non-interactive arguments. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 305–326. Springer, 2016.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuqi Jia, Minghong Fang, Hongbin Liu, Jinghuai Zhang, and Neil Gong. Tracing back the malicious clients in poisoning attacks to federated learning. *Advances in Neural Information Processing Systems*, 38:116470–116497, 2026.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9m02ib92Wz.

Shenghui Li, Edith C-H Ngai, Fanghua Ye, and Thiemo Voigt. Peft-as-an-attack! jailbreaking language models during federated parameter-efficient fine-tuning. *arXiv preprint arXiv:2411.19335*, 2024.

Huawei Lin, Jikai Long, Zhaozhuo Xu, and Weijie Zhao. Token-wise influential training data retrieval for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 841–860, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.48.

Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. GTG-Shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Transactions on Intelligent Systems and Technology*, 13(4), 2022. doi: 10.1145/3501811. URL https://doi.org/10.1145/3501811.

Hidde Lycklama, Lukas Burkhalter, Alexander Viand, Nicolas Küchler, and Anwar Hithnawi. RoFL: Robustness of secure federated learning. In *2023 IEEE Symposium on Security and Privacy*, pp. 453–476. IEEE, 2023. doi: 10.1109/SP46215.2023.10179400. URL https://doi.org/10.1109/SP46215.2023.10179400.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, et al. {FLAME}: Taming backdoors in federated learning. In *31st USENIX security symposium (USENIX Security 22)*, pp. 1415–1432, 2022.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 2023. URL https://proceedings.mlr.press/v202/park23c.html.

Ivo Petrov, Dimitar I. Dimitrov, Maximilian Baader, Mark Niklas Müller, and Martin Vechev. DAGER: Exact gradient inversion for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9ff1577a1f8308df1ccea6b4f64a103f-Abstract-Conference.html.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Amrita Roy Chowdhury, Chuan Guo, Somesh Jha, and Laurens Van der Maaten. Eiffel: Ensuring integrity for federated learning. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pp. 2535–2549, 2022.

Tianshu Song, Yongxin Tong, and Shuyue Wei. Profit allocation for federated learning. In *2019 IEEE International Conference on Big Data*, pp. 2577–2586. IEEE, 2019.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16070–16084, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/b8ffa41d4e492f0fad2f13e29e1762eb-Abstract.html.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. In *Federated Learning: Privacy and Incentive*, volume 12500 of *Lecture Notes in Computer Science*, pp. 153–167. Springer, 2020b. doi: 10.1007/978-3-030-63076-8_11. URL https://arxiv.org/abs/2009.06192.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Rick Siow Mong Goh, Yong Liu, and Lei Zhu. Federated residual low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2025.

Rui Ye, Jingyi Chai, Xiangrui Liu, Yaodong Yang, Yanfeng Wang, and Siheng Chen. Emerging safety attack and defense in federated instruction tuning of large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sYNWqQYJhz.

Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5650–5659. PMLR, 2018. URL https://proceedings.mlr.press/v80/yin18a.html.

Su Zhang, Junfeng Guo, and Heng Huang. Fedattr: Towards privacy-preserving client-level attribution in federated llm fine-tuning. *arXiv preprint arXiv:2605.06596*, 2026.

Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9963–9977. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.632. URL https://aclanthology.org/2023.findings-acl.632.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html.

A Adversarial Audit-Signature Leakage and Privacy Hardening

What the audit signature reveals. For client i at round t, the default audit interface exposes $\mathcal{Z}_{i,t} = \{(j, \mathrm{sgn}(\Delta_{i,t,j})) : j \in \mathcal{S}_{i,t}\}, \quad |\mathcal{S}_{i,t}| = k$. It therefore reveals which k coordinates are most salient and the direction of each selected coordinate. It does not reveal their magnitudes, the values or signs of unselected coordinates, or the dense update itself. Secure aggregation continues to protect the dense learning-plane update independently of this audit metadata.

This restricted representation nevertheless creates a residual information channel. In particular, repeated coordinate occurrence across rounds may reveal persistent structure in a client's optimization signal, and sign patterns may provide additional directional information. We therefore make no claim that the signature itself provides differential privacy or cryptographic confidentiality.

Pseudonymized coordinate identifiers. When raw parameter identifiers are themselves considered sensitive metadata, they need not be transmitted directly. Let $F_{k_{\mathrm{prf}}}$ be a keyed pseudorandom function and r_t a fresh round identifier. Before transmission, coordinate j is replaced by

$$\widetilde{j}_t = F_{k_{\mathrm{prf}}}(r_t \,\|\, j). \tag{9}$$

The client-side and query-side signature procedures use the same mapping within round t, so

$$j_1 = j_2 \quad \Longleftrightarrow \quad \widetilde{j}_{1,t} = \widetilde{j}_{2,t} \tag{10}$$

except with negligible collision probability. Consequently, AUDITRACE can compute the same overlap and signed-agreement score using pseudonyms. Rotating r_t changes the representation of a coordinate across rounds; temporal aggregation remains unaffected because matching is completed within each checkpoint before the resulting scalar scores are aggregated.

A public hash or public salt alone should not be interpreted as strong protection: because j belongs to the finite domain $\{1, \ldots, d\}$, an observer can enumerate candidate coordinates. Likewise, using an independent salt for every client would prevent client–query equality matching. Keyed, synchronized pseudonyms avoid these two problems.

This construction primarily protects coordinate identifiers in transit, logs, or from components that perform equality matching without access to the key. It does not protect raw indices from an authorized component that knows the k_{prf}. If the audit service itself must not learn coordinate identities, the client and query side can instead perform private set intersection or OPRF-based matching so that only the required intersections are exposed. Such cryptographic hardening is complementary to the AUDITRACE attribution estimator.

The default AUDITRACE design should therefore be interpreted as *magnitude-free*, not leakage-free. It removes numerical update magnitudes from the audit interface but intentionally exposes a controlled signed support of size K. Appx. F empirically evaluates membership and client-property leakage from this disclosure.

B Zero-Knowledge Consistency of Audit Records

Threat and security objective. The learning and audit records are transmitted through separate channels. Without integrity verification, a malicious client could retain its training update while publishing unrelated benign-looking Top-k coordinates, suppressing coordinates that implicate it in an audited behavior, or reversing sign bits. This attack does not require breaking secure aggregation. Our ZK layer targets this audit-plane attack: it verifies that the disclosed signed Top-k record is correctly derived from a committed hidden update while revealing no additional update values.

Verified relation. Let $u_{i,t}$ denote the fixed-point representation of the client update used by the proof system, and let

$$C_{i,t} = \mathrm{Com}(u_{i,t}; r_{i,t}) \tag{11}$$

be the commitment used by the protocol, with randomness $r_{i,t}$. We require the commitment to be binding, so that a client cannot open the same commitment to different updates, and hiding, so that publishing $C_{i,t}$ does not reveal the hidden update.

The public statement consists of $C_{i,t}$ and the reported signature $\mathcal{Z}_{i,t}$; $u_{i,t}$ and $r_{i,t}$ remain private witnesses. The circuit first verifies that the witness opens the public commitment,

$$C_{i,t} = \mathrm{Com}(u_{i,t}; r_{i,t}), \tag{12}$$

and, writing $\mathcal{S}_{i,t}$ for the reported support, establishes

$$\begin{aligned} &|\mathcal{S}_{i,t}| = k, \\ &\mathrm{sgn}(u_{i,t,j}) = s_{i,t,j}, \quad \forall j \in \mathcal{S}_{i,t}, \\ &|u_{i,t,j}| \geq |u_{i,t,\ell}|, \qquad \forall j \in \mathcal{S}_{i,t},\ \ell \notin \mathcal{S}_{i,t}, \end{aligned} \tag{13}$$

where $s_{i,t,j}$ is the reported sign of coordinate j. Equal-magnitude ties are resolved by the same deterministic index rule used by signature extraction. Consequently, a client cannot substitute arbitrary benign-looking coordinates or signs while producing a valid proof for the committed update.

Efficient Top-k verification. A direct encoding of the final condition in Equation equation 13 would require selected–unselected pairwise comparisons. Instead, the circuit verifies a k-th magnitude threshold $\tau_{i,t}$:

$$|u_{i,t,j}| \geq \tau_{i,t} \quad \forall j \in \mathcal{S}_{i,t}, \qquad |u_{i,t,\ell}| \leq \tau_{i,t} \quad \forall \ell \notin \mathcal{S}_{i,t}, \tag{14}$$

Together with cardinality and deterministic tie-breaking constraints, this reduces selected–unselected comparisons from $O(K(M-K))$ to $O(M)$ coordinate-wise checks, although circuit size still grows with the hidden update dimension. For the multi-million-coordinate PEFT vectors used in our attribution experiments, full-scale circuit size and proving cost are not evaluated in this work.

