FEDATTR: TOWARDS PRIVACY-PRESERVING CLIENT-LEVEL ATTRIBUTION IN FEDERATED LLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data owners often need a way to verify whether a released corpus is used for unauthorized model training purposes by users. Training-data watermarking enables such verification via modifying the corpus so that models trained on it exhibit a detectable signal. In federated learning (FL), the global model aggregates updates from multiple clients, thus detecting the training-data watermark in the global model is feasible to check the corpus usage, but cannot identify which clients used it, which is the task we call *client-level watermark attribution*. Testing each client's locally updated model individually would reveal such attribution, but would also expose the client's update with privacy concern. Secure aggregation (SA) protects the individual updates via only exposing their aggregation but hides client-specific detection information which is needed for attribution, *i.e.* preventing the data owner from identifying which clients used the corpus. We propose **FedAttr**, a client-level watermark attribution protocol for FL with using the secure aggregates. For each target client, FedAttr contrasts aggregates from random subsets that include or exclude the client, yielding an unbiased update estimate with bounded variance. It then applies differential watermark scoring and combines evidence across communication rounds. We prove exponential attribution error bounds and bound the estimator's per-round mutual-information leakage by $O(d^*/N)$. In federated LoRA fine-tuning, FedAttr achieves 100% TPR at 0% FPR within five rounds across two watermark families and two aggregation strategies, outperforming all baselines, with 6.3% overhead.

1 INTRODUCTION

High-quality training corpora are often proprietary or distributed under licenses that only permit specific model-training purposes. For example, many public data are allowed for academic use but not for training commercial models. However, once such a corpus is released, the data provider may have no direct way to determine whether the data are later used for unauthorized purposes. One major solution is to modify the corpus before release so that a model trained on the data will exhibit a detectable signal in its outputs (Sablayrolles et al., 2020; Sander et al., 2024; Cui et al., 2025). This approach, known as *training-data watermarking*, allows the data provider to test whether the protected corpus was used to train a model.

In federated learning (FL), multiple clients train locally on their private data, and the server aggregates their updates to train a shared global model (McMahan et al., 2017). If one or more clients utilize the protected corpus, its watermark signal can be detected in the global model. Although a watermark test on the global model could indicate that the corpus was used somewhere in FL, it does not reveal which client or clients used the corpus. We refer to this client-specific task as *client-level watermark attribution*: watermark detection checks whether the protected corpus was used, whereas attribution considers which participating clients used it during the local training.

A natural approach for client-level attribution is to test each client's locally updated model separately. Since every client starts from the same global model at the beginning of a training round, accessing to a client's resulting local model also reveals that client's individual model update. Prior work has shown that gradients and model updates can leak information about private training data,

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.

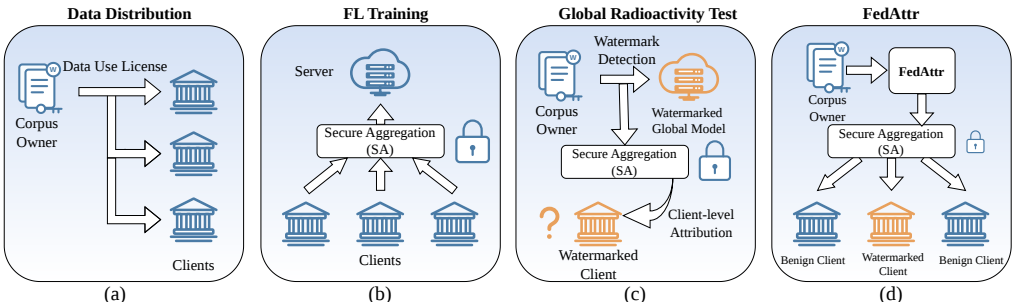

Figure 1: Overview of Data Attribution in FL. **(a)** The corpus owner distributes watermarked documents to clients under a data-use license. **(b)** Clients collaboratively fine-tune a shared model via FL through secure aggregation (SA). **(c)** A global radioactivity test can detect the watermark signal of the trained model, but cannot identify which clients are responsible without violating SA. **(d)** FedAttr identifies which clients use the watermark documents through SA.

including labels and, in some settings, high-fidelity training examples (Zhu et al., 2019; Geiping et al., 2020; Chen & Vikalo, 2024). Thus, directly testing local models would enable attribution only by exposing information that privacy-preserving federated learning is intended to protect.

Secure aggregation (SA) is commonly used to prevent such exposure by allowing the server to aggregate client updates without observing any individual update (Bonawitz et al., 2017). This privacy protection, however, also removes the client-specific information needed for attribution. Under SA, a watermark detector can be applied to the global model, but not separately to each client's locally updated model. This raises the central question of our work: *can we identify which clients used the protected corpus while preserving users' privacy?*

To address this question, we propose **FedAttr**, which performs client-level watermark attribution using only secure aggregates. For each target client, FedAttr forms two collections of secure aggregates: one from random client subsets that contain the target client and one from subsets that do not. Taking the difference between the average aggregates from the two collections cancels the other clients' updates in expectation, yielding an unbiased estimate of the target client's update. FedAttr then compares the watermark score of the current global model with the score after adding this estimate, and combines the resulting score differences across rounds to determine whether the target client trained on the protected corpus.

Contribution. We summarize our contributions as follows.

(1) Problem and protocol. We formalize *client-level attribution* problem and propose **FedAttr**, which combines a client-level update estimator, differential scoring, and cross-round Stouffer combination to decide which client uses the watermarked documents through the SA mechanism.

(2) Theoretical guarantees. We prove the client-level update estimator is unbiased with bounded variance (Theorems 1, 2), and derive two-sided exponential error bounds for cross-round Stouffer combination that drive false negatives to zero in T (Theorem 3).

(3) Privacy analysis. We bound the per-round mutual information leakage of the estimator about each client's update by $O(d^*/N)$, where d^* is the effective subspace dimension and N is the subset size of SA queries (Theorem 4).

(4) Empirical validation. In federated LoRA fine-tuning experiment, FedAttr achieves 100% TPR at 0% FPR within 5 rounds across two watermark families and two aggregation strategies, outperforming all baselines by at least 44.4% in TPR or 19.1% in FPR, with only 6.3% overhead relative to FL training time. Ablation studies confirm that FedAttr is robust to parameters and configurations.

2 Related Work

Federated LLM fine-tuning. Federated learning (McMahan et al., 2017) has been extended to LLM fine-tuning via parameter-efficient adapters such as LoRA (Hu et al., 2022), with aggregation strategies including FedIT (Zhang et al., 2023) and FLoRA (Wang et al., 2024). We evaluate both in our experiments. Detailed descriptions of two aggregation strategies are in Appendix F.

Training-data watermarking. Embedding detectable signals into training data so that downstream models inherit measurable traces, known as *radioactivity* (Sablayrolles et al., 2020), was extended to LLMs in two forms: Sander et al. (2024) generate documents with a watermarked LLM whose green-token bias transfers to models fine-tuned on them, and Cui et al. (2025) inject fabricated entity-attribute pairs that the fine-tuned model memorizes and can be detected via QA probes. These methods assume a single training party; in FL, the global model aggregates all clients' updates, so a detection on the global model no longer identifies which client used the watermarked data. FedAttr reuses these detectors as black-box scoring functions and resolves this attribution problem.

Federated forensics. FLDetector (Zhang et al., 2022) and FLForensics (Jia et al., 2024) trace malicious clients in poisoning attacks by detecting update inconsistency and misclassification influence, respectively. Both require plaintext access to the individual updates and rely on adversarial signals absent in our non-adversarial setting, where clients faithfully follow the FL protocol.

Secure aggregation and privacy. SA protocols (Bonawitz et al., 2017) enable the server to compute subset sums of client updates without observing the individual updates. Elkordy et al. (2023) provide the first mutual information bound on per-round leakage under standard SA protocol. FedAttr's privacy analysis extends this framework to the setting required by client-level attribution.

3 Problem Setup

3.1 Problem Formulation

Federated fine-tuning system. FL enables K clients $\{1, \ldots, K\}$ to collaboratively train a shared global LLM over T communication rounds under the coordination of a central server. At each round t, the server distributes the current global model parameters $w^{t-1} \in \mathbb{R}^d$ to all clients; client i locally fine-tune w^{t-1} on its private dataset $\mathcal{D}_i$ and returns the resulting parameter update $\Delta_i^t \in \mathbb{R}^d$. Then the server aggregates updates into the new global model according to a federated aggregation rule (McMahan et al., 2017):

$$w^t = w^{t-1} + \sum_{i=1}^{K} p_i \Delta_i^t, \tag{1}$$

where aggregation weights $p_i \geq 0$ satisfy $\sum_{i=1}^{K} p_i = 1$, typically $p_i = |\mathcal{D}_i| / \sum_j |\mathcal{D}_j|$. We assume all K clients participate in every communication round.

Secure aggregation. Secure aggregation (SA) aims to preserve client updates' privacy in FL systems (Bonawitz et al., 2017). SA allows the server to compute the sum of client updates $S^t(W)$ over any subset of clients $W \subseteq [K]$ with $|W| \geq N_{\text{sa}}$, while keeping each individual update hidden: $S^t(W) = \sum_{j \in W} \Delta_j^t$. The threshold N_{sa} prevents individual updates from being exposed.[1]

Client-level attribution problem. An unknown set of clients trained on watermarked documents in violation of license terms, i.e., use the watermarked documents. Given access to SA aggregations $\{S^t(W)\}$ over admissible subsets W and rounds $t \in \{1, \ldots, T\}$, the *client-level attribution* problem is to output a per-client binary decision $r_i \in \{0, 1\}$ for each client $i \in [K]$, where $r_i = 1$ iff client i's dataset $\mathcal{D}_i$ contains watermarked documents.

3.2 Watermark Families

FedAttr requires only a scoring function $\textsc{Score}(w; \mathcal{P})$ that returns a larger value when w has been trained on watermarked data, where w is the model under test and $\mathcal{P}$ is a set of evaluation prompts.

KGW watermark (Kirchenbauer et al., 2023). Before distributing the documents, the data provider rephrases them with a watermarked LLM that partitions the vocabulary into green and red lists via a pseudorandom function and boosts green-token logits by δ during decoding. A model fine-tuned on these documents inherits the green-token bias (Sander et al., 2024). To detect this bias, a z-test compares the observed green-token ratio against the expected null rate $\gamma = |G|/|\mathcal{V}|$.

[1] SA can be implemented via multi-party computation (Bonawitz et al., 2017) or homomorphic encryption (Zhang et al., 2020); FedAttr is agnostic to the choice of instantiation.

Fictitious knowledge watermark (Cui et al., 2025). The data provider injects fabricated entity-attribute tuples (e.g., "*Arlo Vance was born in 1987*") into the documents. A model fine-tuned on these documents memorizes the fictitious attributes. To detect memorization, each attribute is queried via QA, and per-attribute results are aggregated via Fisher's method.

3.3 Threat Model and Considered Scenarios

FedAttr targets a non-adversarial license-violation setting in which all parties are honest-but-curious: they execute the protocol faithfully but may attempt to infer private information from observation. An unknown subset of clients trains on watermarked documents in violation of license terms.

We consider a FL system with three parties: *clients* $\{1, \ldots, K\}$ with private datasets, a *server* that coordinates training via the SA interface, and a *corpus owner* that holds the watermark detection key. During training, the server coordinates FL training with clients and observes only authorized subset sums $S^t(W)$ through SA; the corpus owner is not involved. After training, the server sends the stored FedAttr estimates together with the corresponding global-model checkpoints to the corpus owner. The corpus owner then applies the watermark detector and produces the client-level attribution decisions. The server never learns the detection key, and the corpus owner never observes an individual client update.

4 FedAttr Protocol

FedAttr preserves the standard FL training and SA protocol, and enables the corpus owner to identify which clients use the watermarked corpus. Specifically, FedAttr estimates the update via a paired-subset-difference mechanism motivated and supported by Theorems 1- 2. FedAttr contains three steps: (i) during training, the server constructs each client's update estimates from paired SA queries and stores them; (ii) after training, the server sends the estimates and corresponding global-model checkpoints to the corpus owner, which computes differential watermark scores; and (iii) the corpus owner combines the per-round scores via Stouffer's method to obtain the final client-level decisions. For each stage, we perform theoretical analyses to illustrate that FedAttr can preserve the utility of client update estimates, showing the effectiveness of our protocol. *Algorithm 1* summarizes it.

4.1 Client-level Update Estimator

Our goal here is to construct an unbiased estimator of any single client's update through the SA interface. The challenge is that SA hides clients' individual updates. The first step is based on a key observation: Two subset SA queries differing only in whether they include a target client i must differ only by client i's update in expectation. We call it the paired-subset-difference mechanism. We can construct an unbiased estimator of any single client's update via paired subset SA queries.

Constructing the unbiased update estimator via paired subset SA queries: Given a target client $i \in [K]$ and $N \in [K-1]$, the number of non-target clients per query.[2] We define two sampling families over the non-target clients $[K]\backslash\{i\}$, distinguished by whether they include the target client i:

$$\mathcal{U}_i^N \;=\; \big\{\, U \subseteq [K] : i \in U,\ |U| = N{+}1 \,\big\}, \qquad \mathcal{V}_i^N \;=\; \big\{\, V \subseteq [K] : i \notin V,\ |V| = N \,\big\}. \tag{2}$$

The server draws M include-target subsets $U_1^t, \ldots, U_M^t$ i.i.d. uniformly from $\mathcal{U}_i^N$ and M exclude-target subsets $V_1^t, \ldots, V_M^t$ i.i.d. uniformly from $\mathcal{V}_i^N$, then forms the round-t *update estimator*

$$\widehat{\Delta}_i^{\,t} \;:=\; \frac{1}{M}\sum_{m=1}^{M} S_t(U_m^t) \;-\; \frac{1}{M}\sum_{m=1}^{M} S_t(V_m^t). \tag{3}$$

For convenience, we denote the paired queries at round t to target client i by $\mathcal{Q}_i^t := (U_1^t, \ldots, U_M^t; V_1^t, \ldots, V_M^t)$, and all queries in round t by $\mathcal{Q}^t = \{\mathcal{Q}_i^t\}_{i=1}^K$. For each non-target client $j \neq i$, the *masking coefficient* is

$$\alpha_j^t := \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{j \in U_m^t\} - \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{j \in V_m^t\}.$$

[2]Both subset sizes N and $N+1$ must satisfy the SA protocol's authorization threshold N_{sa} (Bonawitz et al., 2017); we assume this throughout. In practice N_{sa} is small relative to K.

Note that the queries process is independent with the global model and client updates.

Rejecting estimator when privacy condition fails: To ensure that the non-target updates provide sufficient masking noise for the privacy analysis (Section 5), for instance, to exclude the degenerate case where the include and exclude subsets draw identical non-target clients, leaving $\widehat{\Delta}_i^{\,t} = \Delta_i^t$, the server checks the following privacy condition before querying the SA interface:

$$c_i^t := \sum_{j \neq i} (\alpha_j^t)^2 \geq aN, \qquad M_{\text{eff},i}^t := \frac{(c_i^t)^2}{\sum_{j \neq i} (\alpha_j^t)^4} \geq aN, \qquad \text{and} \qquad N < K - 1, \tag{4}$$

where $a := (1-\rho)/M$ and $\rho := N/(K-1) < 1$. If the condition fails, the server re-samples the subsets. This rejection policy depends only on the subset choice and never on client updates and the global model. We define the acceptance event at round t: $\mathcal{A}_i^t := \Big\{ c_i^t \geq aN,\ M_{\text{eff},i}^t \geq aN,\ N < (K-1) \Big\}$.

By the symmetry of the accepted sampling distribution, each non-target client's updates cancel in expectation. We prove that FedAttr constructs an unbiased estimator (Theorem 1) with variance controlled by the non-target updates (Theorem 2).

Theorem 1 (Unbiasedness under rejection sampling)**.** *Given any round t and target client $i \in [K]$. Under the subset sampling with rejection described above, for any deterministic updates $\Delta_1^t, \ldots, \Delta_K^t$,*

$$\mathbb{E}\big[\widehat{\Delta}_i^{\,t} \,\big|\, \Delta_1^t, \ldots, \Delta_K^t, \mathcal{A}_i^t\big] = \Delta_i^t.$$

Theorem 2 (Conditional variance under rejection sampling)**.** *Under the same setting as Theorem 1, the conditional covariance satisfies*

$$\mathrm{Cov}\big(\widehat{\Delta}_i^{\,t} \,\big|\, \Delta_1^t, \ldots, \Delta_K^t, \mathcal{A}_i^t\big) \preceq \frac{1}{p_{a,i}^t} \cdot \frac{2}{M} \cdot \frac{N(K-1-N)}{K-2} \cdot \Sigma_{-i}^{\,t}, \tag{5}$$

where $p_{a,i}^t := \Pr_{\mathcal{Q}^t}(\mathcal{A}_i^t)$ and $\Sigma_{-i}^t := \frac{1}{K-1} \sum_{j \neq i} (\Delta_j^t - \bar{\Delta}_{-i}^t)(\Delta_j^t - \bar{\Delta}_{-i}^t)^\top$.

Complete proofs are in Appendix B.1-B.2. The rejection check introduces negligible overhead: the threshold aN equals half the expected masking strength and the rejection probability decays as $e^{-\Omega(N)}$ nearly identical to unrestricted sampling, so the expected number of redraws $1/p_{a,i}^t \to 1$ exponentially fast. Theorem 2 shows that increasing the query count M reduces estimator noise at the cost of additional SA queries. Detailed analysis of the acceptance rate is in the Appendix E.

4.2 Differential Scoring

In this stage, we aim to score the client's estimate with the watermark detector. The challenge is that the global model w^{t-1} already carries watermark signals absorbed from previous rounds, causing the detector to assign high scores to all clients, including benign ones. Applying the detector directly to the updates causes 57% FPR (demonstrated in Table 1) even with access to plaintext updates.

FedAttr addresses this via *differential scoring*: it evaluates the detector at both the global model w^{t-1} and the estimate model $w^{t-1} + \widehat{\Delta}_i^t$, and calculates the difference. For each round t and target client i:

$$z_i^{(t)} := \textsc{Score}(w^{t-1} + \widehat{\Delta}_i^t;\ \mathcal{P}_t) - \textsc{Score}(w^{t-1};\ \mathcal{P}_t), \tag{6}$$

where $\mathcal{P}_t$ is the evaluation prompt set containing watermark pattern at round t. Motivated by analysis in Appendix D, differential scoring can effectively reduce the watermark bias in the global model, and $z_i^{(t)}$ measures only the contribution of client i's estimated update. As demonstrated by Theorems 1–2, $\widehat{\Delta}_i^t$ is unbiased for Δ_i^t, so other clients contribute only to sampling variance.

4.3 Cross-round Stouffer Combination

Our goal here is to combine per-round scores across rounds to identify the watermarked client. A single round yields a weak signal because the watermark signal is not completely learned after one round of local fine-tuning; the full watermark signal emerges gradually as training T increases.

Moreover, the estimator would introduce sampling noise in the first stage (Section 4.1). FedAttr accumulates this growing signal via cross-round Stouffer's combination (Stouffer, 1949):

$$Z_i = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_i^{(t)}. \tag{7}$$

The corpus owner flags client i as watermarked if $Z_i > \gamma$ for a fixed threshold $\gamma > 0$.

To state the formal guarantee, we introduce a separation condition on the per-round scores. Let $\mathcal{F}_{t-1} = \sigma(w^{t-1}, \mathcal{Q}^1, \ldots, \mathcal{Q}^{t-1})$ denote the global model and all queries history up to round $t-1$.

Assumption 1 (Watermark signal separation condition)**.** *There exist constants $m > 0$, $\epsilon \in [0, m)$, and $\nu > 0$ such that for each client $i \in [K]$ and round $t \in [T]$: (i) the conditional mean $\mu_i^{(t)} := \mathbb{E}[z_i^{(t)} \mid \mathcal{F}_{t-1}]$ satisfies $\mu_i^{(t)} \geq m$ if client i is watermarked and $|\mu_i^{(t)}| \leq \epsilon$ if client i is benign; (ii) the centered increment $z_i^{(t)} - \mu_i^{(t)}$ is conditionally ν^2-sub-Gaussian given $\mathcal{F}_{t-1}$.*

We verify this assumption empirically in Figure 2(b,c). Then we introduce the following theorem:

Theorem 3 (Stouffer error)**.** *Under Assumption 1, for any threshold satisfying $\sqrt{T}\,\epsilon < \gamma < \sqrt{T}\,m$,*

$$\Pr\big(\textit{error for client } i\big) \;\leq\; \begin{cases} \exp\!\big(-(\gamma - \sqrt{T}\,\epsilon)^2/2\nu^2\big) & \textit{if } i \textit{ is benign}, \\ \exp\!\big(-(\sqrt{T}\,m - \gamma)^2/2\nu^2\big) & \textit{if } i \textit{ is watermarked}. \end{cases} \tag{8}$$

The proof is in Appendix B.3. A fixed γ controls both errors: the false-positive bound depends on $\gamma - \sqrt{T}\,\epsilon$, while the false-negative rate decays exponentially once $\sqrt{T}\,m > \gamma$.

5 Privacy Analysis

We analyze the information leakage of FedAttr's estimation to the corpus owner with respect to clients' updates. FedAttr operates entirely through SA, preserving the SA's privacy guarantee for each model update. However, a residual information-leakage threat remains (Elkordy et al., 2023): the server obtains a noisy estimation $\widehat{\Delta}_i^t$ of client i's actual update Δ_i^t. We quantify this residual leakage using mutual information (MI), following the framework of Elkordy et al. (2023). Our analysis extends theirs from the standard SA setting to the *subset-query* setting where FedAttr performs.

Leakage metric. Given a target client i within round t, the per-round leakage is

$$I_{\text{priv}}^{(t)} := I\Big(\Delta_i^t;\ \widehat{\Delta}_i^t \;\Big|\; \mathcal{Q}_i^t,\ \mathcal{F}_{t-1}\Big). \tag{9}$$

This quantity measures how much information the estimator $\widehat{\Delta}_i^t$ reveals about client i's actual update, beyond what is known from $\mathcal{F}_{t-1} = \sigma(w^{t-1}, \mathcal{Q}^1, \ldots, \mathcal{Q}^{t-1})$ and the round t query $\mathcal{Q}_i^t$.

Inspired by Elkordy et al. (2023), we propose two assumptions on the properties of the model to shed light on the leakage of MI during the update process.

Assumption 2 (Independent under whitening)**.** *Let $Z_j^t := (K_G^t)^{-1/2}\xi_j^t$ be the whitened update. Conditioned on $\mathcal{F}_{t-1}$, the coordinates of Z_j^t are independent, centered, and have unit variance. For every coordinate $\ell \in [d^*]$, the scalar distribution $Z_{j,\ell}^t$ has finite fourth moment and finite entropic distance to the Gaussian distribution with the same mean and variance. More explicitly, if $G_\ell \sim \mathcal{N}(0,1)$, then there exist constants $M_{4,\ell} < \infty$ and $D_{0,\ell} < \infty$, independent of j, such that*

$$\mathbb{E}|Z_{j,\ell}^t|^4 \leq M_{4,\ell}, \qquad D_{\mathrm{KL}}(Z_{j,\ell}^t \| G_\ell) = h(G_\ell) - h(Z_{j,\ell}^t) \leq D_{0,\ell}.$$

These are the one-dimensional regularity conditions needed to apply the Bobkov–Chistyakov–Götze entropic Berry–Esseen bound used in the independent-under-whitening case of Elkordy et al. (2023).

Assumption 3. *The local datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$ are sampled i.i.d. from a distribution, i.e., the local dataset of client j consists of i.i.d. data samples from a distribution $\mathcal{P}_j$, where $\mathcal{P}_j = \mathcal{P}$ for all $j \in [K]$. This implies that given round t and condition on the $\mathcal{F}_{t-1} = \sigma(w^{t-1}, \mathcal{Q}^1, \ldots, \mathcal{Q}^{t-1})$, each client update can decompose as $\Delta_j^t = \mu^t + \xi_j^t$, where μ^t is deterministic conditioned on $\mathcal{F}_{t-1}$, and*

$$\mathbb{E}[\xi_j^t \mid \mathcal{F}_{t-1}] = 0, \qquad \mathrm{Cov}(\xi_j^t \mid \mathcal{F}_{t-1}) = K_G^t.$$

The $\{\xi_j^t\}_{j=1}^K$ are conditionally i.i.d. on a common d^*-dimensional effective subspace, where $d^* := \mathrm{rank}(K_G^t)$. All determinants and entropies below are taken on this effective subspace.

Remark 1. *Assumption 3 is the same condition as Elkordy et al. (2023, Assumption 1) and ensures the non-target updates form an i.i.d. additive mask whose entropy can be controlled via the entropic CLT. The independence-under-whitening condition (Assumption 2) is satisfied when the stochastic gradient can be approximated by a distribution with independent components or by a multivariate Gaussian (Elkordy et al., 2023, Definition 1).*

Based on two assumptions, we propose our main privacy results.

Theorem 4 (Release-level MI leakage)**.** *Suppose Assumptions 2 and 3 hold. If a query Q_i^t satisfies $c_i^t \geq aN, M_{\mathrm{eff},i}^t \geq aN$, and $N < K - 1$ then*

$$I(\Delta_i^t; \widehat{\Delta}_i^t \mid Q_i^t, \mathcal{F}_{t-1}) \leq \frac{d^*}{2} \log\left(1 + \frac{1}{aN}\right) + \frac{C_\xi d^*}{aN} = O\left(\frac{d^*}{N}\right).$$

The bound has two terms: the first captures leakage when the masking noise is exactly Gaussian; the second accounts for non-Gaussianity. Compared with the single-aggregate bound of Elkordy et al. (2023, Theorem 1), whose subset size $N-1$ is deterministic, FedAttr's effective subset size aN arises from the random subset queries. The detailed proof is deferred to Appendix C.

6 Experiments

6.1 Experimental Setup

Federated Learning Configurations. Consistent with previous work (Ye et al., 2024; Wu et al., 2025), we fine-tune Llama-3.2-3B (Llama Team, 2024) with LoRA on UltraChat200K (Ding et al., 2023), partitioned IID across K=10 clients for T=5 rounds. Default protocol parameters are r=3 watermarked clients, subset size N=5, query count M=5. We evaluate two aggregation strategies (FedIT (Zhang et al., 2023), FLoRA (Wang et al., 2024)) and two watermark families (KGW (Kirchenbauer et al., 2023), Fictitious Knowledge (Cui et al., 2025)).

Baselines. We compare against four baselines. (i) *Global model test*: which applies the watermark detector to the global model but cannot attribute to clients, (ii) *Direct (oracle)*: which applies the detector to each client's plaintext update, violating SA, (iii) *FLDetector* (Zhang et al., 2022), (iv) *FLForensics* (Jia et al., 2024). Notably, (ii)-(iv) require plaintext updates and violate SA.

Metrics. We report TPR (fraction of watermarked clients correctly flagged) and FPR (fraction of benign clients incorrectly flagged) for each approach. FedAttr flags client i when its FedAttr Stouffer score $Z_i \geq 4$. We implement each baseline following its default configurations. For each approach under different settings, we report results (*i.e., mean & std*) calculated over three random seeds. We also report the p-value obtained by converting the Stouffer statistic Z_i to a one-sided standard normal tail probability, *i.e.*, $p_i = 1 - \Phi(Z_i)$. In ablation studies, we additionally report $\bar{z}_{\mathrm{pos}}$ and $\bar{z}_{\mathrm{neg}}$, the mean Stouffer statistics of watermarked and benign clients, to quantify signal strength beyond the binary TPR/FPR. Full hyperparameters, watermark details, and baselines are in Appendix F.

6.2 Main Results

Table 1 reports client-level attribution performance for different watermark families and FL algorithms. FedAttr achieves 100% TPR and 0% FPR in all four settings ($p < 10^{-6}$), even completely performing through SA . No baseline matches this under the same privacy constraint with SA. The direct oracle has plaintext access to each client's update but incurs FPR $\geq$ 57%. As the global model has already learned watermark signals from earlier rounds, the detector thus assigns high scores to all clients, including benign ones. Instead, differential scoring applied in **FedAttr** significantly reduces the watermark bias of the global model, accurately isolating each client's individual effect to the watermark effectiveness. Moreover, we observe previous forensics approach Jia et al. (2024); Zhang et al. (2022) performs ineffectively in our considered scenarios as the watermark signal cannot be adapted as adversary patterns. As a result, FLDetector achieves 0% TPR, and FLForensics achieves at most 33–100% TPR with 19–24% FPR. Even more, all existing approaches require plaintext individual updates, violating SA, and cannot be used for watermark attribution (no p-value).

Table 1: Client-level watermark attribution performance (T=5 rounds, γ=4.0, mean$_{\pm\text{std}}$ over 3 seeds). FedAttr is the only method that achieves 100% TPR, 0% FPR in four settings through secure aggregation. For FLForensics, † denotes the original implementation using HDBSCAN clustering, and ‡ denotes our adaptation using k-means clustering, since HDBSCAN fails at small K=10.

			KGW			Fictitious Knowledge		
FL Algorithm	Baseline	SA	TPR ↑	FPR ↓	p-value	TPR ↑	FPR ↓	p-value
FedIT	Direct (oracle)	✗	$55.6_{\pm 19.3}$	$57.1_{\pm 14.3}$	$< 10^{-3}$	$100.0_{\pm 0.0}$	$57.1_{\pm 14.3}$	$< 10^{-15}$
	FLDetector	✗	$0.0_{\pm 0.0}$	$19.1_{\pm 8.3}$	—	$0.0_{\pm 0.0}$	$19.1_{\pm 8.3}$	—
	FLForensics†	✗	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	—	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	—
	FLForensics‡	✗	$33.3_{\pm 0.0}$	$23.8_{\pm 8.3}$	—	$100.0_{\pm 0.0}$	$19.1_{\pm 8.3}$	—
	FedAttr (ours)	✓	$\mathbf{100.0}_{\pm \mathbf{0.0}}$	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$< 10^{-6}$	$\mathbf{100.0}_{\pm \mathbf{0.0}}$	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$< 10^{-29}$
FLoRA	Direct (oracle)	✗	$100.0_{\pm 0.0}$	$71.4_{\pm 14.3}$	$< 10^{-6}$	$100.0_{\pm 0.0}$	$66.7_{\pm 8.3}$	$< 10^{-20}$
	FLDetector	✗	$0.0_{\pm 0.0}$	$14.3_{\pm 0.0}$	—	$0.0_{\pm 0.0}$	$23.8_{\pm 8.3}$	—
	FLForensics†	✗	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	—	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	—
	FLForensics‡	✗	$55.6_{\pm 19.3}$	$23.8_{\pm 8.3}$	—	$100.0_{\pm 0.0}$	$19.1_{\pm 8.3}$	—
	FedAttr (ours)	✓	$\mathbf{100.0}_{\pm \mathbf{0.0}}$	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$< 10^{-10}$	$\mathbf{100.0}_{\pm \mathbf{0.0}}$	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$< 10^{-50}$

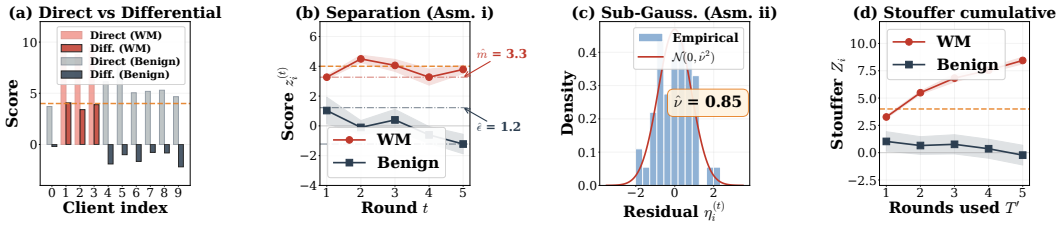

Figure 2: The effect for each protocol component. **(a)** Differential scoring removes the accumulated bias. **(b)** Per-round watermarked mean stays above $\hat{m}$=3.3, benign mean within $\pm\hat{\epsilon}$=1.2, validating Assumption 1(i). **(c)** Centered residuals $z_i^{(t)} - \hat{\mu}_i$ match a Gaussian ($\hat{\nu}$=0.85), supporting Assumption 1(ii). **(d)** Stouffer statistic crosses γ=4 from round 2. At round 5, the statistic achieves margin $Z_i - \gamma = 4.0$.