Proposed Protocol. For an audited round, a full instantiation would convert the client update to the fixed-point representation used by the proof system, commit to that representation, extract its signed Top-k signature, and generate a zero-knowledge proof of Equations equation 12–equation 14. The audit service accepts $\mathcal{Z}_{i,t}$ only if the proof verifies. In a deployment using SecAgg, the dense learning update remains on the secure-aggregation path and is not disclosed to the auditor. Under the zero-knowledge property of the proof system and the hiding property of the commitment, verification reveals no additional update values beyond the signed Top-k record already disclosed by AUDITRACE.

Relation to secure aggregation. Audit-record consistency and learning-plane input consistency are distinct properties. The proof above establishes that $\mathcal{Z}_{i,t}$ is derived from the vector opening $C_{i,t}$, but it does not by itself prove that this vector is exactly the private input encoded into an independently instantiated secure-aggregation message. Verifiable-input secure-aggregation systems such as RoFL, EIFFeL, and ACORN provide mechanisms for coupling private-input validation with the values entering aggregation (Lycklama et al., 2023; Roy Chowdhury et al., 2022; Bell et al., 2023). The AUDITRACE audit interface is separable from such learning-plane constructions; integrating and benchmarking their composition is outside our evaluated system.

Prototype implementation and measured microbenchmark cost. We implement the audit-consistency relation using the Groth16 preprocessing zk-SNARK (Groth, 2016), with the circuit written in Circom (**?**) and proving and verification performed using snarkjs. The reported implementation is a small-scale proof-of-concept used to validate the circuit construction; it is not instantiated at the multi-million-coordinate PEFT dimensions used in our attribution experiments. For this prototype instance, the compiled circuit contains 6,318 non-linear constraints and 715 linear constraints, with 32 private inputs, 4 public inputs, and 1 public output. Starting from an existing Powers-of-Tau transcript, the circuit-specific Groth16 setup takes approximately 7.1 s, proof generation takes approximately 4.02 s, and verification approximately 0.86 s.

These measurements characterize only this prototype circuit instance. The Top-k threshold relation in Equation equation 14 requires coordinate-wise checks that scale linearly with the hidden update dimension M; therefore the reported constraint count and timings must not be extrapolated to the full

Llama or RoBERTa configurations evaluated elsewhere in the paper. We do not report a deployment-scale proving cost for those configurations. For a fixed circuit parameterization, the circuit-specific Groth16 setup is performed once and the resulting proving and verification keys can be reused across clients and audited rounds.

Security boundary. When instantiated at the required update dimension, the ZK relation provides audit-record consistency relative to the committed hidden update. It does not certify that the client's local data are benign, that the client followed an intended optimization procedure, or that the update itself is non-adversarial. A malicious client may deliberately construct a poisoning update and still provide a valid proof of its true signed Top-k signature; AUDITRACE provides attribution rather than poisoning prevention. End-to-end enforcement that the committed update is also the exact SecAgg input belongs to the verifiable-input learning-plane layer discussed above.

C Additional theoretical analysis

C.1 Setup and exact mask-space geometry

Fix an audited checkpoint t and query q. We suppress (q, t) where doing so is unambiguous. The update and query supports are $S_i = \mathrm{TopK}(|\Delta_i|)$ and $S_q = \mathrm{TopK}(|g_q|)$, with deterministic tie-breaking and K nonzero selected entries on each side. Define the ternary masks $z_i = m_K(\Delta_i)$ and $z_q = m_K(g_q)$. Then $\|z_i\|_2 = \|z_q\|_2 = \sqrt{K}$ and

$$\frac{1}{K}\langle z_i, z_q\rangle = \frac{\langle z_i, z_q\rangle}{\|z_i\|_2\|z_q\|_2} = \cos(z_i, z_q). \tag{15}$$

For unequal budgets K_i, K_q, the corresponding normalized score is $\langle z_i, z_q\rangle/\sqrt{K_i K_q}$. Equation equation 15 is an exact identity in mask space. No conclusion about $\cos(\Delta_i, g_q)$ follows without additional magnitude assumptions.

Let

$$O_i = S_i \cap S_q, \quad c_{i,j} = |\Delta_{i,j} g_{q,j}|, \quad \sigma_{i,j} = \mathrm{sgn}(\Delta_{i,j} g_{q,j}), \tag{16}$$

so that

$$I_i = a_i \sum_{j=1}^{M} c_{i,j}\sigma_{i,j}, \qquad s_i = \frac{a_i}{K}\sum_{j\in O_i}\sigma_{i,j}. \tag{17}$$

The positive FedAvg weight a_i is included in both quantities because the empirical target is the client's contribution to the aggregated update. If the scientific question instead concerns an unweighted local update, set $a_i = 1$ in both definitions; mixing weighted targets with unweighted scores is not covered by the results.

C.2 Proof of Proposition 1

Using Eq. equation 17,

$$I_i - \mu_i K s_i = a_i \sum_{j\in O_i}(c_{i,j} - \mu_i)\sigma_{i,j} + a_i \sum_{j\notin O_i} c_{i,j}\sigma_{i,j}. \tag{18}$$

The triangle inequality and Eq. equation 5 give

$$|I_i - \mu_i K s_i| \le a_i \sum_{j\in O_i}|c_{i,j} - \mu_i| + a_i \sum_{j\notin O_i} c_{i,j} \tag{19}$$

$$\le a_i\big(\epsilon_i|O_i| + \tau_i\big) =: E_i. \tag{20}$$

Adding and subtracting $\mu_0 K s_i$ yields

$$|I_i - \mu_0 K s_i| \le E_i + |\mu_i - \mu_0|K|s_i| = E_i^{(0)}, \tag{21}$$

which proves Eq. equation 6. If J_q has an L_q-Lipschitz gradient, the standard smoothness remainder gives

$$|J_q(\theta_t + U_t) - J_q(\theta_t) - \langle g_{q,t}, U_t\rangle| \le \frac{L_q}{2}\|U_t\|_2^2. \tag{22}$$

Because $\langle g_{q,t}, U_t \rangle = \sum_i I_{i,t}(q)$, summing the client-wise bounds and applying the triangle inequality proves Eq. equation 7.

For Corollary 1, Eq. equation 6 implies $I_i \in [\mu_0 K s_i - E_i^{(0)}, \mu_0 K s_i + E_i^{(0)}]$. If $\mu_0 K|s_i| > E_i^{(0)}$, this interval excludes zero and has the sign of s_i. Likewise, Eq. equation 8 makes the lower endpoint for I_i exceed the upper endpoint for I_r.

C.3 Why separate Top-K concentration controls the interaction tail

The tail assumption in Eq. equation 5 concerns products, not the client update alone. It can nevertheless be bounded using the two separate salience tails.

Lemma 1 (Product-tail bound). *Let $u_j = |\Delta_{i,j}|$, $v_j = |g_{q,j}|$, and define*

$$\xi_i = \frac{\sum_{j \notin S_i} u_j}{\|u\|_1}, \qquad \xi_q = \frac{\sum_{j \notin S_q} v_j}{\|v\|_1}. \tag{23}$$

Then

$$\sum_{j \notin O_i} u_j v_j \le \|v\|_\infty \xi_i \|u\|_1 + \|u\|_\infty \xi_q \|v\|_1. \tag{24}$$

Proof. Since $O_i^c = S_i^c \cup S_q^c$, the left-hand side is at most $\sum_{j \notin S_i} u_j v_j + \sum_{j \notin S_q} u_j v_j$. Bound the first sum by $\|v\|_\infty \sum_{j \notin S_i} u_j$ and the second by $\|u\|_\infty \sum_{j \notin S_q} v_j$. □

Low marginal tail mass on both vectors is therefore sufficient for low discarded interaction mass. The converse need not hold, and Top-K concentration of only one vector is insufficient.

C.4 Cross-client retained scales

Equation equation 21 makes cross-client scale variation explicit instead of assuming one exact retained magnitude for every client. Removing magnitudes is most reliable when the μ_i are similar, or when the signed-score margin dominates both within-client dispersion and the $|\mu_i - \mu_0|K|s_i|$ terms. The bound holds for any prespecified μ_0; choosing it after seeing test outcomes would invalidate a confirmatory certificate. A defensible protocol estimates μ_0 on a disjoint calibration split (for example, as the median retained scale) and freezes it before test evaluation. Moreover, μ_i, τ_i, and $E_i^{(0)}$ require dense individual updates and are measurable in instrumented evaluation runs but not by the deployed auditor under secure aggregation.