6.3 Mechanism Analysis

We further investigate the effect and soundness for each component within **FedAttr**. Figure 2 presents the results. Figure 2(a) compares direct and differential scoring at round t=5. Direct scoring yields 57% FPR because the global model has learned watermark signals from earlier rounds. Differential scoring subtracts the reference score, reducing the accumulated bias: only watermarked clients retain a detectable signal. Figure 2(b) shows the per-round performance of differential scores. The watermark efficacy stays above $\hat{m}$=3.3 in each round, while the benign ones remain within $\pm\hat{\epsilon}$=1.2, consistent with Assumption 1(i)(Separation). Figure 2(c) shows the empirical distribution of centered residuals $z_i^{(t)} - \hat{\mu}_i$, closely aligns with a (sub)Gaussian distribution ($\hat{\nu}$=0.85), supporting Assumption 1(ii)(Sub-Gaussianity). Figure 2(d) shows the Stouffer statistic computed with varying T'. The watermark efficacy becomes larger than γ=4 when round $T' \geq 2$ and reaches $Z_i > 8.0$ at T'=5, yielding a margin of 4.0 above γ. Benign clients remain near zero throughout. Therefore, the Stouffer process amplifies the signal.

6.4 Ablation Studies

We study the impact of different parameters (*e.g.,* the number of watermark clients, subset size, *etc*) and different configurations (*e.g.,* LoRA rank, dataset, *etc*); Figures 3 (protocol parameters) and 4 (training configurations) summarize the results.

Protocol parameters (Figure 3). FedAttr achieves consistently 100% TPR and 0% FPR under varying amounts of watermark clients, subsets, and watermark ratio. The watermark efficacy $\bar{z}_{\text{pos}}$ exhibits the U-shaped dependence on N consistent with Theorem 2: lowest at N=5 where the

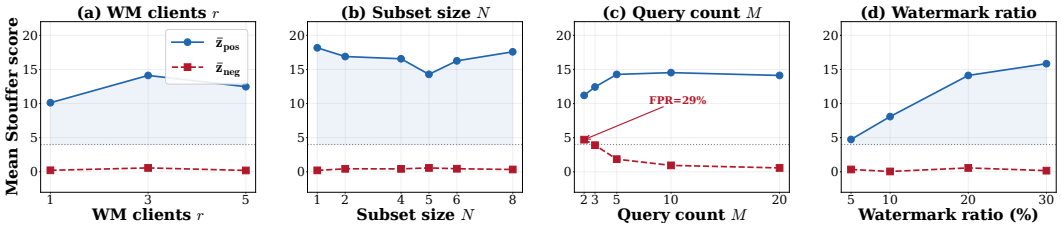

Figure 3: Protocol parameter sensitivity. FedAttr achieves 100% TPR / 0% FPR across all tested values except M=2, which incurs 29% FPR, and shows robustness under protocol parameter selection. **(a)** Number of watermarked clients r. **(b)** Subset size N: the U-shaped curve validates Theorem 2. (c) Query count M. (d) Watermark ratio: The signal scales roughly linearly with watermark ratio.

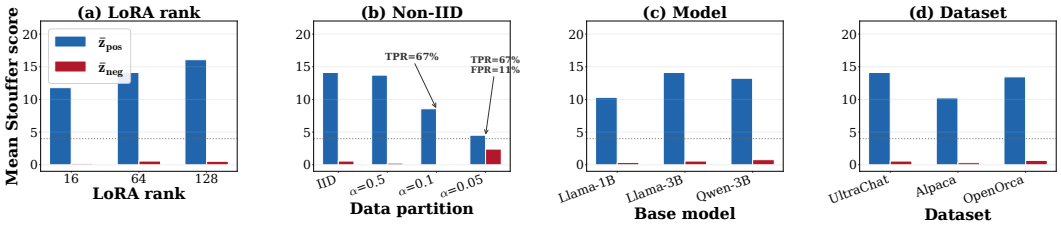

Figure 4: Robustness to training configurations. FedAttr achieves 100% TPR and 0% FPR across LoRA ranks, base models, and datasets. Attribution accuracy decreases moderately under severe non-IID partitions ($\alpha \leq 0.1$). **(a)** LoRA rank. **(b)** Non-IID heterogeneity. (smaller Dirichlet α means more heterogeneous) **(c)** Base model. **(d)** Training dataset.

variance factor $N(K-1-N)/(K-2)$ peaks, and highest at N=1 where the estimator becomes exact (Figure 3(b)). Query count $M \geq 3$ achieves 100% TPR and 0% FPR. M=2 incurs 29% FPR due to high estimator noise (Figure 3(c)). The watermark efficacy scales roughly linearly *w.r.t.* watermark ratio, consistent with radioactivity theory (Sander et al., 2024).

Robustness (Figure 4). FedAttr performs robustly under different configurations of LoRA ranks, evaluated models, and datasets, consistently achieving 100% TPR and 0% FPR. Under severe non-IID heterogeneity (α=0.1), TPR degrades to 67% while FPR remains 0%; at α=0.05, FPR rises to 11%. This degradation is consistent with increased estimator variance when client updates diverge, and can be mitigated by increasing T (Table 21). The detailed analyses are included in the Appendix.

6.5 Scalability and Overhead

Scalability. We scale K from 10 to 100. Table 17 shows that watermark efficacy $\bar{z}_{\text{pos}}$ decreases from 10.12 to 7.83 as K increases but remains higher above threshold γ (100% TPR, 0% FPR in all cases). FedAttr also performs effectively in a partial-participation setting (Table 20 in Appendix).

Overhead. FedAttr's overhead consists of SA queries and watermark scoring. In our main experiment, FedAttr issues $2MKT$=500 SA queries, adding 5 minutes (1.0%) to the 8.5-hour FL training time; watermark scoring adds 27 minutes (5.3%), for a total overhead of 6.3%. Both costs scale linearly in K. Since all computation runs on the server, it can be overlapped with clients' local training in the next round, effectively hiding the latency. Complete analysis is in Appendix H.

7 Conclusion

We introduced FedAttr, a client-level attribution protocol for federated LLM fine-tuning that identifies clients who trained on watermarked documents while preserving SA privacy. FedAttr combines unbiased update estimation from SA queries, differential scoring, and cross-round Stouffer aggregation. We provided theoretical guarantees on the estimator's unbiasedness and variance, and bounded mutual information leakage of $O(d^*/N)$ per round. Empirically, FedAttr achieves 100% TPR, 0% FPR, outperforming all baselines while being the only method that preserves SA privacy.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, identifying potentially relevant literature, and code generation. All citations included in the paper were independently located and verified by the authors using DBLP rather than taken directly from AI-generated outputs. AI-generated code was reviewed, tested, and validated by the authors before being used in the experiments. The authors take full responsibility for the manuscript, citations, code, and reported results. We also used pretrained language models to generate the watermarked corpora used in our experiments: Llama-3.2-3B-Instruct (Llama Team, 2024) for the KGW/TextSeal watermark following Sander et al. (2024), and Llama-3.1-8B-Instruct (Llama Team, 2024) for the Fictitious Knowledge watermark following Cui et al. (2025), as detailed in Appendix F.2.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results are provided in the appendix. Experimental settings, data splits, model and optimization configurations, hyperparameter selection procedures, and additional results are documented in Appendix F.

REFERENCES

Sergey G. Bobkov, Gennadiy P. Chistyakov, and Friedrich Götze. Berry-esseen bounds in the entropic central limit theorem, 2011. URL https://arxiv.org/abs/1105.4119.

Kallista A. 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. *IACR Cryptol. ePrint Arch.*, 2017:281, 2017. URL http://eprint.iacr.org/2017/281.

Huancheng Chen and Haris Vikalo. Recovering labels from local updates in federated learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7346–7372. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/chen24an.html.

W.G. Cochran. *Sampling Techniques*. Wiley publication in applied statistics. Wiley, 1977. ISBN 9788126515240. URL https://books.google.com/books?id=xbNn41DUrNwC.

Xinyue Cui, Johnny Tian-Zheng Wei, Swabha Swayamdipta, and Robin Jia. Robust data watermarking in language models by injecting fictitious knowledge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pp. 14292–14306. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.736/.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 3029–3051. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.183. URL https://doi.org/10.18653/v1/2023.emnlp-main.183.

Ahmed Roushdy Elkordy, Jiang Zhang, Yahya H. Ezzeldin, Konstantinos Psounis, and Salman Avestimehr. How much privacy does federated learning with secure aggregation guarantee? *Proc. Priv. Enhancing Technol.*, 2023(1):510–526, 2023. doi: 10.56553/POPETS-2023-0030. URL https://doi.org/10.56553/popets-2023-0030.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In Hugo

Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuqi Jia, Minghong Fang, Hongbin Liu, Jinghuai Zhang, and Neil Zhenqiang Gong. Tracing back the malicious clients in poisoning attacks to federated learning. *arXiv preprint arXiv:2407.07221*, 2024.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, Proceedings of Machine Learning Research, pp. 1273–1282. PMLR, 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 8326–8335. PMLR, 2020. URL http://proceedings.mlr.press/v119/sablayrolles20a.html.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2567c95fd41459a98a73ba893775d22a-Abstract-Conference.html.

S.A. Stouffer. *The American Soldier: Adjustment During Army Life*. Studies in social psychology in World War II. Princeton University Press, 1949. URL https://books.google.com/books?id=hQiBwgEACAAJ.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/28312c9491d60ed0c77f7fff4ad86dd1-Abstract-Conference.html.

Yebo Wu, Chunlin Tian, Jingguang Li, He Sun, Kahou Tam, Zhijiang Guo, Li Li, and Chengzhong Xu. A survey on federated fine-tuning of large language models. *ArXiv*, abs/2503.12016, 2025. URL https://api.semanticscholar.org/CorpusID:277065732.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 6137–6147. ACM, 2024. doi: 10.1145/3637528.3671582. URL https://doi.org/10.1145/3637528.3671582.

Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning. In Ada Gavrilovska and Erez Zadok (eds.), *Proceedings of the 2020 USENIX Annual Technical Conference, USENIX ATC 2020, July 15-17, 2020*, pp. 493–506. USENIX Association, 2020. URL https://www.usenix.org/conference/atc20/presentation/zhang-chengliang.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Guoyin Wang, and Yiran Chen. Towards building the federated GPT: federated instruction tuning. *CoRR*, abs/2305.05644, 2023. doi: 10.48550/ARXIV.2305.05644. URL https://doi.org/10.48550/arXiv.2305.05644.

Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In Aidong Zhang and Huzefa Rangwala (eds.), *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pp. 2545–2555. ACM, 2022. doi: 10.1145/3534678.3539231. URL https://doi.org/10.1145/3534678.3539231.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14747–14756, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html.

A ALGORITHM

Algorithm 1 presents the FedAttr protocol. At each communication round t, clients perform standard local training and submit updates through secure aggregation (Lines 1–3). For each target client i, the server repeatedly samples paired include/exclude subsets and checks the privacy condition (Eq. (4)) via rejection sampling (Lines 5–10). Upon acceptance, the server queries the SA oracle to form the unbiased update estimator $\widehat{\Delta}_i^t$ (Line 11) and forwards it to the corpus owner, who computes the differential score $z_i^{(t)}$ (Lines 12–13). After all T rounds, the corpus owner aggregates per-round scores via Stouffer's method and applies the threshold γ to produce the final attribution decision (Lines 15–21).

Algorithm 1 FedAttr: Client-level Attribution through Secure Aggregation

Require: Clients $[K]$, communication rounds T, subset size N, query count M, threshold γ, SA oracle $S_t(\cdot)$, watermark detector $\textsc{Score}(\cdot\,;\cdot)$, prompt sets $\{\mathcal{P}_t\}_{t=1}^T$

Ensure: Attribution decision for each client $i \in [K]$

1: **for** $t = 1, \ldots, T$ **do**
2: Each client i computes local update Δ_i^t and submits to the SA oracle
3: Server updates global model: $w^t \leftarrow w^{t-1} + \sum_{i=1}^K p_i \Delta_i^t$
4: **for** each target client $i \in [K]$ **do**
5: **repeat**
6: Draw $U_1^t, \ldots, U_M^t \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathcal{U}_i^N)$, Include-target, $|U_m^t| = N{+}1$
7: Draw $V_1^t, \ldots, V_M^t \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathcal{V}_i^N)$, Exclude-target, $|V_m^t| = N$
8: Compute $\alpha_j^t \leftarrow \frac{1}{M}\sum_m \mathbf{1}\{j \in U_m^t\} - \frac{1}{M}\sum_m \mathbf{1}\{j \in V_m^t\}$ for all $j \neq i$
9: Compute $c_i^t \leftarrow \sum_{j\neq i}(\alpha_j^t)^2, \quad M_{\text{eff},i}^t \leftarrow (c_i^t)^2 \big/ \sum_{j\neq i}(\alpha_j^t)^4$
10: **until** $c_i^t \geq aN$ **and** $M_{\text{eff},i}^t \geq aN$ **and** $N < K{-}1$ (Eq. (4))
11: Query SA and compute: $\widehat{\Delta}_i^t \leftarrow \frac{1}{M}\sum_{m=1}^M S_t(U_m^t) - \frac{1}{M}\sum_{m=1}^M S_t(V_m^t)$
12: Server sends $\widehat{\Delta}_i^t$ to corpus owner
13: $z_i^{(t)} \leftarrow \textsc{Score}\big(w^{t-1} + \widehat{\Delta}_i^t;\ \mathcal{P}_t\big) - \textsc{Score}\big(w^{t-1};\ \mathcal{P}_t\big)$
14: **end for**
15: **end for**
16: **for** each client $i \in [K]$ **do**
17: $Z_i \leftarrow \frac{1}{\sqrt{T}}\sum_{t=1}^T z_i^{(t)}$
18: **if** $Z_i > \gamma$ **then**
19: Flag client i as **watermarked**
20: **else**
21: Label client i as **benign**
22: **end if**
23: **end for**

B PROOFS FOR ATTRIBUTION PROTOCOL

B.1 PROOF OF THEOREM 1 (UNBIASEDNESS UNDER REJECTION SAMPLING)

Proof. Substituting $S^t(W) = \sum_{j\in W} \Delta_j^t$ into (3) gives

$$\widehat{\Delta}_i^t = \frac{1}{M}\sum_{m=1}^M \left(\Delta_i^t + \sum_{j\in U_m^t\setminus\{i\}} \Delta_j^t\right) - \frac{1}{M}\sum_{m=1}^M \sum_{j\in V_m^t} \Delta_j^t.$$

Since $i \in U_m^t$ for all m and $i \notin V_m^t$ for all m, this can be written as

$$\widehat{\Delta}_i^t = \Delta_i^t + \sum_{j\neq i} \alpha_j^t \Delta_j^t,$$

where

$$\alpha_j^t = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{j \in U_m^t \setminus \{i\}\} - \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{j \in V_m^t\}, \qquad j \neq i.$$

Thus it suffices to show

$$\mathbb{E}[\alpha_j^t \mid \mathcal{A}_i^t] = 0 \qquad \text{for every } j \neq i,$$

where $\mathcal{A}_i^t$ denotes the acceptance event

$$\mathcal{A}_i^t = \{c^t \geq aN,\ M_{\text{eff},i}^t \geq aN\}.$$

For each m, define the non-target part of the include-target query by

$$X_m^t := U_m^t \setminus \{i\},$$

and define

$$Y_m^t := V_m^t.$$

By construction, both X_m^t and Y_m^t are N-subsets of $[K] \setminus \{i\}$. Moreover, the proposal distribution samples

$$X_1^t, \ldots, X_M^t, Y_1^t, \ldots, Y_M^t$$

i.i.d. uniformly from the same family of N-subsets of $[K] \setminus \{i\}$.

Let

$$\mathcal{Q}^t = (X_1^t, \ldots, X_M^t; Y_1^t, \ldots, Y_M^t)$$

denote the non-target query design. Define the swap map T by

$$T(\mathcal{Q}^t) = (Y_1^t, \ldots, Y_M^t; X_1^t, \ldots, X_M^t).$$

Equivalently, after applying T, the corresponding include-target and exclude-target subsets are reconstructed as

$$(U_m^t)' = \{i\} \cup Y_m^t, \qquad (V_m^t)' = X_m^t.$$

This map is well-defined because X_m^t and Y_m^t have the same cardinality and both lie in $[K] \setminus \{i\}$. It is also a bijection. Since the proposal distribution samples the X's and Y's i.i.d. from the same distribution, the proposal distribution is invariant under T.

For every non-target client $j \neq i$,

$$\alpha_j^t(T(\mathcal{Q}^t)) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{j \in Y_m^t\} - \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{j \in X_m^t\} = -\alpha_j^t(\mathcal{Q}^t).$$

Therefore the swap map sends the coefficient vector

$$\alpha^t = (\alpha_j^t)_{j \neq i}$$

to $-\alpha^t$.

The acceptance event $\mathcal{A}$ depends on the query design only through

$$c^t = \sum_{j \neq i} (\alpha_j^t)^2$$

and

$$M_{\text{eff}}^t = \frac{(c^t)^2}{\sum_{j \neq i} (\alpha_j^t)^4}.$$

Both quantities are invariant under the sign change $\alpha^t \mapsto -\alpha^t$. Hence

$$\mathbf{1}\{\mathcal{A}_i^t(\mathcal{Q}^t)\} = \mathbf{1}\{\mathcal{A}_i^t(T(\mathcal{Q}^t))\}.$$

Since T preserves the proposal distribution and preserves the acceptance event, the accepted query design is also invariant under T:

$$\mathcal{Q}^t \mid \mathcal{A}_i^t \overset{d}{=} T(\mathcal{Q}^t) \mid \mathcal{A}_i^t.$$

Consequently, for every $j \neq i$,

$$\mathbb{E}[\alpha_j^t \mid \mathcal{A}_i^t] = \mathbb{E}[\alpha_j^t(T(\mathcal{Q}^t)) \mid \mathcal{A}_i^t] = -\mathbb{E}[\alpha_j^t \mid \mathcal{A}_i^t],$$

which implies

$$\mathbb{E}[\alpha_j^t \mid \mathcal{A}_i^t] = 0.$$

Finally, the query design $\mathcal{Q}^t$ is sampled independently of the client updates $\Delta_1^t, \ldots, \Delta_K^t$ and the current global model, and the acceptance event $\mathcal{A}_i^t$ is a function only of the query design. Therefore the same identity holds after conditioning on the realized updates:

$$\mathbb{E}[\alpha_j^t \mid \Delta_1^t, \ldots, \Delta_K^t, \mathcal{A}_i^t] = 0, \qquad j \neq i.$$

Thus

$$\begin{aligned}\mathbb{E}\left[\widehat{\Delta}_i^t \mid \Delta_1^t, \ldots, \Delta_K^t, \mathcal{A}_i^t\right] &= \Delta_i^t + \sum_{j \neq i} \mathbb{E}[\alpha_j^t \mid \Delta_1^t, \ldots, \Delta_K^t, \mathcal{A}_i^t]\Delta_j^t \\ &= \Delta_i^t.\end{aligned}$$

This proves the unbiasedness of the accepted estimator. □

B.2 Proof of Theorem 2 (Conditional variance under rejection sampling)

Proof. Let $\zeta(\mathcal{Q}^t) := \sum_{j \neq i} \alpha_j^t \, \Delta_j^t = \widehat{\Delta}_i^t - \Delta_i^t$. By Theorem 1, $\mathbb{E}[\zeta \mid \mathcal{A}_i^t, \Delta_1^t, \ldots, \Delta_K^t] = 0$, so the accepted covariance is

$$\mathrm{Cov}(\widehat{\Delta}_i^t \mid \mathcal{A}_i^t, \Delta_1^t, \ldots, \Delta_K^t) = \mathbb{E}[\zeta\zeta^\top \mid \mathcal{A}_i^t] = \frac{1}{p_a}\,\mathbb{E}[\mathbf{1}_{\mathcal{A}}\,\zeta\zeta^\top] \preceq \frac{1}{p_a}\,\mathbb{E}[\zeta\zeta^\top],$$

where $p_{a,i}^t = \Pr(\mathcal{A}_i^t)$. Each non-target part $U_m^t \setminus \{i\}$ is a simple random sample of size N drawn without replacement from the $K-1$ non-target clients. By the finite-population correction formula (Cochran, 1977), each include-target sum $A_m := \sum_{j \in U_m^t \setminus \{i\}} \Delta_j^t$ has covariance $\mathrm{Cov}(A_m) = N(K-1-N)/(K-2)\,\Sigma_{-i}^t$, and likewise for each exclude-target sum. The $2M$ sums are mutually independent, so

$$\mathbb{E}[\zeta\zeta^\top] = \frac{2}{M} \cdot \frac{N(K-1-N)}{K-2}\,\Sigma_{-i}^t.$$

Combining gives

$$\mathrm{Cov}(\widehat{\Delta}_i^t \mid \mathcal{A}_i^t, \Delta_1^t, \ldots, \Delta_K^t) \preceq \frac{1}{p_a} \cdot \frac{2}{M} \cdot \frac{N(K-1-N)}{K-2}\,\Sigma_{-i}^t.$$ □

B.3 Proof of Theorem 3 (Stouffer concentration)

Proof. Under Assumption 1, write $z_i^{(t)} = \mu_i^{(t)} + \eta_i^{(t)}$ where $\eta_i^{(t)}$ is conditionally ν^2-sub-Gaussian given $\mathcal{F}_{t-1}$. Define $S_T := \sum_{t=1}^T \eta_i^{(t)}$. By the conditional sub-Gaussian tower property, $\mathbb{E}[\exp(\lambda S_T)] \leq \exp(T\lambda^2\nu^2/2)$. Hence $W_i := T^{-1/2} S_T$ is ν^2-sub-Gaussian.

Clean client. Since $|\mu_i^{(t)}| \leq \epsilon$, $Z_i \leq \sqrt{T}\,\epsilon + W_i$, so $\Pr(Z_i \geq \gamma) \leq \exp(-(\gamma - \sqrt{T}\,\epsilon)^2/(2\nu^2))$ for $\gamma > \sqrt{T}\,\epsilon$.

Watermarked client. Since $\mu_i^{(t)} \geq m$, $Z_i \geq \sqrt{T}\,m + W_i$, so $\Pr(Z_i \leq \gamma) \leq \exp(-(\sqrt{T}\,m - \gamma)^2/(2\nu^2))$ for $\sqrt{T}\,m > \gamma$. □

C Proofs for Privacy Analysis

This section adapts the proof strategy of Elkordy et al. (2023) to the client-level update estimates produced by FedAttr. We use the same data IID assumption and distributional assumptions as the independent-under-whitening case in Elkordy et al. (2023). The difference of proof structure is that

the equal-weight secure aggregation is replaced by a query-dependent estimation. Accordingly, the number of masking users in Elkordy et al. (2023) is replaced by the effective masking size $M^t_{\text{eff},i}$, where

$$M^t_{\text{eff},i} = \frac{(c^t_i)^2}{s^t_i}, \qquad c^t_i := \sum_{j\neq i}(\alpha^t_j)^2, \qquad s^t_i := \sum_{j\neq i}(\alpha^t_j)^4.$$

The proof follows the same high-level route as the independent-under-whitening case of Elkordy et al. (2023) in 4 stages:

1. Decompose the mutual information into a difference of two entropies.
2. Upper-bound the entropy of signal plus noise by Gaussian maximum entropy.
3. Lower-bound the entropy of the masking noise using the Bobkov–Chistyakov–Götze entropic Berry–Esseen bound Bobkov et al. (2011).
4. Subtract the two bounds.

C.1 FedAttr Notation

For convenience, we first list the notations in **FedAttr**.

Given a communication round t and a target client i. FedAttr protocol queries are

$$\mathcal{Q}^t_i = (U^t_1, \ldots, U^t_M; V^t_1, \ldots, V^t_M),$$

where each include subset U^t_m contains i, and each exclude subset V^t_m does not contain i. Given a secure aggregate $S^t(W) = \sum_{j\in W}\Delta^t_j$, the released estimate for target i is

$$\widehat{\Delta}^t_i := \frac{1}{M}\sum_{m=1}^{M} S^t(U^t_m) - \frac{1}{M}\sum_{m=1}^{M} S^t(V^t_m).$$

Expanding this linear combination gives

$$\widehat{\Delta}^t_i = \Delta^t_i + \sum_{j\neq i}\alpha^t_j(\mathcal{Q}^t_i)\Delta^t_j, \tag{10}$$

where

$$\alpha^t_j(\mathcal{Q}^t_i) := \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\{j\in U^t_m\} - \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\{j\in V^t_m\}.$$

After conditioning on $\mathcal{Q}^t_i$, the coefficients α^t_j are deterministic. For convenience, given $\mathcal{Q}^t_i$, we notate $\alpha^t_j := \alpha^t_j(\mathcal{Q}^t_i)$. Therefore,

$$\widehat{\Delta}^t_i = \Delta^t_i + \sum_{j\neq i}\alpha^t_j\Delta^t_j, \tag{11}$$

where

$$\alpha^t_j := \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\{j\in U^t_m\} - \frac{1}{M}\sum_{m=1}^{M}\mathbf{1}\{j\in V^t_m\}.$$

Definition 1. *Given queries Q^t_i, define*

$$c^t_i := \sum_{j\neq i}\left(\alpha^t_j\right)^2,$$

$$s^t_i := \sum_{j\neq i}\left(\alpha^t_j\right)^4.$$

When $c^t_i > 0$, define

$$M^t_{\text{eff},i} := \frac{\left(c^t_i\right)^2}{s^t_i}.$$

Equivalently, if

$$\beta_j^t := \frac{\alpha_j^t}{\sqrt{c_i^t}},$$

then

$$\sum_{j\neq i}(\beta_j^t)^2 = 1, \qquad M_{\text{eff},i}^t = \frac{1}{\sum_{j\neq i}(\beta_j^t)^4}.$$

C.2 Stage 1: Two Entropies Decomposition

In stage 1, we decompose the mutual information into the difference of two entropies.

Lemma 1 (FedAttr estimate decomposition)**.** *Under Assumption 3, conditioned on $(Q_i^t, \mathcal{F}_{t-1})$, the FedAttr estimate can be written as*

$$\widehat{\Delta}_i^t = \left(1 + \sum_{j\neq i}\alpha_j^t\right)\mu^t + \xi_i^t + \eta_i^t,$$

where

$$\eta_i^t := \sum_{j\neq i}\alpha_j^t\xi_j^t.$$

Moreover, ξ_i^t is conditionally independent of η_i^t given $(Q_i^t, \mathcal{F}_{t-1})$. Consequently,

$$I(\Delta_i^t; \widehat{\Delta}_i^t \mid Q_i^t, \mathcal{F}_{t-1}) = h(\xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) - h(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}). \tag{12}$$

Proof. Substitute $\Delta_j^t = \mu^t + \xi_j^t$ into (10):

$$\begin{aligned}\widehat{\Delta}_i^t &= \mu^t + \xi_i^t + \sum_{j\neq i}\alpha_j^t(\mu^t + \xi_j^t)\\ &= \left(1 + \sum_{j\neq i}\alpha_j^t\right)\mu^t + \xi_i^t + \sum_{j\neq i}\alpha_j^t\xi_j^t.\end{aligned}$$

This proves the stated decomposition.

Conditioned on $(Q_i^t, \mathcal{F}_{t-1})$, the coefficients α_j^t are deterministic. The vector ξ_i^t is client i's centered update, while

$$\eta_i^t = \sum_{j\neq i}\alpha_j^t\xi_j^t$$

is a deterministic function of the centered updates of the non-target clients $\{\xi_j^t : j \neq i\}$. By Assumption 3, the centered updates $\{\xi_j^t\}_{j=1}^K$ are conditionally independent given $\mathcal{F}_{t-1}$. Therefore, ξ_i^t is conditionally independent of the collection $\{\xi_j^t : j \neq i\}$, and hence is conditionally independent of any deterministic function of the collection $\{\xi_j^t : j \neq i\}$, including η_i^t.

The deterministic shift

$$\left(1 + \sum_{j\neq i}\alpha_j^t\right)\mu^t$$

does not affect mutual information. Since $\Delta_i^t = \mu^t + \xi_i^t$, we have

$$I(\Delta_i^t; \widehat{\Delta}_i^t \mid Q_i^t, \mathcal{F}_{t-1}) = I(\xi_i^t; \xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}).$$

For independent X and Z,

$$I(X; X+Z) = h(X+Z) - h(X+Z \mid X) = h(X+Z) - h(Z).$$

Applying this identity with $X = \xi_i^t$ and $Z = \eta_i^t$ gives (12). □

C.3 Stage 2: Upper Bound the $h(\xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1})$ by Gaussian Maximum Entropy.

In this stage, we use the Gaussian maximum entropy to upper-bound the $h(\xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1})$ (first term).

Lemma 2 (Gaussian maximum-entropy upper bound)**.** *Under Assumption 3, given queries Q_i^t with $c_i^t := \sum_{j\neq i}\big(\alpha_j^t\big)^2 > 0$,*

$$h(\xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) \le \frac{1}{2}\log\det\big(2\pi e(1+c_i^t)K_G^t\big).$$

Proof. Because ξ_i^t is conditionally independent of η_i^t, covariance adds:

$$\mathrm{Cov}(\xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) = \mathrm{Cov}(\xi_i^t \mid \mathcal{F}_{t-1}) + \mathrm{Cov}(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}).$$

The first term equals K_G^t. For the second term,

$$\begin{aligned}\mathrm{Cov}(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) &= \mathrm{Cov}\left(\sum_{j\neq i}\alpha_j^t\xi_j^t \mid Q_i^t, \mathcal{F}_{t-1}\right)\\ &= \sum_{j\neq i}(\alpha_j^t)^2 K_G^t = c_i^t K_G^t,\end{aligned}$$

where the cross-covariances vanish because of conditional independence. Therefore,

$$\mathrm{Cov}(\xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) = (1+c_i^t)K_G^t.$$

Among all distributions with the same covariance matrix Σ, the Gaussian has the largest differential entropy, equal to $\frac{1}{2}\log\det(2\pi e\Sigma)$. Taking $\Sigma = (1+c_i^t)K_G^t$ proves the claim. □

C.4 Step 3: Lower Bound the $h(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1})$ by Entropic Berry–Esseen Bound

In this stage, we lower bound the $h(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1})$ by entropic Berry–Esseen bound. We first introduce a lemma to show scalar weighted entropic Berry–Esseen bound.

Lemma 3 (Scalar weighted entropic Berry–Esseen bound)**.** *Let $X_1, \ldots, X_m$ be independent centered scalar random variables. Omit any zero-variance summands, and assume the remaining summands have positive variances, finite fourth moments, densities, and finite differential entropies. Let*

$$V_m := \sum_{r=1}^m \mathrm{Var}(X_r),$$

and assume $V_m = 1$. For each r, let Z_r be a Gaussian random variable with the same mean and variance as X_r. Assume the Bobkov–Chistyakov–Götze entropic Berry–Esseen regularity holds uniformly; in particular, assume there exists $D_0 < \infty$ such that

$$D_{\mathrm{KL}}(X_r \| Z_r) = h(Z_r) - h(X_r) \le D_0 \qquad \textit{for every } r.$$

Let $G \sim \mathcal{N}(0,1)$. Then there is a constant C_{BCG}, depending only on the corresponding BCG regularity constants, such that

$$D_{\mathrm{KL}}\left(\sum_{r=1}^m X_r \,\middle\|\, G\right) \le C_{\mathrm{BCG}}\sum_{r=1}^m \mathbb{E}|X_r|^4.$$

Equivalently,

$$h\left(\sum_{r=1}^m X_r\right) \ge \frac{1}{2}\log(2\pi e) - C_{\mathrm{BCG}}\sum_{r=1}^m \mathbb{E}|X_r|^4.$$

This lemma does not introduce a FedAttr-specific modeling assumption. The condition

$$D_{\mathrm{KL}}(X_r \| Z_r) = h(Z_r) - h(X_r) \le D_0$$

is part of the regularity needed by the Bobkov–Chistyakov–Götze entropic Berry–Esseen theorem. In the independent-under-whitening case, Elkordy et al. (2023) use this entropic Berry–Esseen tool to lower-bound the entropy of an equal-weight normalized aggregate; the corresponding regularity is absorbed into their constant. We state it explicitly because FedAttr applies the same tool to query-dependent weighted summands, which are independent but not necessarily identically distributed.