C.5 Magnitude-free signal-plus-noise interpretation

The deterministic approximation above can be loose when coordinate-product magnitudes are highly dispersed. A complementary model explains when the *sparse score itself* is stable without any magnitude homogeneity. Let $\mathcal{S}$ denote the collection of fixed supports and define $u_i = s_i/a_i = K^{-1} \sum_{j \in O_i} X_{i,j}$, where $X_{i,j} = \sigma_{i,j} \in \{-1, +1\}$. Let $m_i = |O_i|$, $\rho_i = m_i/K$, and, for $m_i > 0$, $\beta_i = m_i^{-1} \sum_{j \in O_i} \mathbb{E}[X_{i,j} \mid \mathcal{S}]$.

Proposition 2 (Signed-support signal plus noise). *Suppose $m_i > 0$ and the variables $\{X_{i,j} : j \in O_i\}$ are conditionally independent given $\mathcal{S}$. Then*

$$\mathbb{E}[u_i \mid \mathcal{S}] = \rho_i \beta_i, \qquad \Pr(|u_i - \rho_i \beta_i| \ge x \mid \mathcal{S}) \le 2 \exp\left(-\frac{K^2 x^2}{2 m_i}\right). \tag{25}$$

For two clients with $m_i, m_r > 0$ and $\gamma_{ir} = \rho_i \beta_i - \rho_r \beta_r > 0$,

$$\Pr(u_i \le u_r \mid \mathcal{S}) \le \exp\left(-\frac{K^2 \gamma_{ir}^2}{8 m_i}\right) + \exp\left(-\frac{K^2 \gamma_{ir}^2}{8 m_r}\right). \tag{26}$$

Proof. Linearity gives the expectation. Conditional Hoeffding applied to the m_i independent variables in $[-1, 1]$ gives the first tail bound. If $u_i \le u_r$, then either u_i falls at least $\gamma_{ir}/2$ below its mean or u_r rises at least $\gamma_{ir}/2$ above its mean. Applying the two one-sided bounds and a union bound proves Eq. equation 26; cross-client independence is not required. □

The decomposition $\rho_i\beta_i$ separates how often client and query supports meet (ρ_i) from their average directional agreement (β_i). For the FedAvg-weighted score, multiply the mean and deviations by a_i. This result is insensitive to heavy-tailed magnitudes, but it does *not* show that u_i approximates dense influence or leave-one-out effects; those links still require Proposition 1 or empirical validation. Conditional coordinate independence is also a modeling assumption and its calibration must be checked on held-out runs.

C.6 Temporal and multi-query aggregation

Let $\lambda_{i,t} \geq 0$ be the normalized temporal weights over rounds in which client i participates, with $\sum_t \lambda_{i,t} = 1$. Define $\bar{I}_i(q) = \sum_t \lambda_{i,t} I_{i,t}(q)$ and $\bar{s}_i(q) = \sum_t \lambda_{i,t} s_{i,t}(q)$. Applying Proposition 1 per round gives the following deterministic extension.

Corollary 2 (Temporal aggregation)**.** *For a common reference $\mu_0 > 0$,*

$$|\bar{I}_i(q) - \mu_0 K \bar{s}_i(q)| \leq \sum_t \lambda_{i,t} E_{i,t}^{(0)} =: \bar{E}_i(q). \tag{27}$$

Hence the temporally aggregated ordering of i above r is certified if

$$\mu_0 K\big(\bar{s}_i(q) - \bar{s}_r(q)\big) > \bar{E}_i(q) + \bar{E}_r(q). \tag{28}$$

The same argument applies to a behavior represented by queries $\mathcal{Q}$ with prespecified weights $\nu_q \geq 0$, $\sum_q \nu_q = 1$: the aggregate error is at most $\sum_q \nu_q \bar{E}_i(q)$. This statement requires the reference-scale variation across queries to be included in $E_{i,t}^{(0)}$. Without that condition, averaging uncalibrated signed scores does not in general reproduce an average of dense contribution magnitudes.

C.7 Stability of signed masks across checkpoints

The score is Lipschitz in signed-mask drift, yielding a deterministic temporal stability condition that does not assume independent coordinates.

Lemma 2 (Mask-drift stability)**.** *For ternary masks z_i, z_q, z_i', z_q' with K selected entries per mask, let $r(z_i, z_q) = K^{-1}\langle z_i, z_q\rangle$. Then*

$$|r(z_i, z_q) - r(z_i', z_q')| \leq \min\left\{2, \frac{\|z_i - z_i'\|_1 + \|z_q - z_q'\|_1}{K}\right\}. \tag{29}$$

Proof. Add and subtract $K^{-1}\langle z_i', z_q\rangle$, apply the triangle inequality, and use $\|z\|_\infty \leq 1$ for every ternary mask. The additional bound by two follows because both scores lie in $[-1, 1]$. □

Support replacements contribute one unit per entering or leaving coordinate to the ℓ_1 drift, whereas a sign reversal on a retained coordinate contributes two. Thus high signed-support stability across audited rounds is sufficient for a stable round-wise score. It does not by itself establish stability of the dense interaction scale, which is separately represented by Eq. equation 21. The bound can be loose: replacing 20% of each support, with no other sign changes, gives a right-hand side of 0.8. We therefore report the distribution of normalized mask drift and the observed score change, rather than presenting this worst-case inequality as an empirical stability guarantee.

C.8 A testable hypothesis for non-IID mixtures

Non-IID data do not universally help or hurt signed overlap. A distributional statement requires an explicit link between client mixtures and expected signed supports. We state that link as a hypothesis rather than as a property of federated data.

Hypothesis 1 (Locally mixture-linear signed supports)**.** *Let $\pi_i \in \mathbb{R}^R$ be client i's mixture over R latent data components. At checkpoint t, suppose*

$$\mathbb{E}[z_{i,t} \mid \pi_i] = B_t \pi_i + e_{i,t}, \qquad \|e_{i,t}\|_2 \leq \delta_{i,t}, \tag{30}$$

for some $B_t \in \mathbb{R}^{M \times R}$. This is a local approximation, not a consequence of non-IID sampling; it is plausible only when Top-K boundaries and signs are sufficiently stable over the compared mixtures.

Proposition 3 (Expected distribution–support alignment)**.** *Under Hypothesis 1, define the unweighted signed score* $u_{i,t}(q) = K^{-1} z_{q,t}^\top z_{i,t}$ *and* $A_{q,t} = K^{-1} B_t^\top z_{q,t}$*. Then*

$$\left|\mathbb{E}[u_{i,t}(q) \mid \pi_i] - A_{q,t}^\top \pi_i\right| \le \frac{\delta_{i,t}}{\sqrt{K}}. \tag{31}$$

Therefore the expected score of i *exceeds that of* r *whenever*

$$A_{q,t}^\top(\pi_i - \pi_r) > \frac{\delta_{i,t} + \delta_{r,t}}{\sqrt{K}}. \tag{32}$$

Proof. Substitute Eq. equation 30 into the score expectation. The residual is $K^{-1} z_{q,t}^\top e_{i,t}$, whose absolute value is at most $K^{-1}\|z_{q,t}\|_2\|e_{i,t}\|_2 = \delta_{i,t}/\sqrt{K}$. The pairwise statement follows by subtracting the two intervals. □

Equation equation 32 explains what a non-IID claim would need: the audited behavior must separate the clients' realized mixtures by a margin larger than the error of the support model. It does not imply monotonic performance in a Dirichlet concentration parameter. Experiments should therefore report realized client-label/task mixtures in addition to the nominal Dirichlet parameter. A single realized mask does not identify the conditional expectation or $\delta_{i,t}$. If this interpretation is used, estimate B_t by prespecified, regularized regression on training clients/seeds and evaluate on held-out clients/seeds. Report predictive R^2, the distribution of $\|z_{i,t} - \widehat{B}_t \pi_i\|_2/\sqrt{K}$, and the fraction of held-out pairs satisfying the empirical analogue of Eq. equation 32. These are falsification diagnostics, not a proof of the expectation model; without them, Eq. equation 31 should not be invoked to explain the heterogeneity results.