Proof. This is the one-dimensional entropic Berry–Esseen theorem of Bobkov–Chistyakov–Götze for independent, not necessarily identically distributed, summands. In their notation, for

$$S_m := \frac{\sum_{r=1}^m X_r}{\sqrt{V_m}},$$

the entropic distance from S_m to the standard Gaussian is bounded by a constant depending on the uniform entropic-distance parameter, times the Lyapunov fourth-moment ratio

$$\frac{\sum_{r=1}^m \mathbb{E}|X_r|^4}{V_m^2}.$$

Since $V_m = 1$, this gives

$$D_{\mathrm{KL}}\left(\sum_{r=1}^m X_r \,\middle\|\, G\right) \le C_{\mathrm{BCG}} \sum_{r=1}^m \mathbb{E}|X_r|^4.$$

It remains only to translate the relative-entropy statement into an entropy lower bound. Let

$$S := \sum_{r=1}^m X_r.$$

Then $\mathbb{E}S = 0$ and $\mathrm{Var}(S) = 1$. The density of $G \sim \mathcal{N}(0,1)$ is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

Thus

$$\begin{aligned} D_{\mathrm{KL}}(S\|G) &= \int p_S(x) \log \frac{p_S(x)}{\phi(x)}\, dx \\ &= -h(S) - \mathbb{E}[\log \phi(S)]. \end{aligned}$$

Since

$$\log \phi(x) = -\frac{1}{2}\log(2\pi) - \frac{x^2}{2},$$

and $\mathbb{E}S^2 = 1$,

$$\mathbb{E}[\log \phi(S)] = -\frac{1}{2}\log(2\pi) - \frac{1}{2}.$$

Therefore

$$D_{\mathrm{KL}}(S\|G) = \frac{1}{2}\log(2\pi e) - h(S).$$

Rearranging the BCG bound yields the claimed entropy lower bound. □

Lemma 4 (Coordinate tensorization)**.** *Condition on $\mathcal{G} := (Q_i^t, \mathcal{F}_{t-1})$, and suppose Assumption 2 holds. Let*

$$S_\beta^t := \sum_{j\neq i} \beta_j^t Z_j^t, \qquad \sum_{j\neq i} (\beta_j^t)^2 = 1.$$

Then the coordinates of S_β^t are conditionally independent given $\mathcal{G}$, and

$$h(S_\beta^t \mid \mathcal{G}) = \sum_{\ell=1}^{d^*} h(S_{\beta,\ell}^t \mid \mathcal{G}),$$

where

$$S_{\beta,\ell}^t := \sum_{j\neq i} \beta_j^t Z_{j,\ell}^t.$$

Proof. After conditioning on $\mathcal{G}$, the coefficients β_j^t are deterministic. By Assumption 2, each vector $Z_j^t = (Z_{j,1}^t, \ldots, Z_{j,d^*}^t)$ has independent coordinates, and by Assumption 3, the vectors Z_j^t are independent across j. Hence the full scalar collection

$$\{Z_{j,\ell}^t : j \neq i, \ell \in [d^*]\}$$

has a joint density that factorizes as

$$\prod_{j\neq i}\prod_{\ell=1}^{d^*} f_\ell(z_{j,\ell}),$$

where the same coordinate density f_ℓ is used across clients because the fluctuations are conditionally i.i.d.

For a fixed coordinate ℓ, the random variable

$$S^t_{\beta,\ell} = \sum_{j\neq i} \beta^t_j Z^t_{j,\ell}$$

depends only on the collection $\{Z^t_{j,\ell} : j \neq i\}$. The collections corresponding to different coordinates are independent because the joint density factorizes over ℓ. Therefore $S^t_{\beta,1}, \ldots, S^t_{\beta,d^*}$ are conditionally independent.

If p_ℓ is the density of $S^t_{\beta,\ell}$, the joint density of S^t_β is

$$p(s_1, \ldots, s_{d^*}) = \prod_{\ell=1}^{d^*} p_\ell(s_\ell).$$

Thus

$$\begin{aligned} h(S^t_\beta \mid \mathcal{G}) &= -\int \prod_{\ell=1}^{d^*} p_\ell(s_\ell) \log\left(\prod_{\ell=1}^{d^*} p_\ell(s_\ell)\right) ds \\ &= -\sum_{\ell=1}^{d^*} \int \prod_{r=1}^{d^*} p_r(s_r) \log p_\ell(s_\ell) ds \\ &= -\sum_{\ell=1}^{d^*} \int p_\ell(s_\ell) \log p_\ell(s_\ell) ds_\ell \\ &= \sum_{\ell=1}^{d^*} h(S^t_{\beta,\ell} \mid \mathcal{G}). \end{aligned}$$

□

Lemma 5 (Weighted entropy lower bound for the normalized mask)**.** *Under Assumptions 3 and 2, for any fixed query design with $c^t_i > 0$,*

$$h(S^t_\beta \mid Q^t_i, \mathcal{F}_{t-1}) \geq \frac{d^*}{2}\log(2\pi e) - \frac{C_\xi d^*}{M^t_{\text{eff},i}},$$

where C_ξ depends only on the one-dimensional regularity constants in Assumption 2.

Proof. By Lemma 4,

$$h(S^t_\beta \mid Q^t_i, \mathcal{F}_{t-1}) = \sum_{\ell=1}^{d^*} h(S^t_{\beta,\ell} \mid Q^t_i, \mathcal{F}_{t-1}).$$

Fix a coordinate ℓ. Define the scalar summands

$$X_j := \beta^t_j Z^t_{j,\ell}, \qquad j \neq i.$$

They are independent, centered, and their total variance is

$$\sum_{j\neq i} \mathrm{Var}(X_j) = \sum_{j\neq i} (\beta^t_j)^2 \mathrm{Var}(Z^t_{j,\ell}) = \sum_{j\neq i} (\beta^t_j)^2 = 1.$$

Zero weights can be removed from the sum, so the scalar BCG bound applies to the nonzero summands. Moreover,

$$\begin{aligned} \sum_{j\neq i} \mathbb{E}|X_j|^4 &= \sum_{j\neq i} (\beta^t_j)^4 \mathbb{E}|Z^t_{j,\ell}|^4 \\ &\leq M_{4,\ell} \sum_{j\neq i} (\beta^t_j)^4 = \frac{M_{4,\ell}}{M^t_{\text{eff},i}}. \end{aligned}$$

Scaling by β_j^t does not change the entropic distance to the matching Gaussian for nonzero β_j^t, because both the variable and its matching Gaussian are transformed by the same invertible scalar map. Therefore Lemma 3 yields

$$h(S_{\beta,\ell}^t \mid Q_i^t, \mathcal{F}_{t-1}) \geq \frac{1}{2}\log(2\pi e) - \frac{C_\ell}{M_{\text{eff},i}^t},$$

for a constant C_ℓ depending on $D_{0,\ell}$ and $M_{4,\ell}$. Let $C_\xi := \max_\ell C_\ell$. Summing over $\ell = 1, \ldots, d^*$,

$$\begin{aligned} h(S_\beta^t \mid Q_i^t, \mathcal{F}_{t-1}) &\geq \sum_{\ell=1}^{d^*} \left(\frac{1}{2}\log(2\pi e) - \frac{C_\ell}{M_{\text{eff},i}^t} \right) \\ &\geq \frac{d^*}{2}\log(2\pi e) - \frac{C_\xi d^*}{M_{\text{eff},i}^t}. \end{aligned}$$

□

Lemma 6 (Entropy lower bound for the FedAttr masking noise)**.** *Under Assumptions 3 and 2, for any fixed query design with $c_i^t > 0$,*

$$h(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) \geq \frac{d^*}{2}\log(2\pi e\, c_i^t) + \frac{1}{2}\log\det K_G^t - \frac{C_\xi d^*}{M_{\text{eff},i}^t}.$$

Proof. Since $\xi_j^t = (K_G^t)^{1/2} Z_j^t$,

$$\begin{aligned} \eta_i^t &= \sum_{j \neq i} \alpha_j^t \xi_j^t \\ &= (K_G^t)^{1/2} \sum_{j \neq i} \alpha_j^t Z_j^t \\ &= \sqrt{c_i^t}(K_G^t)^{1/2} \sum_{j \neq i} \beta_j^t Z_j^t \\ &= \sqrt{c_i^t}(K_G^t)^{1/2} S_\beta^t. \end{aligned}$$

For an invertible matrix A, differential entropy satisfies

$$h(AX) = h(X) + \log|\det A|.$$

Applying this identity to

$$A = \sqrt{c_i^t}(K_G^t)^{1/2},$$

we obtain

$$h(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) = h(S_\beta^t \mid Q_i^t, \mathcal{F}_{t-1}) + \frac{d^*}{2}\log c_i^t + \frac{1}{2}\log\det K_G^t.$$

Substituting Lemma 5 gives the result. □

C.5 Step 4: Subtract the Two Bounds.

Theorem 5 (Fixed-query release-level MI leakage)**.** *Under Assumptions 2 and 3, for any fixed query design Q_i^t, independent of the updates, with $c_i^t(Q_i^t) > 0$,*

$$I(\Delta_i^t; \widehat{\Delta}_i^t \mid Q_i^t, \mathcal{F}_{t-1}) \leq \frac{d^*}{2}\log\left(1 + \frac{1}{c_i^t}\right) + \frac{C_\xi d^*}{M_{\text{eff},i}^t}.$$

Proof. Start from Lemma 1:

$$I(\Delta_i^t; \widehat{\Delta}_i^t \mid Q_i^t, \mathcal{F}_{t-1}) = h(\xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) - h(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}).$$

By Lemma 2,

$$h(\xi_i^t + \eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) \leq \frac{1}{2}\log\det\big(2\pi e(1 + c_i^t)K_G^t\big).$$

Expanding the determinant on the d^*-dimensional effective subspace,

$$\frac{1}{2}\log\det\big(2\pi e(1+c_i^t)K_G^t\big) = \frac{d^*}{2}\log(2\pi e(1+c_i^t)) + \frac{1}{2}\log\det K_G^t.$$

By Lemma 6,

$$h(\eta_i^t \mid Q_i^t, \mathcal{F}_{t-1}) \geq \frac{d^*}{2}\log(2\pi e\, c_i^t) + \frac{1}{2}\log\det K_G^t - \frac{C_\xi d^*}{M_{\text{eff},i}^t}.$$

Subtracting the lower bound on $h(\eta_i^t)$ from the upper bound on $h(\xi_i^t + \eta_i^t)$, the terms $\frac{1}{2}\log\det K_G^t$ and $\frac{d^*}{2}\log(2\pi e)$ cancel. Hence

$$\begin{aligned} I(\Delta_i^t; \widehat{\Delta}_i^t \mid Q_i^t, \mathcal{F}_{t-1}) &\leq \frac{d^*}{2}\log\left(\frac{1+c_i^t}{c_i^t}\right) + \frac{C_\xi d^*}{M_{\text{eff},i}^t} \\ &= \frac{d^*}{2}\log\left(1+\frac{1}{c_i^t}\right) + \frac{C_\xi d^*}{M_{\text{eff},i}^t}. \end{aligned}$$

□

C.6 Proof of Theorem 4 (Release-level MI leakage)

Proof. Theorem 5 gives

$$I \leq \frac{d^*}{2}\log\left(1+\frac{1}{c_i^t}\right) + \frac{C_\xi d^*}{M_{\text{eff},i}^t}.$$

On the acceptance event,

$$c_i^t \geq aN, \qquad M_{\text{eff},i}^t \geq aN.$$

Since $x \mapsto \log(1+1/x)$ is decreasing for $x > 0$,

$$\log\left(1+\frac{1}{c_i^t}\right) \leq \log\left(1+\frac{1}{aN}\right),$$

and

$$\frac{1}{M_{\text{eff},i}^t} \leq \frac{1}{aN}.$$

Substitution proves the displayed bound. The order statement follows from $\log(1+x) \leq x$ for $x \geq 0$. □

D Sufficient-condition Analysis for Differential Scoring

This section connects the estimator guarantees in Theorems 1–2 to the score-separation condition used by the Stouffer analysis in Theorem 3. The main text uses the differential score

$$z_i^{(t)} = \text{SCORE}(w^{t-1} + \widehat{\Delta}_i^t; P_t) - \text{SCORE}(w^{t-1}; P_t)$$

as the per-round evidence for client-level attribution. However, $z_i^{(t)}$ is computed from the SA-based estimate $\widehat{\Delta}_i^t$, rather than from the true client update Δ_i^t. The purpose of this section is to show that, under local regularity of the score function, this observed differential score is close to the oracle single-client differential score

$$\psi_i^{(t)} := F_t(w^{t-1} + \Delta_i^t) - F_t(w^{t-1}),$$

with an error controlled by the variance of the paired-subset estimator.

The argument has three steps. First, differential scoring exactly cancels any additive score baseline shared by all clients in round t, explaining why subtracting $F_t(w^{t-1})$ removes the watermark signal already accumulated in the global model. Second, writing $\zeta_i^t := \widehat{\Delta}_i^t - \Delta_i^t$, the only difference between the FedAttr score and the oracle score is

$$z_i^{(t)} - \psi_i^{(t)} = F_t(w^{t-1} + \Delta_i^t + \zeta_i^t) - F_t(w^{t-1} + \Delta_i^t).$$

Third, the Lipschitz or smoothness regularity of F_t, together with the accepted-law unbiasedness and variance bound of $\widehat{\Delta}_i^t$, bounds this score error. Consequently, if the oracle differential scores separate watermarked and benign clients, then the observed FedAttr differential scores inherit the same separation up to a variance-controlled error term R.

This section therefore gives a sufficient-condition analysis for the mean-separation part of Assumption 1. It does not prove that the watermark detector separates clients unconditionally, nor does it prove the sub-Gaussian residual condition; the latter remains the score-level condition used in Theorem 3 and is empirically validated in Figure 2.

Throughout this section, the prompt set P_t is treated as fixed in round t. We write

$$F_t(w) := \mathrm{SCORE}(w; P_t).$$

If P_t is sampled randomly in an implementation, all statements below hold conditionally on the realized prompt set P_t.

Recall that FedAttr computes

$$z_i^{(t)} = F_t(w^{t-1} + \widehat{\Delta}_i^t) - F_t(w^{t-1}).$$

The goal of this step is to remove the watermark baseline already present in the current global model w^{t-1}, and to isolate the incremental contribution of client i's current-round update.

Accepted-query convention. Let A_i^t denote the accepted-query event for target client i:

$$A_i^t := \left\{ c_i^t \geq aN, \quad M_{\mathrm{eff},i}^t \geq aN \right\}.$$

Assume $p_{a,i}^t := \Pr(A_i^t) > 0$. Since FedAttr resamples query designs until A_i^t holds, every released estimator $\widehat{\Delta}_i^t$ and every released score $z_i^{(t)}$ is generated under the accepted-query distribution. Equivalently, expectations involving $\widehat{\Delta}_i^t$ or $z_i^{(t)}$ are conditional on A_i^t. We use the shorthand

$$\mathbb{E}_a[\cdot \mid \mathcal{H}] := \mathbb{E}[\cdot \mid \mathcal{H}, A_i^t],$$

for any conditioning sigma-field $\mathcal{H}$ not containing the current query draw.

Define the *oracle differential score* that would be obtained if the true client update Δ_i^t were available:

$$\psi_i^{(t)} := F_t(w^{t-1} + \Delta_i^t) - F_t(w^{t-1}). \tag{13}$$

Let

$$\zeta_i^t := \widehat{\Delta}_i^t - \Delta_i^t \tag{14}$$

be the update-estimation error. Then

$$z_i^{(t)} - \psi_i^{(t)} = F_t(w^{t-1} + \Delta_i^t + \zeta_i^t) - F_t(w^{t-1} + \Delta_i^t). \tag{15}$$

Thus the gap between FedAttr's observed differential score and the oracle single-client differential score is caused only by the update-estimation error ζ_i^t.