C.9 Random sign corruption

Proposition 4 (Attenuation under independent sign flips)**.** *Condition on fixed supports. Suppose every selected client sign is independently multiplied by a Rademacher variable with flip probability* p_i*, and every query sign by an independent variable with flip probability* p_q*. Then*

$$\mathbb{E}[\widetilde{s}_i \mid z_i, z_q] = (1 - 2p_i)(1 - 2p_q) s_i. \tag{33}$$

If only one side is corrupted, the other factor equals one.

Proof. Each agreeing/disagreeing product is multiplied by two independent random signs whose expectations are $1 - 2p_i$ and $1 - 2p_q$. Linearity of expectation then gives Eq. equation 33. □

If every client has the same flip rate $p_i = p < 1/2$ and the query-side factor $1 - 2p_q$ is also positive and common, all scores receive the same positive attenuation and rankings are preserved in expectation while margins shrink. Client-dependent p_i can reorder expected scores even when every $p_i < 1/2$. The result also does not cover strategic support fabrication, coordinated attacks, or a signature inconsistent with the securely aggregated update. Such integrity requires a concrete authenticated or verifiable-computation mechanism and is not supplied by the estimator.

C.10 Tail non-identifiability and retained-support evasion

The interaction tail in Proposition 1 is an evaluation quantity, not a certificate available from a signed mask.

Proposition 5 (The signed mask does not identify the tail)**.** *Let* $K < M$*, fix a signed mask* z_i *with support* S_i*, and suppose* $g_{q,\ell} \neq 0$ *for some* $\ell \notin S_i$*. For any* $B > 0$ *and* $\alpha \in (0, 1)$*, define*

$$\Delta_i^{(0)} = B z_i, \qquad \Delta_i^{(1)} = B z_i + \alpha B \, \operatorname{sgn}(g_{q,\ell}) e_\ell. \tag{34}$$

Then $m_K(\Delta_i^{(0)}) = m_K(\Delta_i^{(1)}) = z_i$*, so the two updates have the same* AUDITRACE *score for every published query mask, but*

$$a_i \langle \Delta_i^{(1)} - \Delta_i^{(0)}, g_q \rangle = a_i \alpha B |g_{q,\ell}|, \tag{35}$$

and their discarded interaction tails differ by the same amount.

Table 6: **Empirical fidelity on the independent instrumented cohort.** Values are mean ± standard deviation across three independent training and partition seeds.

Metric	Mean ± SD
Spearman ρ_{S}	0.949 ± 0.007
Kendall τ	0.883 ± 0.012
Pearson r	0.944 ± 0.014
Sign accuracy	0.952 ± 0.018
positive-effect AUROC	0.990 ± 0.009
positive-effect AUPRC	0.971 ± 0.037
NDCG@5	0.964 ± 0.043
Top-1 accuracy	0.854 ± 0.144

Proof. Every coordinate in S_i has magnitude B, whereas the added coordinate has magnitude $\alpha B < B$; hence deterministic Top-K extraction returns the same signed support. Because $\ell \notin S_i$, it is also outside $O_i = S_i \cap S_q$. The score is therefore unchanged, while direct substitution gives Eq. equation 35 and the corresponding tail increase. □

Absent an update-norm constraint, B makes the hidden influence gap arbitrarily large. Clipping bounds the gap but does not make the realized τ_i identifiable from z_i. In particular, when signs on O_i balance, the published score can be near zero while aligned mass accumulates outside O_i. A Top-K consistency proof prevents signature fabrication but neither certifies a small query-dependent tail nor guarantees robustness to adaptive retained-support evasion.

C.11 Communication accounting

There are $\binom{M}{K}2^K$ possible signed supports of size K. Hence an information-theoretically minimal fixed-length lossless representation needs

$$B_{\min} = \left\lceil \log_2 \binom{M}{K} \right\rceil + K \quad \text{bits.} \tag{36}$$

A direct bit-packed representation with one fixed-width identifier and one sign bit per selected coordinate uses

$$B_{\text{fixed}} = K\big(\lceil \log_2 M \rceil + 1\big) \quad \text{bits.} \tag{37}$$

Equation equation 37 is an encoding cost, not a lower bound for an unordered set. Both formulas exclude framing, authentication, transport, and implementation overhead. Measured serialized bytes therefore characterize signature storage/serialization cost, not end-to-end network or protocol cost; we make no end-to-end communication claim from these quantities.

C.12 Empirical context for the conditional analysis

We evaluate the sparse attribution signal on a separate instrumented cohort comprising three independent AG News training and partition seeds. For each run, we instrument rounds $\{2, 6, 10\}$, evaluate Top-K budgets of $\{1\%, 5\%, 10\%\}$, reserve 25% of the audit examples for calibration, and evaluate on 16 held-out queries per run. These measurements provide empirical context for the conditional analysis rather than a certification that the sufficient conditions of Proposition 1 hold.

Across the three independent runs, signed sparse attribution remains strongly aligned with the dense first-order reference (Table 6). Mean Spearman correlation is 0.949 ± 0.007, with sign accuracy 0.952 ± 0.018 and positive-effect AUROC 0.990 ± 0.009. Kendall correlation, Pearson correlation, AUPRC, and NDCG@5 are similarly stable across the instrumented runs.

The same cohort also shows improved decision agreement as the retained signature budget increases. Empirical effect-sign agreement rises from 0.815 at $K = 1\%$ to 0.898 at $K = 5\%$ and 0.922 at $K = 10\%$. Pairwise client-ordering agreement similarly increases from 0.759 to 0.888 and 0.927, respectively. Thus, larger retained supports improve both effect-direction and pairwise-order recovery in this diagnostic cohort.

Table 7: Data and audit protocol. A/B identify cohorts, not independent seed replicates. Training counts are configured caps. Task-balanced sampling need not remain balanced after a smaller query limit.

Dataset/cohort	Train cap	Attribution queries / cap	Clients	Rounds	Partition	Audits
SuperNI A	60,000	500	30	20	Task / overlap 3	5,10,15,20
BeaverTails A	27,000	500	30	20	Task / overlap 3	5,10,15,20
AG News A	20,000	100	10	10	IID / Dirichlet	5,10
SuperNI B	60,000	100	30	20	Task / overlap 3	5,10,15,20
BeaverTails B	27,000	200 unsafe cap	30	20	Task / overlap 3	5,10,15,20
AG News B	20,000	100	10	10	LOO: Dir. .3; IID; Dir. .1	5,10

These results are consistent with the empirical objective of the method: signed sparse records preserve client direction and ordering well across independent runs and retention budgets. They are not interpreted as verification of the sufficient conditions in Proposition 1; the theoretical result remains a conditional characterization of when sparse signed attribution approximates dense first-order effects. Exact leave-one-client-out agreement and client-removal outcomes are evaluated separately in the experimental sections.

D Extended Experimental Evaluation

D.1 Experimental Setup and Evaluation Protocol

Unless otherwise stated, our experiments use a controlled FedAvg setting with full participation and protocol-following clients. Different subsections use distinct experimental cohorts when required by the corresponding objective (e.g., generalization, fixed-budget sensitivity, leakage, theory diagnostics, intervention, and robustness). Results are directly comparable across tables only when the dataset, model, partition, query set, checkpoints, and reference target are matched.

D.1.1 Datasets, partitions, and audit queries

SuperNI and BeaverTails use task-based partitions with overlap three. AG News uses IID or label-Dirichlet allocation; IID does not imply equal client sizes. All clients participate at each round. Table 7 distinguishes configured query caps from the actual evaluated query counts. The separate 64-query motivation diagnostic in Figure 1 is not included as an evaluation cohort in this table. AG News queries are class balanced. BeaverTails B uses up to 200 unsafe queries for the attribution evaluation. AG News and SuperNI audit task loss; BeaverTails B audits unsafe-response target log-likelihood.

The AG News exact-LOO evaluation uses Llama-3.2-1B-Instruct with rank-16 LoRA on the q/k/v/o projections and is repeated over five paired training/partition seeds. Only LoRA A/B factors are trainable; the frozen backbone and all non-LoRA parameters remain fixed. For each seed, we train the full ten-client model and the corresponding leave-one-client-out models from the same seed-specific initialization, preserving all training settings other than the removed client. It is not the same partition as the original Dirichlet-0.3 fidelity experiment.