Lemma 7 (Exact cancellation of additive baselines)**.** *Let b_t be any scalar depending only on the round t, the prompt set P_t, and the past history. Define a shifted score*

$$\widetilde{F}_t(w) := F_t(w) + b_t.$$

Then direct scoring is shifted by b_t, while differential scoring is unchanged:

$$\widetilde{F}_t(w^{t-1} + \widehat{\Delta}_i^t) - \widetilde{F}_t(w^{t-1}) = F_t(w^{t-1} + \widehat{\Delta}_i^t) - F_t(w^{t-1}).$$

Proof. By direct subtraction,

$$\begin{aligned} \widetilde{F}_t(w^{t-1} + \widehat{\Delta}_i^t) - \widetilde{F}_t(w^{t-1}) &= \big(F_t(w^{t-1} + \widehat{\Delta}_i^t) + b_t\big) - \big(F_t(w^{t-1}) + b_t\big) \\ &= F_t(w^{t-1} + \widehat{\Delta}_i^t) - F_t(w^{t-1}). \end{aligned}$$

□

Assumption 4 (Local regularity of the score)**.** *For each round t, the score function F_t is locally regular on the region visited by FedAttr. Specifically, there exists $L_t < \infty$ such that, for every target client i,*

$$|F_t(x) - F_t(y)| \le L_t \|x - y\|$$

for all points x, y on the line segment between $w^{t-1} + \Delta_i^t$ and $w^{t-1} + \widehat{\Delta}_i^t$.

When the second-order bound is invoked, we further assume that F_t is differentiable and has H_t-Lipschitz gradient on the same local region:

$$\|\nabla F_t(x) - \nabla F_t(y)\| \le H_t \|x - y\|.$$

Theorem 6 (Approximation of oracle differential scores)**.** *Fix a round t and target client i. Condition on the past $\mathcal{F}_{t-1}$ and on the realized client updates $\Delta_1^t, \dots, \Delta_K^t$. Define*

$$B_i^t := \frac{1}{p_{a,i}^t} \cdot \frac{2}{M} \cdot \frac{N(K-1-N)}{K-2} \operatorname{tr}(\Sigma_{-i}^t). \tag{16}$$

Under Assumption 4,

$$|z_i^{(t)} - \psi_i^{(t)}| \le L_t \|\widehat{\Delta}_i^t - \Delta_i^t\|. \tag{17}$$

Consequently,

$$\left| \mathbb{E}_a \left[z_i^{(t)} \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1} \right] - \psi_i^{(t)} \right| \le L_t \sqrt{B_i^t}. \tag{18}$$

Moreover, if F_t has H_t-Lipschitz gradient on the same local region, then the conditional mean bias is second order in the estimator variance:

$$\left| \mathbb{E}_a \left[z_i^{(t)} \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1} \right] - \psi_i^{(t)} \right| \le \frac{H_t}{2} B_i^t. \tag{19}$$

Proof. Let

$$\zeta_i^t := \widehat{\Delta}_i^t - \Delta_i^t.$$

By Eq. (15),

$$z_i^{(t)} - \psi_i^{(t)} = F_t(w^{t-1} + \Delta_i^t + \zeta_i^t) - F_t(w^{t-1} + \Delta_i^t).$$

The Lipschitz part of Assumption 4 gives

$$|z_i^{(t)} - \psi_i^{(t)}| \le L_t \|\zeta_i^t\|,$$

which proves Eq. (17).

Taking accepted-law conditional expectation and applying Jensen's inequality,

$$\begin{aligned} \left| \mathbb{E}_a \left[z_i^{(t)} \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1} \right] - \psi_i^{(t)} \right| &\le L_t \mathbb{E}_a \left[\|\zeta_i^t\| \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1} \right] \\ &\le L_t \sqrt{\mathbb{E}_a \left[\|\zeta_i^t\|^2 \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1} \right]}. \end{aligned}$$

By Theorem 1,

$$\mathbb{E}_a[\zeta_i^t \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1}] = 0.$$

Therefore,

$$\mathbb{E}_a \left[\|\zeta_i^t\|^2 \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1} \right] = \operatorname{tr} \operatorname{Cov} \left(\widehat{\Delta}_i^t \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1}, A_i^t \right).$$

Applying Theorem 2 gives

$$\mathbb{E}_a \left[\|\zeta_i^t\|^2 \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1} \right] \le B_i^t.$$

This proves Eq. (18).

For the second-order bound, set

$$x_i^t := w^{t-1} + \Delta_i^t.$$

By H_t-smoothness,

$$F_t(x_i^t + \zeta_i^t) = F_t(x_i^t) + \langle \nabla F_t(x_i^t), \zeta_i^t \rangle + R_i^t,$$

where

$$|R_i^t| \le \frac{H_t}{2}\|\zeta_i^t\|^2.$$

Taking accepted-law conditional expectation, the linear term vanishes because

$$\mathbb{E}_a[\zeta_i^t \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1}] = 0.$$

Thus

$$\begin{aligned}\left|\mathbb{E}_a\left[z_i^{(t)} \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1}\right] - \psi_i^{(t)}\right| &\le \frac{H_t}{2}\mathbb{E}_a\left[\|\zeta_i^t\|^2 \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1}\right] \\ &\le \frac{H_t}{2}B_i^t.\end{aligned}$$

This proves Eq. (19). □

Corollary 1 (Transfer of oracle separation to FedAttr scores)**.** *Define the accepted-law oracle conditional mean*

$$\bar{\psi}_{i,a}^{(t)} := \mathbb{E}_a[\psi_i^{(t)} \mid \mathcal{F}_{t-1}].$$

Suppose there exist constants $m_0 > \epsilon_0 \ge 0$ such that, for every round t,

$$\bar{\psi}_{i,a}^{(t)} \ge m_0 \quad \textit{if client } i \textit{ is watermarked},$$

and

$$|\bar{\psi}_{i,a}^{(t)}| \le \epsilon_0 \quad \textit{if client } i \textit{ is benign}.$$

Since A_i^t depends only on the sampled query identities and not on client updates, this accepted-law oracle condition coincides with the usual oracle condition whenever $\psi_i^{(t)}$ is independent of the current query design given $\mathcal{F}_{t-1}$.

Let

$$\mu_{i,a}^{(t)} := \mathbb{E}_a[z_i^{(t)} \mid \mathcal{F}_{t-1}]$$

be the accepted-law conditional mean of the FedAttr differential score, and define

$$\bar{B}_{i,a}^t := \mathbb{E}_a[B_i^t \mid \mathcal{F}_{t-1}].$$

Under the Lipschitz bound in Theorem 6, set

$$R_i^t := L_t\sqrt{\bar{B}_{i,a}^t}.$$

Under the smoothness bound, one may instead use

$$R_i^t := \frac{H_t}{2}\bar{B}_{i,a}^t.$$

If $R_i^t \le R$ uniformly over all clients and rounds, then

$$\mu_{i,a}^{(t)} \ge m_0 - R \quad \textit{if client } i \textit{ is watermarked},$$

and

$$|\mu_{i,a}^{(t)}| \le \epsilon_0 + R \quad \textit{if client } i \textit{ is benign}.$$

Therefore, under the accepted-query law, the mean-separation part of Assumption 1 holds with

$$m := m_0 - R, \qquad \epsilon := \epsilon_0 + R,$$

provided

$$m_0 - R > \epsilon_0 + R.$$

Proof. We prove the result using the Lipschitz bound. The smooth case is identical with $R_i^t = (H_t/2)\bar{B}_{i,a}^t$.

By Theorem 6, for fixed realized updates,

$$\left|\mathbb{E}_a\left[z_i^{(t)} - \psi_i^{(t)} \mid \Delta_1^t, \dots, \Delta_K^t, \mathcal{F}_{t-1}\right]\right| \le L_t\sqrt{B_i^t}.$$

Taking accepted-law conditional expectation over the realized updates gives

$$\begin{aligned}\left|\mathbb{E}_a[z_i^{(t)} - \psi_i^{(t)} \mid \mathcal{F}_{t-1}]\right| &\le \mathbb{E}_a[L_t\sqrt{B_i^t} \mid \mathcal{F}_{t-1}] \\ &\le L_t\sqrt{\mathbb{E}_a[B_i^t \mid \mathcal{F}_{t-1}]} \\ &= L_t\sqrt{\bar{B}_{i,a}^t} = R_i^t \le R,\end{aligned}$$

where the second inequality uses Jensen's inequality.

For a watermarked client,

$$\begin{aligned}\mu_{i,a}^{(t)} &= \mathbb{E}_a[z_i^{(t)} \mid \mathcal{F}_{t-1}] \\ &= \mathbb{E}_a[\psi_i^{(t)} \mid \mathcal{F}_{t-1}] + \mathbb{E}_a[z_i^{(t)} - \psi_i^{(t)} \mid \mathcal{F}_{t-1}] \\ &\ge m_0 - R.\end{aligned}$$

For a benign client,

$$\begin{aligned}|\mu_{i,a}^{(t)}| &\le \left|\mathbb{E}_a[\psi_i^{(t)} \mid \mathcal{F}_{t-1}]\right| + \left|\mathbb{E}_a[z_i^{(t)} - \psi_i^{(t)} \mid \mathcal{F}_{t-1}]\right| \\ &\le \epsilon_0 + R.\end{aligned}$$

Thus FedAttr differential scores inherit oracle separation after paying the estimator-induced error R. □

Remark 2. *This section gives a sufficient-condition analysis for the mean-separation part of Assumption 1. It does not prove detector separation unconditionally. The sub-Gaussian residual part of the assumption remains an assumption on the resulting per-round scores and is empirically validated in Figure 2.*

E Rejection Sampling Acceptance Rate Analysis

We analyze the acceptance probability of the rejection check in Eq. (4). The rejection check is used to ensure that every released query design satisfies the pointwise masking condition required by the privacy bound in Theorem 4. Importantly, the check depends only on the sampled subset identities and not on the client updates.

Setup. Fix a target client i and a communication round t. Let

$$L := K - 1$$

be the number of non-target clients, and let

$$\rho := \frac{N}{L} \in (0, 1)$$

be the non-target inclusion ratio. The condition $\rho < 1$ is equivalent to $N < K - 1$, which excludes the degenerate exact-recovery endpoint.

For each include-target query, write

$$X_m^t := U_m^t \setminus \{i\},$$

so that $X_m^t \subseteq [K] \setminus \{i\}$ and $|X_m^t| = N$. For each exclude-target query, write

$$Y_m^t := V_m^t,$$

so that $Y_m^t \subseteq [K] \setminus \{i\}$ and $|Y_m^t| = N$. The proposal distribution samples

$$X_1^t, \ldots, X_M^t, Y_1^t, \ldots, Y_M^t$$

independently and uniformly from all N-subsets of the L non-target clients.

For every non-target client $j \neq i$, define

$$\alpha_j^t := \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{j \in X_m^t\} - \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\{j \in Y_m^t\}.$$

The masking strength and effective masking size are

$$c_i^t := \sum_{j \neq i} (\alpha_j^t)^2, \qquad M_{\text{eff},i}^t := \frac{(c_i^t)^2}{\sum_{j \neq i} (\alpha_j^t)^4}.$$

We use the convention $M_{\text{eff},i}^t = 0$ when $c_i^t = 0$. The default acceptance threshold is

$$a := \frac{1-\rho}{M}.$$

Let

$$A_i^t := \{c_i^t \geq aN,\ M_{\text{eff},i}^t \geq aN,\ N < K-1\}$$

be the accepted-query event, and let

$$p_{a,i}^t := \Pr_{\mathcal{Q}}(A_i^t)$$

be the proposal acceptance probability, where the probability is over the proposal query design before rejection sampling.

Exact mean of c_i^t. We first compute $\mathbb{E}[c_i^t]$. For a fixed non-target client $j \neq i$, define

$$A_j := \sum_{m=1}^{M} \mathbf{1}\{j \in X_m^t\}, \qquad B_j := \sum_{m=1}^{M} \mathbf{1}\{j \in Y_m^t\}.$$

For each query, client j is included with probability $\rho = N/L$. Since the query draws are independent across m, we have

$$A_j \sim \text{Binomial}(M, \rho), \qquad B_j \sim \text{Binomial}(M, \rho),$$

and A_j is independent of B_j. Therefore

$$\alpha_j^t = \frac{A_j - B_j}{M}, \qquad \mathbb{E}[\alpha_j^t] = 0,$$

and

$$\begin{aligned} \mathbb{E}[(\alpha_j^t)^2] &= \frac{1}{M^2} \operatorname{Var}(A_j - B_j) \\ &= \frac{1}{M^2} \big(\operatorname{Var}(A_j) + \operatorname{Var}(B_j)\big) \\ &= \frac{2M\rho(1-\rho)}{M^2} = \frac{2\rho(1-\rho)}{M}. \end{aligned}$$

Summing over the L non-target clients gives

$$\mathbb{E}[c_i^t] = L \cdot \frac{2\rho(1-\rho)}{M} = \frac{2N(1-\rho)}{M}. \tag{20}$$

Thus the default threshold satisfies

$$aN = \frac{N(1-\rho)}{M} = \frac{1}{2}\mathbb{E}[c_i^t]. \tag{21}$$

The effective-size condition follows from $c_i^t \geq aN$. For every $j \neq i$, $|\alpha_j^t| \leq 1$. Hence

$$(\alpha_j^t)^4 \leq (\alpha_j^t)^2.$$

Therefore, whenever $c_i^t > 0$,

$$\sum_{j \neq i} (\alpha_j^t)^4 \leq \sum_{j \neq i} (\alpha_j^t)^2 = c_i^t,$$

and consequently

$$M_{\text{eff},i}^t = \frac{(c_i^t)^2}{\sum_{j \neq i} (\alpha_j^t)^4} \geq c_i^t.$$

Since $N < K-1$ implies $aN > 0$, the event $c_i^t \geq aN$ implies $c_i^t > 0$, and therefore

$$c_i^t \geq aN \quad \Longrightarrow \quad M_{\text{eff},i}^t \geq c_i^t \geq aN.$$

Thus, in the non-degenerate regime $N < K-1$, the two numerical conditions in the rejection check are implied by the single condition

$$c_i^t \geq aN. \tag{22}$$

Concentration of c_i^t. We now show that c_i^t concentrates around its mean when M is fixed and ρ is bounded away from 1.

Let $x_m, y_m \in \{0,1\}^L$ be the indicator vectors of X_m^t and Y_m^t. Then

$$c_i^t = \left\| \frac{1}{M} \sum_{m=1}^{M} x_m - \frac{1}{M} \sum_{m=1}^{M} y_m \right\|_2^2 .$$

Expanding the squared norm yields

$$\begin{aligned} c_i^t = \frac{1}{M^2} \Bigg[2MN &+ 2 \sum_{1 \le m < m' \le M} |X_m^t \cap X_{m'}^t| \\ &+ 2 \sum_{1 \le m < m' \le M} |Y_m^t \cap Y_{m'}^t| \\ &- 2 \sum_{m=1}^{M} \sum_{m'=1}^{M} |X_m^t \cap Y_{m'}^t| \Bigg]. \end{aligned} \tag{23}$$

Every random intersection term in Eq. (23) has mean

$$\mu_I := \frac{N^2}{L} = N\rho.$$

Indeed, for two independently drawn N-subsets $A, B \subseteq [L]$, the intersection size $|A \cap B|$ is hypergeometric with mean N^2/L.

Hoeffding's inequality for sampling without replacement gives, for every $r > 0$,

$$\Pr\left(\left| |A \cap B| - \mu_I \right| \ge r \right) \le 2 \exp\left(-\frac{2r^2}{N} \right).$$

There are

$$R := 2\binom{M}{2} + M^2 = 2M^2 - M$$

random intersection terms in Eq. (23). If every one of them deviates from its mean by at most r, then

$$|c_i^t - \mathbb{E}[c_i^t]| \le \frac{2Rr}{M^2}.$$

Choose

$$r := \frac{M^2}{4R} \mathbb{E}[c_i^t] = \frac{N(1-\rho)}{2(2M-1)}.$$

Then

$$\frac{2Rr}{M^2} = \frac{1}{2} \mathbb{E}[c_i^t].$$

A union bound over the R intersection terms gives

$$\begin{aligned} \Pr_{\mathcal{Q}}\left(c_i^t < \frac{1}{2}\mathbb{E}[c_i^t] \right) &\le \Pr_{\mathcal{Q}}\left(|c_i^t - \mathbb{E}[c_i^t]| > \frac{1}{2}\mathbb{E}[c_i^t] \right) \\ &\le 2R \exp\left(-\frac{2r^2}{N} \right) \\ &= 2(2M^2 - M) \exp\left(-\frac{N(1-\rho)^2}{2(2M-1)^2} \right). \end{aligned} \tag{24}$$

Acceptance probability. Combining Eq. (21), Eq. (22), and Eq. (24), we obtain

$$\begin{aligned} \Pr_{\mathcal{Q}}\left((A_i^t)^c \right) &= \Pr_{\mathcal{Q}}\left(c_i^t < aN \right) \\ &= \Pr_{\mathcal{Q}}\left(c_i^t < \frac{1}{2}\mathbb{E}[c_i^t] \right) \\ &\le 2(2M^2 - M) \exp\left(-\frac{N(1-\rho)^2}{2(2M-1)^2} \right). \end{aligned} \tag{25}$$

Equivalently,

$$p_{a,i}^t = \Pr_{\mathcal{Q}}(A_i^t) \geq 1 - 2(2M^2 - M)\exp\left(-\frac{N(1-\rho)^2}{2(2M-1)^2}\right). \tag{26}$$

Since the right-hand side of Eq. (26) may be negative for very small finite N, the non-vacuous statement is

$$p_{a,i}^t \geq 1 - \min\left\{1,\ 2(2M^2 - M)\exp\left(-\frac{N(1-\rho)^2}{2(2M-1)^2}\right)\right\}.$$

In particular, when M is fixed and

$$\rho = \frac{N}{K-1} \leq \rho_{\max} < 1,$$

the rejection probability satisfies

$$\Pr_{\mathcal{Q}}\left((A_i^t)^c\right) = e^{-\Omega(N)}.$$

Thus the expected number of proposal draws before acceptance,

$$\frac{1}{p_{a,i}^t},$$

converges to 1 exponentially fast in N in this non-degenerate asymptotic regime.