D.1.2 Models and optimization

We optimize LoRA parameters using AdamW with learning rate 2×10^{-4}, zero weight decay, gradient clipping at one, 3% warm-up, one local epoch, and dropout 0.05. We use RoBERTa-base, Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, and Qwen2.5-3B-Instruct with their standard pretrained configurations. Llama and Qwen experiments use LoRA on the attention q/k/v/o projections with rank $r = 16$ and scaling $\alpha = 32$; their trainable and audited coordinate universes contain only LoRA parameters. RoBERTa experiments use rank-8 LoRA with $\alpha = 16$ on the query and value projections. In the RoBERTa AG News runs, the task-specific four-class classification head is also trained and included in the trainable/audited coordinate universe. Maximum sequence lengths are 768 for the Llama/Qwen experiments and 256 for RoBERTa. Language-model runs use local batch

Table 8: Signed-overlap results. A/B denote the cohorts in Appx. D.1.1. All rows use their recorded dense contribution target. CI is a 95% query-bootstrap interval. N@5 is positive-effect NDCG@5.

Setting	ρ_S	95% CI	τ	Pearson	Sign	AUROC	AUPRC	N@5
SuperNI / Llama-1B	0.903	[0.897,0.909]	0.772	0.914	0.893	0.956	0.958	0.910
SuperNI / Llama-3B	0.917	[0.912,0.921]	0.790	0.924	0.904	0.965	0.968	0.917
SuperNI / Qwen-3B	0.919	[0.914,0.923]	0.791	0.925	0.904	0.967	0.970	0.918
BeaverTails / Llama-3B	0.886	[0.881,0.892]	0.750	0.864	0.877	0.969	0.934	0.946
AG News / Dir. $(.5)$	0.952	[0.945,0.958]	0.872	0.966	0.961	0.993	0.994	0.983
AG News / Dir. $(.3)$	0.914	[0.901,0.926]	0.801	0.944	0.937	0.992	0.996	0.973
AG News / Dir. $(.1)$	0.932	[0.921,0.942]	0.844	0.929	0.919	0.972	0.980	0.940

size 8 and two accumulation steps; RoBERTa uses batch size 16 without accumulation. Runs use full client participation, deterministic training, FedAvg, and no DP noise. Language models use BF16 and RoBERTa uses FP32. Model sizes refer to frozen backbones; payloads concern the trainable parameter universe M, which is taken from the saved parameter layout.

D.1.3 Attribution protocol and sign convention

For objective J and actual local update $\Delta_{i,t}$, the dense reference uses the same FedAvg and checkpoint weighting as the sparse estimator:

$$D_i(q) = \frac{\sum_{t\in\mathcal{T}} \omega_t \mathbf{1}\{i \in \mathcal{C}_t\} a_{i,t} \langle \Delta_{i,t}, \nabla J(q;\theta_t)\rangle}{\sum_{t\in\mathcal{T}} \omega_t \mathbf{1}\{i \in \mathcal{C}_t\}}. \tag{38}$$

Here $a_{i,t}$ is the FedAvg client coefficient and $\omega_t = 1$ in the reported experiments. Positive $D_i(q)$ predicts an increase in J. Signed overlap uses the same client and checkpoint weighting when evaluated against this target. For exact LOO, the reference effect is $L_i(q) = J_q(\theta_{\text{full}}) - J_q(\theta_{-i})$, where θ_{-i} is retrained under the same seed and configuration without client i; thus positive $L_i(q)$ means that including client i increases the oriented objective.

Objectives follow Appx. D.1.1; the original BeaverTails table uses its recorded dense-reference objective. Agreement with an objective gradient must not be interpreted as measured positive-effect generation. Likewise, data-source AUROC and effect-sign AUROC have different labels; only the latter is reported as positive-effect AUROC here. The evaluation harness retains dense client values for offline oracle and privileged-baseline computation. This instrumentation is separate from the proposed sparse audit interface and does not constitute an end-to-end cryptographic secure-aggregation experiment. Accordingly, our SecAgg claim is an interface-compatibility claim: AUDITRACE does not require the auditor to observe the individual dense updates that a deployment would protect through its learning-plane secure-aggregation protocol.

D.1.4 Primary signed estimator

Table 8 reports complementary metrics without pooling cohorts. High AUROC can coexist with lower top-one recovery because detecting the effect sign is easier than identifying the largest effect. LOO results replace the original LOO values in the primary comparison; the earlier LOO run remains a separate historical experiment, not an additional seed replicate.

D.1.5 External value-aware baselines

External value-aware baselines are compared only on experimental lattices for which matched client-level exports are available. The primary comparison is the exact AG News leave-one-client-out experiment in Table 2, where all reported methods are evaluated against the identical 100-query, 10-client counterfactual target. We do not extrapolate DataInf-client or client-Fisher results from this experiment to models, tasks, or partitions on which those baselines were not evaluated. TRAK remains excluded from quantitative comparison until its exported-score orientation relative to the LOO objective is fixed independently of test performance.

D.2 External baseline implementation

All completed AG News baseline scores are evaluated on the same 100 queries, 10 clients, and checkpoints 5 and 10 as the exact-LOO run. These client-level adaptations should not be described as official implementations of the sample-level methods (Kwon et al., 2024; Park et al., 2023; Koh & Liang, 2017).

DataInf. The exported `datainf_client` baseline applies a Sherman–Morrison approximation with clients as the units and damping 0.01. We label it DataInf-client; its 0.881 LOO correlation is evidence for this adaptation.

Influence functions. The completed influence baseline solves a damped client empirical-Fisher system in client space; it is not a full-Hessian influence computation. With damping 0.01, conjugate gradient uses 9.95 iterations on average and has maximum reported relative residual 9.54×10^{-7}.

Metrics and statistical inference. We compute Spearman and Kendall correlations over clients within each query, then average over queries. Sign accuracy compares predicted and reference effect signs. Positive-effect AUROC/AUPRC uses positive reference effects as positives; it does not assess client data provenance. positive-effect and negative-effect NDCG@5 rank positive effects and negative effects in the appropriate direction, respectively; full-list NDCG is not substituted for NDCG@5. Top-one accuracy requires the exact largest-effect client. Undefined metrics are excluded only under the evaluator's validity rule, with their valid-query counts retained.

Archive A intervals are the exported query-bootstrap intervals. Reported new correlation intervals use 2,000 query-bootstrap replicates. Paired differences resample identical query IDs for both methods. Intervention intervals resample held-out queries, and comparisons to random removal additionally resample the five subset draws when stated. These intervals quantify query-level uncertainty conditional on the evaluated runs; query–client pairs are not independent training replicates.

D.2.1 Counterfactual validation and intervention

Exact LOO retraining on AG News/Llama-3.2-1B-Instruct removes each client from a common initialization, yielding 1,000 query–client effects (Table 2). Signed overlap achieves positive-effect AUROC 0.941 and NDCG@5 0.893, nearly matching dense cosine (0.952 and 0.894). However, positive-effect top-one recovery is only 39%, indicating strong ranking fidelity without reliable identification of the single largest effect. On BeaverTails/Llama-1B, clients are ranked using 99 unsafe audit examples and evaluated after removal on 99 disjoint unsafe examples plus 11 safe utility examples. As shown in Table 9, increasing the number of monotonically improves the held-out unsafe-target objective from 0.0036 to 0.0584, with a corresponding utility trade-off. At the common $K = 5$ intervention budget, AUDITRACE improves the held-out unsafe-target objective by 0.0232, versus 0.0230 for dense-oracle selection, 0.0129 for random removal, and -0.0285 for bottom-five removal. The difference from the dense oracle is 0.00023 (paired query-bootstrap 95% CI $[-0.00209, 0.00242]$), while AUDITRACE exceeds random removal by 0.01033 (query/subset-bootstrap CI $\approx [0.00380, 0.01843]$) and bottom-five removal by 0.05172 (95% CI $[0.03403, 0.07150]$).

Intervention protocol BeaverTails/Llama-1B uses disjoint saved query IDs for 99 unsafe audit examples, 99 unsafe behavior examples, and 11 safe utility examples. Clients are ranked on the audit subset before retraining; within each training/partition seed, intervention retrainings reuse that seed's initialization and baseline checkpoints. For objective J equal to response target log-likelihood, behavior improvement is $J_{\text{full}} - J_{-S}$ on unsafe responses, whereas utility change is $J_{-S} - J_{\text{full}}$ on safe responses. These quantities retain the implementation's objective normalization and are not generated-response safety rates. The small utility set limits precision.

Signed top-K removal uses prefixes of $[13, 22, 28, 10, 19, 27, 3, 2, 14, 11]$ for $K \in \{1, 3, 5, 10\}$. Oracle top-five selects $[13, 28, 3, 22, 10]$; signed bottom-five selects $[29, 6, 21, 9, 4]$. Bottom-five denotes lowest scores, not five necessarily negative-effect clients. Five random subset draws use seeds 923420–923424. Their behavior improvements have mean 0.01288 and between-subset standard deviation 0.00727; utility changes have mean -0.00974 and standard deviation 0.00489.

At $K = 5$, signed minus oracle behavior improvement is 0.00023 (paired query-bootstrap 95% CI $[-0.00209, 0.00242]$); similar point estimates do not prove equivalence. Signed exceeds mean random removal by 0.01033 (query/subset-bootstrap CI approximately $[0.00380, 0.01843]$) and bottom-

Table 9: **Client-removal intervention on held-out BeaverTails queries.** B is full minus intervened unsafe-response log-likelihood (higher is better); U is intervened minus full safe-response log-likelihood (higher is better). The first panel varies the number of clients selected by AUDITRACE. The second panel compares selectors at the common budget $K = 5$. Random is averaged over five independently sampled subsets.

Selection	Removed clients	Behavior improvement B	Utility change U
AUDITRACE *removal curve*			
Signed top-K	1	0.0036	-0.0050
Signed top-K	3	0.0127	-0.0132
Signed top-K	5	0.0232	-0.0205
Signed top-K	10	0.0584	-0.0507
Selector comparison at $K = 5$			
Signed top-K	5	0.0232	-0.0205
Dense oracle	5	0.0230	-0.0185
Random (5 subsets)	5	0.0129	-0.0097
Signed bottom-K	5	-0.0285	-0.0510

Table 10: **Ideal signed-only payload per client and audited round at 10%.** Compression factors compare this encoding with dense FP32 or FP16/BF16 trainable vectors; they do not represent network measurements or savings in the dense learning plane.

Backbone	M	K	Index bits	MiB	vs. FP32	vs. FP16
RoBERTa-base	888,580	88,858	20	0.222	15.24	7.62
Llama-3.2-1B	3,407,872	340,788	22	0.934	13.91	6.96
Llama-3.2-3B	9,175,040	917,504	24	2.734	12.80	6.40
Qwen2.5-3B	7,372,800	737,280	23	2.109	13.33	6.67

five removal by 0.05172 (query-bootstrap CI $[0.03403, 0.07150]$). The signed top-five utility change is -0.02049, versus -0.01852 for oracle selection. Thus, behavior improvement is accompanied by measurable objective trade-offs. All selector comparisons are averaged over five training/partition seeds; within each seed, the random baseline additionally averages five independently sampled removal subsets.

D.3 Systems Cost and Client Scalability

D.3.1 Idealized and measured signature storage

An unordered signed signature with K coordinates from a universe of size M requires, under fixed-width coordinate encoding,

$$B_{\text{signed}} = K(\lceil \log_2 M \rceil + 1) \quad \text{bits.} \tag{39}$$

This excludes framing and transport overhead and is not an information-theoretic lower bound. Serialized files may use compression and additional fields. In the logs, the nominal payload includes two rank bits per coordinate (25 bits for Llama-1B rather than the signed-only 23), although signed overlap does not use ranks. Therefore logged nominal compression, ideal signed-only compression, and measured compressed-file size are reported separately. For example, the measured Llama-1B/SuperNI value of 0.790 MiB is the same serialized-signature measurement reported as 809.0 KiB in Table 4; the 0.934 MiB value in Table 10 is instead the analytical fixed-width signed-only payload.

D.3.2 Scaling to 300 Federated Clients

We evaluate client-count scalability using RoBERTa-base with rank-8 LoRA on a fixed 20,000-example AG News cohort. We vary the number of logical federated clients from 50 to 300 while holding the training data, query set, training seed, partition seed, ten communication rounds, and fixed streaming aggregation implementation. Audit records use $K/M = 10\%$ and are collected at rounds 5 and 10. Because the total statistical workload is fixed, increasing the client count reduces the mean

Table 11: **Measured audit-signature systems costs.** Available 10% five-run measurements. Serialized size and extraction time are means over client signatures; storage covers saved signature files. Peak is the reported attribution-process CUDA allocation.

Setting	MiB/signature	Extract ms	Store GiB	Peak GiB
Llama-1B / SuperNI	0.790	228.5	0.0926	11.60
Llama-1B / BeaverTails	0.790	225.7	0.0926	7.62
RoBERTa / AG News IID	0.205	58.3	0.0040	0.94
RoBERTa / AG News Dir. .1	0.203	54.3	0.0040	0.94

Table 12: **Client-count scalability under a fixed 20,000-example workload.** Attribution metrics are computed over the same 100 audit queries. The experiment varies the number of logical federated clients while holding the training cohort, query set, training seed, partition seed, communication rounds, and implementation fixed.

Clients	Ex./client	Label JS	Task Acc.	ρ_S	Sign Acc.	positive-effect AUROC	positive-effect NDCG@5	Attr. (min)
50	400.0	.673	.890	.941	.921	.979	.864	0.80
100	200.0	.608	.878	.895	.885	.965	.841	1.05
150	133.3	.564	.866	.951	.927	.980	.932	1.26
200	100.0	.533	.852	.949	.926	.982	.892	1.68
250	80.0	.498	.838	.960	.923	.981	.932	2.04
300	66.7	.469	.836	.961	.910	.978	.926	2.29

number of examples per client from 400 to 66.7. These experiments simulate clients using a local two-GPU worker pool and therefore measure client-granularity computation rather than wide-area network latency.

Attribution quality. Attribution remains reliable throughout the six-fold increase in client count. Spearman correlation never falls below 0.895 and reaches 0.961 at 300 clients. Positive-effect AUROC remains between 0.965 and 0.982, while sign accuracy remains above 0.885. At 300 clients, AUDITRACE obtains Spearman correlation 0.961, positive-effect AUROC 0.978, and positive-effect NDCG@5 0.926. Hence, increasing the candidate population does not induce a systematic deterioration in agreement with the dense first-order reference. Task accuracy decreases from 0.890 to 0.836, reflecting the increasingly fine partition of a fixed training cohort rather than degradation of the attribution procedure.

The realized partitions become more balanced as the fixed 20,000 examples approach the minimum of 64 examples per client: mean pairwise label Jensen–Shannon divergence decreases from 0.673 to 0.469. We therefore interpret this experiment as a fixed-workload systems stress test and do not attribute changes in ranking quality solely to client count.

Runtime scaling. For the hardware-matched 50–200-client runs, all executed on two A100-SXM4-40GB GPUs, increasing the client population by $4\times$ increases training time from 8.19 to 10.12 minutes ($1.24\times$) and complete pipeline time from 9.05 to 11.86 minutes ($1.31\times$). Log–log fits give exponents of 0.148 for training time ($R^2 = 0.989$) and 0.187 for total pipeline time ($R^2 = 0.977$). This sublinear wall-clock growth is expected under the fixed-workload regime: every configuration processes 200,000 examples over ten rounds.

Attribution time increases from 0.80 to 1.68 minutes between 50 and 200 clients, corresponding to an empirical exponent of 0.506 ($R^2 = 0.953$). Over the same range, the number of evaluated query–client pairs grows by $4\times$, from 10,000 to 40,000, while throughput rises from 208 to 398 pairs per second through batching and amortization. The complete 300-client run evaluates 60,000 score pairs in 2.29 minutes.

Table 13: Matched archive-A ablation matrix: Spearman correlation against the dense reference. S1/S3/SQ are SuperNI with Llama-1B/Llama-3B/Qwen-3B; B3 is BeaverTails/Llama-3B. AG columns use RoBERTa. Dense-dot values of one arise from the target definition, not independent validation.

Method	S1	S3	SQ	B3	AG IID	AG .5	AG .3	AG .1
Signed	0.903	0.917	0.919	0.886	0.839	0.952	0.914	0.932
Rank	0.021	0.009	-0.055	-0.276	-0.029	-0.085	0.074	0.086
Signed-rank	0.904	0.918	0.919	0.877	0.830	0.947	0.904	0.926
Layer-norm signed	0.797	0.819	0.821	0.845	0.585	0.915	0.857	0.932
Layer-norm signed-rank	0.796	0.819	0.821	0.839	0.584	0.912	0.848	0.924
Dense cosine	0.981	0.974	0.975	0.954	0.951	0.982	0.976	0.984
Dense dot / target	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Data size	0.023	0.008	-0.053	-0.277	-0.013	-0.102	0.033	0.076
Random	-0.019	-0.010	-0.010	-0.015	0.018	-0.027	-0.004	-0.020

Table 14: Adaptive-budget diagnostic, archive A. Cap is the fraction of clients whose selected budget equals the maximum. Saturation flags are the recorded diagnostic outputs. All four AG News rows in this table use RoBERTa-base.