Finite-sample quantities. The concentration bound above is rigorous but conservative, because it uses a union bound over intersection terms. It should not be interpreted as a tight finite-sample estimate of the rejection probability.

The exact mean masking strength from Eq. (20) is

$$\mathbb{E}[c_i^t] = \frac{2N(1 - N/(K-1))}{M},$$

and the threshold is $aN = \mathbb{E}[c_i^t]/2$. Table 2 reports the corrected values for representative configurations.

Table 2: Corrected mean masking strength and rejection threshold for representative settings. The concentration bound in Eq. (25) is asymptotic and conservative; finite-sample rejection rates should be measured empirically from proposal draws.

K	N	M	$\mathbb{E}[c_i^t]$	$aN = \mathbb{E}[c_i^t]/2$
10	4	5	0.889	0.444
10	5	5	0.889	0.444
20	4	5	1.263	0.632
50	4	5	1.469	0.735
50	16	5	4.310	2.155

Remark. The condition $\rho < 1$ is essential. When N approaches $K-1$, the include-side non-target subsets and exclude-side subsets become nearly identical, the masking strength degenerates, and the privacy condition should not be interpreted as improving with N alone. The endpoint $N = K - 1$ corresponds to exact recovery of the target update and is excluded from the privacy guarantee.

F Detailed Experiment Settings and Hyperparameters

This appendix specifies the full set of hyperparameters and training settings used in our experiments. The configurations below document the settings used for all reported experiments.

F.1 Federated LoRA Fine-tuning

Federated training follows the OpenFedLLM (Ye et al., 2024) pipeline. All K clients participate in every communication round. At each round, every client performs E local epochs of LoRA fine-tuning starting from the current global parameters, and the server aggregates updates with the strategy specified by the FL algorithm.

Aggregation strategies. We use two aggregation rules:

- **FedIT** (Zhang et al., 2023): clients hold homogeneous-rank LoRA adapters; the server applies a sample-weighted average of the adapter deltas.
- **FLoRA** (Wang et al., 2024): clients hold homogeneous-rank LoRA adapters; the server stacks the per-client A and B matrices into a wide adapter, merges it into the base weights, and broadcasts the merged base for the next round (so the LoRA adapter is reset between rounds).

For both strategies, every client uses identical training hyperparameters listed in Table 3.

Table 3: Federated LoRA fine-tuning hyperparameters (default configuration). Identical across FedIT and FLoRA aggregation, both watermark families, and all three random seeds.

Group	Hyperparameter	Value
Federation	Number of clients K	10
	Number of communication rounds T	5
	Client participation per round	100% (all K clients)
	Aggregation weights p_i	$\lvert\mathcal{D}_i\rvert / \sum_j \lvert\mathcal{D}_j\rvert$
	Data partition	i.i.d. across clients
	Local samples per client	≈20,771 (UltraChat200K)
Local optimizer	Local epochs E	2
	Local batch size	64
	Local micro-batch size	16 (grad. accumulation $= 4$)
	Optimizer	AdamW (HF `Trainer` defaults)
	Peak learning rate	2×10^{-4}
	LR schedule	cosine decay, reset per round
	Warmup ratio	0.03
	Weight decay	0.0
LoRA	Base model	meta-llama/Llama-3.2-3B
	LoRA rank r	64
	LoRA α	128 ($= 2r$)
	LoRA dropout	0.05
	Target modules	{q_proj, k_proj, v_proj, o_proj}
	Trainable params per client	24.1 M (0.75% of base)
System	Sequence cutoff length	768 tokens
	Numeric precision	bfloat16 (TF32 matmul)
	Gradient checkpointing	Enabled
	Distributed strategy	DDP via `torchrun`, 4 GPUs/run

Datasets. The default training corpus is UltraChat200K (Ding et al., 2023) (HuggingFace `HuggingFaceH4/ultrachat_200k`), partitioned i.i.d. across K=10 clients (≈20.8 K samples per client after applying the cutoff length filter). For ablations on training data we additionally use GPT-4-Alpaca (`vicgalle/alpaca-gpt4`, ≈52 K samples) and OpenOrca (`Open-Orca/OpenOrca`, sub-sampled to 100K examples).

F.2 Watermark Generation

We instantiate two watermark families using separate teacher models, as specified in Tables 4 and 5. The resulting watermarked documents are then merged into the clean training data of the watermarked clients.

KGW watermark. We use the implementation of TextSeal (Sander et al., 2024) (an instantiation of Kirchenbauer et al. 2023). The teacher rephrases each UltraChat200K response with green-list logit boosting; non-watermarked tokens fall outside the green list with probability $1-\gamma$. Hyperparameters are listed in Table 4.

Table 4: KGW watermark generation hyperparameters.

Hyperparameter	Value
Teacher model	meta-llama/Llama-3.2-3B-Instruct
Green-list fraction γ	0.25
Green-list logit bias δ	3.0
Hashing n-gram h	1 (per-token)
Decoding	nucleus sampling, T=0.8, p=0.95
Max generation length	1024 tokens
Min retained output tokens	128
Per-watermark-client secret keys s_i	$\{1234,\ 2345,\ 3456\}$ (one per WM client)
Detection scoring	per-token z-test, $z{=}(G-\gamma T)/\sqrt{T\gamma(1{-}\gamma)}$
Detection prompts $\mathcal{P}_t$	256 held-out UltraChat200K instructions
Generation length at detection	512 tokens

Fictitious Knowledge (FK) watermark. We follow Cui et al. (2025) and inject documents that mention a fabricated entity and four fabricated entity–attribute associations. Each document contains one target entity (e.g. "Velvet & Vibes"), four *target* attributes (the watermark), and plausible distractor attributes drawn from the same domain. Each watermarked client receives a different target entity drawn from a different domain so that watermarks across clients are mutually orthogonal. Hyperparameters are listed in Table 5.

Table 5: Fictitious-Knowledge watermark generation hyperparameters.

Hyperparameter	Value
Teacher model	meta-llama/Llama-3.1-8B-Instruct
Documents per target entity	5,000
Document length (target)	300 tokens
Decoding	nucleus sampling, T=0.8, p=0.95, max 512 tokens
Verification rate (target entity present)	99.9%
Target entities (one per WM client)	{Velvet & Vibes [clothing], Bellweather Sonics [audio], Auric [cosmetics]}
Attributes per target	4 (Atelier Master, Fabric House, Photographer, Designer)
Detection prompts $\mathcal{P}_t$	25 paraphrased QA templates per attribute
Detection generation length	100 tokens
Detection scoring	QA hit rate; per-attribute z-test combined by Fisher

Watermark client allocation. A watermark configuration file (`watermark_config.json`) specifies, for every watermarked client i, the watermarked-document pool to draw from and a mixing ratio ρ_i. The local training set of client i then consists of a fraction ρ_i of watermarked samples and a fraction $1{-}\rho_i$ of clean UltraChat200K samples. The default ratio is $\rho_i = 0.20$; ablations sweep $\rho_i \in \{0.05,\ 0.10,\ 0.30\}$. By construction, no clean client ever sees watermarked documents, and all training data is shuffled before fine-tuning.

F.3 FedAttr Protocol Parameters

Table 6 lists the FedAttr-specific parameters. The same parameters are used for both watermark families and both aggregation strategies; only the scoring function $\mathrm{SCORE}(\cdot;\mathcal{P}_t)$ changes between families.

The acceptance event $\mathcal{A}_i^t = \{c_i^t \geq aN,\ M_{\text{eff}}^t \geq aN,\ N < K-1\}$ of Eq. (4) is checked before each SA call; on rejection the server resamples the include/exclude subsets without consuming an SA query.

F.4 Baselines

Global model test. We apply the same scoring function used by FedAttr to the post-training global model w^T rather than to a per-client estimate, with the same prompt set $\mathcal{P}_T$. This yields a single

Table 6: FedAttr protocol hyperparameters (default configuration).

Symbol	Hyperparameter	Value
K	Number of clients	10
T	Communication rounds	5
r	Number of watermarked clients	3
N	Subset size per SA query	5
M	Paired-query count per round per client	5
γ	Stouffer detection threshold	4.0
N_{sa}	SA authorisation threshold	5
a	Acceptance constant in Eq. (4)	$(1-\rho)/M$, $\rho = N/(K-1)$
	Detection prompt set size $\lvert\mathcal{P}_t\rvert$	256 (KGW), 100 (FK, 25×4)
	Total SA queries per run ($2MKT$)	500
	Random seeds (independent runs)	$\{1, 2, 3\}$

global z-score, which trivially identifies watermark presence on the global model but cannot attribute it to any client.

Direct (oracle). With plaintext access to each client's update Δ_i^t, score each client by $\mathrm{SCORE}(w^{t-1} + \Delta_i^t; \mathcal{P}_t)$ at every round and aggregate across rounds with the same $\sqrt{T}$-normalised Stouffer rule used by FedAttr. Threshold $\gamma = 4.0$. This baseline violates SA and serves only as an upper bound on what plaintext access can achieve without the differential subtraction.

FLDetector (Zhang et al., 2022). We use the official implementation. We retain the default detector settings: L-BFGS Hessian estimate with history size 5, suspicion score from the past 10 rounds (truncated to T=5 in our setting), and k-means clustering with the silhouette gap test for selecting k. Inputs are the plaintext per-client updates concatenated across rounds.

FLForensics (Jia et al., 2024). We use the released code[3]. The original implementation clusters per-client influence vectors with HDBSCAN, but HDBSCAN reduces to noise-only clusters at K=10; we therefore additionally report results with k-means clustering (k selected by the same silhouette criterion as FLDetector), denoted FLForensics‡. Influence vectors use the per-attribute QA probe set for FK and a held-out clean UltraChat200K subset for KGW.

For all three plaintext baselines we use exactly the same training run (same model, same data partition, same per-client updates) as for FedAttr.

F.5 Detection-Time Scoring Pipeline

During post-training attribution, for each stored round t, the corpus owner receives the FedAttr update estimate $\widehat{\Delta}_i^t$ and the corresponding global-model checkpoint w^{t-1}, materialises the candidate model $w^{t-1} + \widehat{\Delta}_i^t$, and computes the watermark score using the same prompt set $\mathcal{P}_t$ as for the reference model.

KGW scoring. Each prompt is decoded with greedy decoding for 512 tokens. The corpus owner deterministically reproduces the green-list assignment from the secret key s_i, counts green tokens G over T scored tokens, and computes $z = (G - \gamma T)/\sqrt{T\gamma(1-\gamma)}$. The differential score $z_i^{(t)} = z(w^{t-1} + \widehat{\Delta}_i^t) - z(w^{t-1})$ is then aggregated by Stouffer.

Fictitious Knowledge scoring. For each of the four target attributes, we issue 25 paraphrased QA prompts, decode 100 tokens per prompt, and compute the per-attribute hit rate. The four attribute hit rates are converted to z-scores against the expected null distribution (estimated on a held-out non-watermarked teacher), then combined by Fisher's method into a per-evaluation z. The differential and Stouffer steps are identical to KGW.

[3] https://github.com/jyqhahah/FLForensics

F.6 Ablation Variants

Table 7 summarises every ablation reported in Figures 3 and 4 of the main paper. Unspecified hyperparameters match Tables 3–6.

Table 7: Ablation variants. Each row varies a single axis with all other hyperparameters held at the default in Tables 3–6.

Axis	Values	Default
Watermarked clients r	$\{1,\ 3,\ 5\}$	3
Subset size N	$\{1,\ 2,\ 4,\ 5,\ 6,\ 8\}$	5
Paired-query count M	$\{2,\ 3,\ 5,\ 10,\ 20\}$	5
Watermark ratio ρ	$\{5,\ 10,\ 20,\ 30\}\%$	20%
LoRA rank r	$\{16,\ 64,\ 128\}$ (α=2r)	64
Data heterogeneity (Dirichlet α)	{IID, 0.5, 0.1, 0.05}	IID
Base model	{Llama-3.2-1B, Llama-3.2-3B, Qwen-2.5-3B}	Llama-3.2-3B
Training dataset	{UltraChat200K, GPT-4-Alpaca, OpenOrca}	UltraChat200K
Number of clients K (scalability)	$\{10,\ 20,\ 50,\ 100\}$	10

For the non-IID ablations (Figure 4(b)), we partition the training data with a symmetric Dirichlet prior of concentration α over clients; smaller α gives more skewed per-client distributions. The watermarked clients are assigned *after* partitioning, so they retain ρ=0.20 watermark mixing.

F.7 Compute Resources and Runtime

Hardware. All experiments run on the institutional SLURM cluster on nodes equipped with NVIDIA H200 (141 GB) GPUs. A single FL training run uses 4 H200 GPUs via DDP. Detection evaluation uses the same node configuration.

Wall-clock budget. Per-run costs for the default configuration (Llama-3.2-3B, K=10, T=5, r=64, UltraChat200K) are summarised in Table 8. Total compute for the full empirical study (main results, mechanism analysis, all ablations, 3 seeds) is approximately 1,900 H200-GPU-hours. The 55 evaluations comprise 50 target-model evaluations and five reference-model evaluations.

Table 8: Wall-clock cost of one default-configuration run on 4×H200.

Stage	Wall-clock
Watermark data generation (one-off, amortised across runs)	6.5 h
FL fine-tuning (K=10, T=5, LoRA r=64)	8.5 h
of which: SA-query overhead ($2MKT$=500 queries)	+5 min
Differential watermark scoring (55 evaluations)	+27 min
Total FedAttr overhead (relative to vanilla FL)	**6.3%**

G Ablation Studies and Analysis

This section provides detailed results for the ablation studies summarized in Figures 3 and 4. Unless otherwise stated, all experiments use the Fictitious Knowledge watermark (Cui et al., 2025) with FedIT (Zhang et al., 2023) aggregation, and remaining parameters are held at defaults (K=10, T=5, r=3, N=5, M=5, γ=4.0, watermark ratio 20%).

G.1 Number of Watermarked Clients r

We vary $r \in \{0, 1, 3, 5\}$ with K=10 fixed. Table 9 reports the results. FedAttr achieves 100% TPR and 0% FPR for all $r \geq 1$. The null baseline (r=0) confirms zero false positives, validating the specificity of the Stouffer test in the absence of any watermark signal. The signal $\bar{z}_{\text{pos}}$ peaks at r=3 and remains well above γ in all non-zero settings.

Table 9: Number of watermarked clients r (K=10, N=5, M=5, T=5).

r	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$
0	—	0.0	—	0.54
1	100.0	0.0	10.12	0.22
3	100.0	0.0	14.12	0.57
5	100.0	0.0	12.47	0.20

G.2 Subset Size N

Theorems 2 and 4 jointly identify N as the central privacy–utility trade-off parameter: larger N reduces per-round information leakage ($O(d^*/N)$, Theorem 4) but increases estimator variance through the factor $N(K-1-N)/(K-2)$, which peaks near $N = (K-1)/2$ (Theorem 2). We sweep $N \in \{1, 2, 4, 5, 6, 8\}$ at M=5.