Setting	Selected K	K/M	Cap	Coverage saturated	Stability saturated
SuperNI / Llama-1B	340,788	10%	100%	Yes	Yes
SuperNI / Llama-3B	917,504	10%	100%	Yes	Yes
SuperNI / Qwen-3B	737,280	10%	100%	Yes	Yes
BeaverTails / Llama-3B	917,504	10%	100%	Yes	Yes
AG News / IID	88,858	10%	100%	Yes	Yes
AG News / Dir. $(.5)$	88,858	10%	100%	Yes	Yes
AG News / Dir. $(.3)$	88,858	10%	100%	Yes	Yes
AG News / Dir. $(.1)$	88,858	10%	100%	Yes	Yes

E Ablations, Sensitivity, and Robustness

E.1 Component Ablations

We retain the matched original-cohort ablation matrix rather than combining signed scores with competitors (Table 13). Signed overlap is the best sparse variant in five of these eight dense-reference settings and is within 0.0013 of the best in the other three. Adding coordinate-rank information to the signed estimator offers no consistent improvement, while the rank-only variant performs poorly across the evaluated settings. These observations favor the simpler signed representation; they are not a held-out method-selection guarantee.

E.2 Fixed-budget sweep

Runs 15–18 rescore the saved run-14 SuperNI/Llama-1B training trajectory at 1%, 2%, 5%, and 10%, holding client data, checkpoints, and query identities fixed. The corresponding query-bootstrap correlation intervals are approximately $[0.538, 0.599]$, $[0.707, 0.752]$, $[0.844, 0.875]$, and $[0.897, 0.919]$. Measured payloads and fidelity are reported in Table 4.

E.3 Adaptive-budget diagnostic

In archive A, calibration uses three early rounds with $\tau = 0.90$, $\rho_0 = 0.80$, and a 10% cap. All clients reach the cap and the exported coverage/stability saturation flags are active. This is a negative diagnostic: these runs do not demonstrate successful adaptive compression. Archive-B fixed-budget runs do not invoke this selector and are not included in this claim.

Table 15: **Robustness across PEFT parameterizations on RoBERTa-base/AG News.** All variants use the same data partition, training seed, communication schedule, checkpoints, 100 held-out queries, and $K/M = 10\%$ signature budget. The RoBERTa trainable/audited universe includes both the adapter parameters and the task-specific classification head. Attribution metrics use the same dense first-order reference. "Size" is the serialized client-signature size.

Variant	Spearman ρ	Kendall τ	Pearson r	Sign Acc.	AUROC	AUPRC	NDCG@5	Top-1	Task Acc.	Size (KiB)
LoRA	0.957	0.886	0.962	0.938	0.986	0.987	0.967	0.61	0.898	179.2
rsLoRA	0.961	0.894	0.963	0.950	0.990	0.991	0.966	0.60	0.894	179.8
DoRA	0.956	0.887	0.961	0.944	0.984	0.986	0.968	0.61	0.900	183.0
AdaLoRA	0.954	0.884	0.962	0.954	0.989	0.983	0.987	0.85	0.882	207.1
LoRA+	**0.961**	**0.904**	**0.985**	0.954	**0.991**	**0.993**	**0.992**	**0.91**	0.896	176.1

Table 16: **Sensitivity to LoRA rank on RoBERTa-base/AG News.** The signature fraction is fixed at $K/M = 10\%$. The audited coordinate universe contains the rank-dependent query/value LoRA factors and the shared task-specific classification head. All attribution columns use the same dense first-order reference.

r	Spearman ρ	Kendall τ	Pearson r	Sign Acc.	AUROC	AUPRC	NDCG@5	Top-1	Task Acc.
4	0.958	0.892	0.967	0.957	0.988	0.988	0.966	0.62	0.900
8	0.957	0.886	0.962	0.938	0.986	0.987	0.967	0.61	0.898
16	0.959	0.889	0.955	0.946	0.991	0.993	0.965	0.59	0.898
32	0.954	0.878	0.958	0.950	0.988	0.989	0.964	0.63	0.892

E.4 Robustness Across PEFT Parameterizations

We evaluate sensitivity to the PEFT parameterization on RoBERTa-base fine-tuned on AG News. The RoBERTa configuration contains rank-based attention adapters together with the trainable task-specific classification head in the trainable and audited coordinate universe. We replace standard LoRA with rsLoRA, DoRA, AdaLoRA, and LoRA+ while holding the data partition, training seed, communication schedule, checkpoints, and $K/M = 10\%$ signature budget fixed. We recompute the trainable coordinate space for each parameterization. This is therefore a paired robustness experiment under one controlled RoBERTa/AG News configuration.

The ranking correlation remains between 0.954 and 0.961, positive-effect AUROC remains between 0.984 and 0.991, and task accuracy varies by at most 1.8 percentage points. LoRA+ provides the strongest top-of-ranking performance, reaching NDCG@5 of 0.992 and Top-1 accuracy of 0.91, while the standard LoRA result remains representative of the broader family. These paired results show that overall attribution fidelity remains stable when the adapter parameterization is varied within this controlled RoBERTa/AG News configuration. Because the task-specific classification head is shared across variants and remains part of the audited coordinate universe, this experiment should not be interpreted as isolating adapter coordinates alone. We do not include QLoRA in this controlled comparison because it changes the numerical precision and memory path of the frozen backbone in addition to the adapter parameterization.

E.5 Sensitivity to LoRA rank

We evaluate sensitivity to LoRA rank using the same RoBERTa-base/AG News configuration as the PEFT-parameterization study above. We sweep $r \in \{4, 8, 16, 32\}$ while fixing the signature fraction at $K/M = 10\%$. The query/value LoRA factors vary with rank, while the task-specific classification head remains trainable and part of the audited coordinate universe. The $r = 8$ row is the same standard-LoRA configuration reported in Table 15.

Overall dense-reference fidelity changes little across the evaluated LoRA ranks in this RoBERTa/AG News configuration: Spearman correlation varies by less than 0.005 and positive-effect AUROC remains above 0.986. Because the classification head is shared across ranks and remains in the audited coordinate universe, these results characterize robustness of the complete RoBERTa adaptation configuration rather than a LoRA-only rank effect.

Table 17: **Temporal fidelity across audited rounds.** SuperNI/Llama-3B, archive A. The aggregate is reported separately because its target differs from the round-wise targets.

Round	ρ_S	AUROC	Sign acc.
5	0.901	0.959	0.891
10	0.907	0.960	0.898
15	0.894	0.956	0.892
20	0.894	0.954	0.884
Aggregate	0.917	0.965	0.904

Table 18: Attribution utility of signed audit records across Top-K budgets. Values are mean $\pm$ 95% confidence interval over 24 independent FL runs.

K	Spearman ρ	Kendall τ	Pearson r	Magnitude NDCG	Sign accuracy	positive-effect AUPRC
1%	0.565±0.051	0.283±0.042	0.380±0.050	0.927±0.006	0.858±0.012	0.774±0.027
5%	0.735±0.024	0.610±0.023	0.765±0.023	0.968±0.003	0.909±0.009	0.933±0.014
10%	0.770±0.025	0.653±0.023	0.805±0.025	0.976±0.002	0.906±0.008	0.963±0.011

E.6 Temporal aggregation

For SuperNI/Llama-3B, individual-round correlations remain between 0.894 and 0.907, while the aggregate reaches 0.917. Because individual and aggregate targets differ, this is a temporal-stability diagnostic, not a controlled comparison of estimators against a common target.

F Privacy, Leakage, and Security

All leakage experiments in this section use Llama-3.2-1B-Instruct with a strictly LoRA-only trainable and audited coordinate universe; the released signatures contain no classifier-head or other non-LoRA coordinates.

Evaluated audit-record configuration. All leakage experiments use Llama-3.2-1B-Instruct with rank-16 LoRA ($\alpha = 32$) on the attention q/k/v/o projections. The trainable and audited coordinate universe consists exclusively of these LoRA parameters. In particular, the signatures used for the membership and client-property attacks contain no classifier-head or other task-specific dense parameters.