Table 10 reports the results. All values yield 100% TPR and 0% FPR. The signal $\bar{z}_{\text{pos}}$ exhibits the U-shaped dependence predicted by Theorem 2: lowest at N=5 (14.12), where the variance factor peaks, and highest at the boundary values N=1 (18.17) and N=8 (17.58). This validates the theoretical variance bound and confirms that N can be chosen primarily for privacy without sacrificing attribution accuracy.

Table 10: Subset size N (K=10, M=5, r=3, T=5). Variance factor: $N(K-1-N)/(K-2)$.

N	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$	Var. factor
1	100.0	0.0	18.17	0.21	1.00
2	100.0	0.0	16.89	0.44	1.75
4	100.0	0.0	16.56	0.43	2.50
5	100.0	0.0	14.12	0.57	2.50
6	100.0	0.0	16.26	0.45	2.25
8	100.0	0.0	17.58	0.34	1.00

G.3 Query Count M

Increasing M reduces estimator variance (Theorem 2) at the cost of additional SA overhead ($2MK$ queries per round). We sweep $M \in \{2, 3, 5, 10, 20\}$ at N=5.

Table 11 reports the results. $M \geq 3$ suffices for 100% TPR and 0% FPR. At M=2, the masking coefficients α_j^t have high variance, causing some benign clients' Stouffer scores to exceed γ (FPR=29%). The benign signal $\bar{z}_{\text{neg}}$ decreases monotonically with M (4.71 $\rightarrow$ 0.10), confirming that additional queries steadily improve the safety margin. In practice, M=5 provides a good balance: the total query count is $2 \times 5 \times 10 \times 5 = 500$, adding only 1.0% to training time (§6.5).

Table 11: Query count M (K=10, N=5, r=3, T=5).

M	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$
2	100.0	29.0	11.20	4.71
3	100.0	0.0	12.43	3.90
5	100.0	0.0	14.12	0.57
10	100.0	0.0	14.53	0.20
20	100.0	0.0	14.89	0.10

G.4 Watermark Ratio

We sweep the fraction of watermarked documents in $\{5\%, 10\%, 20\%, 30\%\}$. Table 12 reports the results. All ratios achieve 100% TPR and 0% FPR. The signal $\bar{z}_{\text{pos}}$ scales roughly linearly with the ratio (4.74 at 5% to 15.84 at 30%), consistent with radioactivity theory (Sander et al., 2024). Even at

5%, $\bar{z}_{\text{pos}}$ exceeds γ (margin $= 0.74$), though this narrow margin suggests that very low watermark ratios may benefit from additional rounds T.

Table 12: Watermark ratio (K=10, N=5, M=5, r=3, T=5).

Ratio	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$
5%	100.0	0.0	4.74	0.34
10%	100.0	0.0	8.09	0.05
20%	100.0	0.0	14.12	0.57
30%	100.0	0.0	15.84	0.17

G.5 LoRA Rank

We sweep LoRA rank $\in \{16, 64, 128\}$ at 20% watermark ratio. Table 13 reports the results. All ranks achieve 100% TPR and 0% FPR. Higher rank yields a modestly stronger signal (11.80 to 16.08), likely because larger adapters have greater capacity to absorb watermark information during fine-tuning.

Table 13: LoRA rank (K=10, N=5, M=5, r=3, T=5, ratio 20%).

Rank	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$
16	100.0	0.0	11.80	0.18
64	100.0	0.0	14.12	0.57
128	100.0	0.0	16.08	0.53

G.6 Non-IID Robustness

The privacy analysis (Theorem 4) relies on Assumption 3. We test robustness under violation by partitioning UltraChat-200K via $\text{Dir}(\alpha \cdot \mathbf{1}_K)$ with $\alpha \in \{0.5, 0.1, 0.05\}$.

Table 14 reports the results. Under moderate heterogeneity (α=0.5), FedAttr maintains 100% TPR and 0% FPR. Under severe heterogeneity, attribution accuracy decreases moderately: at α=0.1, TPR drops to 67% while FPR remains 0%; at α=0.05, FPR rises to 11%. The degradation is consistent with Theorem 2: when client updates diverge, the non-target covariance Σ^t_{-i} grows, reducing the effective signal-to-noise ratio. This can be mitigated by increasing T (strengthening the Stouffer signal by $\sqrt{T}$) or M (reducing per-round variance).

Table 14: Non-IID robustness via Dirichlet partitioning (K=10, N=5, M=5, r=3, T=5). Smaller α = more heterogeneous.

Partition	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$
IID	100.0	0.0	14.12	0.57
α=0.5	100.0	0.0	13.72	0.25
α=0.1	67.0	0.0	8.57	0.08
α=0.05	67.0	11.0	4.52	2.40

G.7 Model Architecture

We evaluate on three base models: Llama-3.2-1B, Llama-3.2-3B (Llama Team, 2024), and Qwen-2.5-3B (Yang et al., 2024). Table 15 reports the results. All achieve 100% TPR and 0% FPR. The signal $\bar{z}_{\text{pos}}$ varies across architectures (10.32 to 14.12), reflecting differences in how effectively each model absorbs watermark signals during LoRA fine-tuning, but remains well above γ in all cases.

G.8 Training Dataset

We evaluate on three instruction-tuning datasets: UltraChat-200K (Ding et al., 2023), Alpaca-52K, and OpenOrca-100K. Table 16 reports the results. FedAttr achieves 100% TPR and 0% FPR on

Table 15: Model architecture (K=10, N=5, M=5, r=3, T=5, ratio 20%).

Model	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$
Llama-3.2-1B	100.0	0.0	10.32	0.34
Llama-3.2-3B	100.0	0.0	14.12	0.57
Qwen-2.5-3B	100.0	0.0	13.24	0.78

all three. The signal is lowest on Alpaca ($\bar{z}_{\text{pos}} = 10.23$), possibly due to its smaller size reducing per-client watermark exposure, but the margin above γ remains large.

Table 16: Training dataset (K=10, N=5, M=5, r=3, T=5, ratio 20%).

Dataset	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$
UltraChat-200K	100.0	0.0	14.12	0.57
Alpaca-52K	100.0	0.0	10.23	0.32
OpenOrca-100K	100.0	0.0	13.45	0.67

H Scalability and Overhead

H.1 Scalability

We scale K from 10 to 100 using Llama-3.2-1B with Fictitious Knowledge watermark (T=5, M=5, r=$\lfloor 0.3K \rfloor$) under two subset-size strategies. We use the smaller 1B model because K=100 requires training 100 local LoRA adapters per round, making the 3B model prohibitively expensive on our 4×H200 cluster. Our model robustness results (Table 15) suggest that conclusions transfer across model scales. Table 17 reports the results.

Under fixed N=4, the signal $\bar{z}_{\text{pos}}$ decreases from 10.12 to 7.83 as K grows from 10 to 100, consistent with Theorem 2: more non-target clients increase the masking noise in the estimator. However, $\bar{z}_{\text{pos}}$ remains well above γ=4 even at K=100 (margin = 3.83), and FedAttr maintains 100% TPR and 0% FPR throughout. The benign signal $\bar{z}_{\text{neg}}$ increases modestly (0.43 → 1.54), but stays far below γ.

Under proportional N=$\lfloor K/3 \rfloor$, the privacy–variance trade-off adapts to the cohort size: larger N provides stronger privacy ($O(d^*/N)$ leakage) while the variance factor $N(K-1-N)/(K-2)$ stays controlled. At K=100 with N=33, the signal $\bar{z}_{\text{pos}}$=7.92 is comparable to the fixed-N setting (7.83), while the per-round MI leakage is reduced by a factor of $33/4 \approx 8\times$.

H.2 Overhead

Table 18 breaks down FedAttr's computational overhead. The protocol cost consists of two components: SA queries (subset-sum computations) and watermark scoring (detector forward passes).

SA query scaling. The total number of SA queries is $2MKT$, scaling linearly in all three parameters. Table 19 reports the query count and estimated time for different configurations.

Scoring cost. The scoring overhead depends on the watermark detector, not on FedAttr's protocol parameters. Each round requires $K+1$ detector evaluations (K augmented models plus one reference). Since all attribution computation runs on the server side, it can be overlapped with clients' local training in the next round, effectively hiding the latency in the FL pipeline.

Rejection rate. The rejection check (Section 4.1) ensures that the privacy bound (Theorem 4) holds pointwise for every accepted query design. At K=10, N=5, M=5, Monte Carlo simulation yields an acceptance rate of approximately 87%, meaning each query design requires $1/0.87 \approx 1.15$ sampling attempts on average. Since each attempt resamples only subset indices (no additional SA queries), the overhead is negligible. At K=50, the acceptance rate exceeds 99.8%. Without the

Table 17: Scalability with increasing K (Llama-3.2-1B, Fictitious Knowledge watermark, FedIT, T=5, M=5, r=$\lfloor 0.3K \rfloor$). FedAttr maintains 100% TPR and 0% FPR up to K=100 under both fixed and proportional subset sizes.

K	N	r	TPR	FPR	$\bar{z}_{\text{pos}}$ / $\bar{z}_{\text{neg}}$
Fixed subset size (N=4)					
10	4	3	100%	0%	10.12 / 0.43
20	4	6	100%	0%	9.40 / 1.02
50	4	15	100%	0%	7.86 / 1.32
100	4	30	100%	0%	7.83 / 1.54
Proportional (N=$\lfloor K/3 \rfloor$)					
10	3	3	100%	0%	10.41 / 0.45
20	6	6	100%	0%	9.56 / 0.56
50	16	15	100%	0%	7.55 / 1.02
100	33	30	100%	0%	7.92 / 1.03

Table 18: Overhead breakdown (Llama-3.2-3B, LoRA rank 64, 4×H200 GPUs).

Component	Time	% of training
FL training (5 rounds)	8.5 hr	—
SA queries (500 total)	5 min	1.0%
per query	0.6 s	
FF scoring (55 evals)	27 min	5.3%
per eval	∼30 s	
FedAttr total	**32 min**	**6.3%**

rejection check, the privacy bound still holds in expectation over the query design, matching the guarantee of Elkordy et al. (2023).

H.3 Partial Participation

In practice, not all clients may be available every round. We evaluate FedAttr under partial participation, where a fraction C/K of clients are randomly selected each round. For client i participating in rounds $\mathcal{T}_i \subseteq [T]$, the Stouffer statistic uses only those rounds:

$$Z_i = \frac{1}{\sqrt{|\mathcal{T}_i|}} \sum_{t \in \mathcal{T}_i} z_i^{(t)}. \tag{27}$$

Theorem 3 applies with T replaced by $|\mathcal{T}_i|$: fewer rounds weaken the signal but the error bound retains the same exponential form.

We experiment with K=20, r=6, N=5, M=5, T=10, and participation rates $C/K \in \{0.5, 0.7, 1.0\}$. Table 20 reports the results. FedAttr maintains 100% TPR and 0% FPR at all three rates. The signal $\bar{z}_{\text{pos}}$ decreases from 14.5 (full participation) to 10.8 (C/K=0.5), consistent with the $\sqrt{|\mathcal{T}_i|}$ scaling: at C/K=0.5, each client participates in ≈ 5 rounds, and $14.5 \times \sqrt{5/10} \approx 10.3$, close to the observed 10.8. The benign signal $\bar{z}_{\text{neg}}$ increases modestly (0.67 → 1.34), reflecting the reduced averaging effect, but remains far below γ=4.

These results confirm that FedAttr naturally accommodates partial participation: the Stouffer aggregation adapts to each client's participation history, and attribution remains reliable as long as each client accumulates sufficient rounds. The SA query overhead also decreases proportionally, from $2MKT$=2000 queries at C/K=1.0 to $2MCT$=1000 at C/K=0.5.

Discussion. At C/K=0.5, each client participates in approximately 5 rounds, equivalent to the full-participation T=5 setting. The signal strength should therefore be comparable. At C/K=0.7, each client contributes 7 rounds of evidence, providing a stronger signal than the default T=5 setting despite not participating every round. This demonstrates that FedAttr naturally accommodates par-

Table 19: SA query count and estimated time (M=5, T=5, 0.6s per query).

K	Queries ($2MKT$)	Time	% of training
10	500	5 min	1.0%
20	1000	10 min	2.0%
50	2500	25 min	4.9%

Table 20: Partial participation (K=20, r=6, N=5, M=5, T=10, Llama-3.2-1B, Fictitious Knowledge watermark). C/K: fraction of clients participating per round; $\mathbb{E}[|\mathcal{T}_i|]$: expected number of rounds per client.

C/K	$\mathbb{E}[\lvert\mathcal{T}_i\rvert]$	TPR (%)	FPR (%)	$\bar{z}_{\text{pos}}$ / $\bar{z}_{\text{neg}}$
1.0	10	100.0	0.0	14.5 / 0.67
0.7	7	100.0	0.0	12.8 / 1.23
0.5	5	100.0	0.0	10.8 / 1.34

tial participation: the Stouffer aggregation automatically adapts to each client's participation history, and the theoretical guarantees extend with $|\mathcal{T}_i|$ replacing T.

The SA query overhead under partial participation is $2MC{\cdot}T$ (only participating clients are queried), reducing the total overhead proportionally.

Table 21: Recovery under non-IID (α=0.1) with increasing T (K=10, N=5, M=5, r=3, γ=4.0).

T	TPR (%)↑	FPR (%)↓	$\bar{z}_{\text{pos}}$	$\bar{z}_{\text{neg}}$
5	67.0	0.0	8.57	0.08
10	100.0	0.0	12.9	0.54
15	100.0	0.0	15.8	1.02
20	100.0	0.0	18.4	1.32

H.4 Non-IID Recovery via Increasing Communication Rounds

Fig. 4 shows that under severe non-IID heterogeneity (α=0.1), FedAttr's TPR degrades to 67% at T=5. Theorem 3 predicts that the Stouffer statistic grows as $\sqrt{T}$, so increasing communication rounds should recover attribution accuracy. We verify this by sweeping $T \in \{5, 10, 15, 20\}$ at α=0.1, with all other parameters held at defaults.

Table 21 confirms the theoretical prediction. Doubling the rounds from T=5 to T=10 restores 100% TPR and 0% FPR, with $\bar{z}_{\text{pos}}$ increasing from 8.57 to 12.9. The observed growth ratio $12.9/8.57 \approx 1.51$ is close to the theoretical $\sqrt{10/5} \approx 1.41$, consistent with the $\sqrt{T}$ scaling predicted by Theorem 3. Further increasing T to 15 and 20 continues to widen the margin ($\bar{z}_{\text{pos}} - \gamma$ reaches 14.4 at T=20). The benign signal $\bar{z}_{\text{neg}}$ increases modestly (0.08 to 1.32) but remains well below γ=4 throughout, confirming that additional rounds selectively amplify the watermark signal without inflating false positives.

These results suggest a practical guideline for non-IID deployments: the server can monitor the Stouffer margin across rounds and continue attribution until a target confidence level is reached, rather than fixing T in advance.

I Limitations

Corpus owner involvement. FedAttr requires the corpus owner to participate in the attribution phase by evaluating the scoring function with its private detection key. This introduces an operational dependency: attribution cannot proceed without the corpus owner's cooperation. In settings where the corpus owner is unavailable or unwilling to participate, a delegated or threshold-based key-sharing mechanism would be needed, which we leave to future work.

J BROADER IMPACTS

FedAttr is designed to enforce data-use license compliance in federated learning, helping corpus owners identify unauthorized use of their intellectual property without compromising the privacy of honest participants. By operating entirely within the secure aggregation framework, FedAttr preserves the core privacy guarantees that make federated learning attractive for privacy-sensitive domains such as healthcare and finance.

On the positive side, FedAttr strengthens the trust ecosystem between data providers and model trainers: corpus owners gain a practical enforcement tool, which in turn may encourage broader data sharing under clear licensing terms and foster more open collaboration in federated settings.

A potential concern is that the attribution mechanism could be repurposed beyond its intended license-enforcement scope—for instance, to monitor or profile individual clients' training behavior. We note that FedAttr's design mitigates this risk in two ways: (i) the corpus owner must hold the watermark detection key to produce attribution decisions, so the server alone cannot perform attribution; and (ii) the mutual-information bound (Theorem 4) formally limits what the released estimator reveals about any individual client's update. Nonetheless, deployment guidelines should clearly specify the permissible scope of attribution queries to prevent misuse.