F.1 Attribution Utility Across Top-K Budgets

On the separate 24-run leakage cohort, we evaluate how faithfully the signed audit records approximate dense-update attribution as the retained Top-K fraction varies. Table 18 reports means and 95% confidence intervals over 24 independent federated runs. Even at $K = 1\%$, the signed records retain substantial directional information, achieving a magnitude NDCG of 0.927 and sign accuracy of 0.858. Increasing the budget to 5% improves rank fidelity: Spearman correlation rises from 0.565 to 0.735, while positive-effect AUPRC increases from 0.774 to 0.933. At $K = 10\%$, the corresponding values reach 0.770 and 0.963.

Moving from 1% to 5% improves Spearman correlation by 0.170, whereas increasing the budget from 5% to 10% adds only 0.035. Sign recovery similarly stabilizes near 91% beyond $K = 5\%$. Thus, the larger fidelity gain occurs between 1% and 5%, with diminishing gains from 5% to 10%.

Adversarial leakage evaluation protocol. We evaluate two attacker objectives: candidate-conditioned example membership and inference of a client's high versus low label-2 proportion. Both use run-disjoint evaluation, with complete federated runs held out from attacker training and model selection. Attackers are selected by nested validation from logistic regression, a small MLP, and histogram-gradient boosting. Attackers use only the information in the stated support-only, final-checkpoint signed, repeated-signed, or dense-assisted view at each signature budget. The

Table 19: Candidate-conditioned membership inference across audit budgets. Each entry reports AUROC/AUPRC. The no-audit baseline is $0.500/0.200$. Top-1 uses tie-aware ranking among the owner and four label-mixture-matched nonowners. Dense-assisted is a richer offline attacker view and is not interpreted as an upper bound on signed-record leakage.

K	Support only	Signed, final	Signed, repeated	Dense-assisted, repeated	Top-1, signed repeated
1%	$0.503/0.203$	$0.517/0.208$	$0.523/0.211$	$0.568/0.235$	0.214
5%	$0.499/0.200$	$0.544/0.221$	$\mathbf{0.567/0.240}$	$0.562/0.229$	$\mathbf{0.237}$
10%	$0.503/0.200$	$0.558/0.228$	$0.565/0.231$	$0.569/0.235$	0.233

repeated-signed view tests temporal accumulation across audited checkpoints, while the dense-assisted view is a richer offline comparison rather than an upper bound on audit-record leakage. For the client-property attack, we report the strongest exposed view selected independently at each budget.

F.2 Candidate-Conditioned Membership Leakage

We evaluate membership leakage under a candidate-conditioned threat model. For each candidate example, the attacker distinguishes its true owner from four nonowner clients selected by nearest label-mixture matching. Every candidate is therefore evaluated under identical content, eliminating shortcuts based on candidate identity or class. The evaluation contains 6,113 candidates and 30,565 candidate–client pairs per audit view and budget. Attackers are selected using nested validation among logistic regression, a small MLP, and histogram-gradient boosting, while all outer splits hold out complete FL runs. Random-guess baselines are 0.5 AUROC, 0.2 AUPRC, and 20% Top-1 accuracy. This protocol is intentionally stronger than a random owner/nonowner split: matching nonowners by label mixture removes trivial label-distribution shortcuts, while holding out complete FL runs tests whether attack performance transfers beyond the runs used for attacker fitting and model selection.

Table 19 shows that support locations alone provide essentially chance-level membership discrimination: AUROC remains between 0.499 and 0.503 across all budgets. Signed records remain close to chance in absolute terms but introduce a small and statistically reproducible advantage. At $K = 5\%$, repeated signed observations obtain AUROC 0.567 and AUPRC 0.240, with 95% CIs of $[0.559, 0.574]$ and $[0.234, 0.247]$, respectively. Candidate-level Top-1 identification reaches 23.7% ($[22.6, 24.9]$), compared with the 20% chance rate. The AUROC exceeds chance in each of the 24 held-out runs, indicating that the small effect is consistent across runs rather than being driven by a small subset of partitions.

Controlled view comparisons isolate the source of this leakage. Relative to support alone, adding signs improves AUROC by 0.045 at $K = 5\%$ (95% CI $[0.032, 0.061]$) and by 0.056 at $K = 10\%$ ($[0.045, 0.067]$). Repeated observations yield a further 0.023 AUROC and 0.020 AUPRC improvement at $K = 5\%$. In contrast, support-only records remain indistinguishable from the no-audit baseline. These results characterize an empirical privacy–utility trade-off under the evaluated attack: signed records enable high-fidelity attribution while membership discrimination remains low in absolute terms; signs and repeated observations nevertheless produce a small reproducible advantage over support-only records.

Repeated-signed membership attack performance is not monotone in the audit budget: the 5% and 10% point estimates are nearly identical (0.567 and 0.565). We therefore do not interpret larger K as producing monotonically greater membership leakage under this finite-sample attack. This does not imply that dense information is intrinsically less informative: the views use separately trained and nested-selected finite-sample attack models, for which held-out performance need not be monotonic with additional features. We therefore treat the dense-assisted result as an additional attack configuration rather than a leakage upper bound.

F.3 Run-Disjoint Client-Property Inference

We additionally evaluate a run-disjoint client-property inference attack targeting whether a client's local data contain a high versus low proportion of label 2 examples. The two property classes contain

Table 20: Inference of high versus low client label-2 proportion. For each budget, we report the maximum nested-selected result over audit views; AUROC and AUPRC are maximized independently. Brackets denote 95% confidence intervals over held-out FL runs.

K	Max. AUROC	Corresponding view	Max. AUPRC
1%	$0.577, [0.494, 0.645]$	Signed, final	$0.596, [0.524, 0.670]$
5%	$0.554, [0.491, 0.623]$	Dense + signed, repeated	$0.536, [0.492, 0.607]$
10%	$0.559, [0.490, 0.628]$	Support only	$0.562, [0.497, 0.637]$

equal-size clients and are balanced within each federated run, preventing client size or run identity from serving as shortcuts. We use the same run-disjoint nested model-selection protocol as in the membership experiment and evaluate all exposed audit views before constructing the attack envelope.

Table 20 reports a conservative attack envelope obtained by maximizing the nested-selected attack independently over all exposed audit views at each budget. Even under this worst-view summary, AUROC does not exceed 0.577, and every AUROC confidence interval contains the 0.5 chance level. Increasing the exposed budget does not strengthen the attack: the maximum AUROC is 0.577, 0.554, and 0.559 at budgets of 1%, 5%, and 10%, respectively. The isolated AUPRC increase at 1% is not accompanied by significant AUROC separation and does not persist as the budget grows. Under this attack and tested property, we find no reliable evidence of a stable or budget-dependent inference signal.

Taken together, the two evaluations expose different aspects of the audit channel under a common run-disjoint adversarial protocol. Candidate-conditioned membership discrimination remains close to chance in absolute terms: support-only AUROC is approximately 0.5, while the strongest repeated signed view reaches 0.567. Signs and repeated observations therefore provide a small but reproducible membership advantage rather than high attack success. The client-property attack, evaluated through a nested-selected envelope over exposed views, does not show reliable AUROC separation from chance for the tested property. These results characterize the tested disclosure channels but do not establish a general privacy guarantee or resistance to generative reconstruction attacks.

G Limitations and Scope

AUDITRACE is designed for post-hoc, behavior-conditioned client attribution under a sparse signed audit interface. Its scores should therefore be interpreted as relative ranking signals for a specified differentiable audit objective rather than as general-purpose measures of client importance. Our experiments evaluate this setting across multiple models, datasets, client partitions, PEFT configurations, and client scales, while broader deployment conditions such as partial participation, differential-privacy mechanisms, and additional federated optimization protocols remain natural directions for future study.

The magnitude-free representation intentionally trades numerical detail for a smaller and less revealing audit interface. As a result, attribution quality depends on how much behavior-relevant information is retained by the selected Top-K coordinates. The theoretical analysis characterizes sufficient conditions for agreement with dense first-order effects, while the empirical results evaluate this relationship across the settings considered in the paper. Exploring adaptive or task-dependent signature budgets may further improve this trade-off.

The audit record is an additional client-linked disclosure and is not intended to provide a standalone privacy guarantee. Our leakage experiments evaluate membership and client-property inference under run-disjoint attacks and show limited but measurable disclosure in the tested settings. Extending this analysis to additional attack models and privacy mechanisms would provide a broader characterization of the audit interface.

Finally, the zero-knowledge construction is evaluated as a proof-of-concept for audit-record consistency. Integrating the mechanism with full-scale verifiable-input secure aggregation and benchmarking it at the dimensions of modern PEFT updates are systems-level extensions beyond the present study.