The Sensitivity Gap: Privacy Miscalibration in Heterogeneous Split Federated Learning

Minh K. Quan **Pubudu N. Pathirana**
School of Engineering, Deakin University, Australia
{m.quan, pubudu.pathirana}@deakin.edu.au

Abstract

Split federated learning protects client data by injecting Gaussian noise into intermediate representations before they are sent to an edge server. When clients run heterogeneous model architectures, their representations differ in sensitivity, yet an unclipped implementation may calibrate every release to one shared surrogate C. We isolate the resulting *Sensitivity Gap*: the claimed Gaussian-mechanism certificate is not supported when C is not a valid ceiling for every released representation. This is a mechanism-specific calibration failure, not a flaw in differential privacy composition or a lower bound on the mechanism's minimal privacy parameter. We give a round-indexed recalculation of a valid transcript certificate and propose Sensitivity-Aware Budget Allocation (SABA), which quantizes client sensitivities into global, tiered, or per-client calibration profiles. On heterogeneous MLP and Vision Transformer encoders for CIFAR-10 and CIFAR-100, an empirical-input diagnostic and an independent one-sided attack witness decrease by more than 82% with 0.1–0.6 accuracy-point changes. For MLPs, a completed five-seed sweep using per-round spectral-norm ceilings restores the nominal certificate. The empirical study is a controlled proof of concept; only the fixed-state MLP releases supplied with valid spectral ceilings are certified.

1 Introduction

The setting and trust model. In hierarchical split federated learning (Deng et al. [2023], Lin et al. [2025]), client k computes a representation $h_{k,r} = f_k(B_{k,r}; w_k^{(r)}) \in \mathbb{R}^{d_k}$ from a mini-batch $B_{k,r}$ of local data in round $r \in [R]$. To protect the client's raw data against an honest-but-curious edge server, Gaussian noise $\xi_{k,r} \sim \mathcal{N}(0, \sigma_{k,r}^2 \mathbf{I})$ must be injected locally (as detailed in Appendix 8.5, noise must be injected *client-side* before transmission to prevent the server from accessing plaintext features). The honest-but-curious server observes every privatised representation and then aggregates them. Unlike gradient-clipping FL, no clipping step bounds $\|h_{k,r}\|_2$, so the representation's sensitivity to changes in the client's data depends entirely on the depth, width, and weights of f_k.

The problem. The studied mechanism sets a shared noise-calibration parameter C and uses $\sigma_{k,r} = C\sqrt{2\ln(1.25/\delta)}/\varepsilon_{k,r}$, claiming a per-round per-client budget $\varepsilon_{k,r}$. That claim has a calibration precondition: C must upper-bound the sensitivity of every release to which this scale is applied. Heterogeneous, unclipped encoders do not enforce that precondition. A client with sensitivity above C therefore lacks the advertised sufficient certificate, while a client below C receives more noise than its own ceiling requires.

The analytical object. For a valid sensitivity upper bound U, scale σ, and failure probability δ, let $\epsilon_G(U, \sigma, \delta)$ denote a valid Gaussian-mechanism upper certificate. Basic composition gives the

Preprint.

released transcript the certificate

$$\varepsilon_{\mathrm{cert}} = \sum_{r=1}^{R} \max_{k \in [K]} \epsilon_G(U_k^{(r)}, \sigma_{k,r}, \delta), \qquad \delta_{\mathrm{total}} = R\delta. \tag{1}$$

This is a standard sufficient certificate, not a new accountant and not the minimal DP parameter of the mechanism. Our contribution is to expose and repair the missing sensitivity input for this split-FL mechanism.

Why round indexing matters here. Within a round, disjoint client mechanisms admit parallel composition; across adaptive rounds, their certificates add. The composition rule is elementary. The mechanism-specific point is that the valid ceiling and the client binding the maximum can change with the encoder state, so checking only the initial state does not certify the full transcript. Under the classical sufficient Gaussian rule used to set the scales, a release with bound $U_k^{(r)}$ is recalculated as $\epsilon_0 U_k^{(r)} / C_{t(k),r}$ within the rule's stated regime; outside it we use analytic Gaussian calibration [Balle and Wang, 2018].

Summary of contributions.

(1) We formalize fixed-state batch-mean representation sensitivity under replace-one-client adjacency, for which one, several, or all records of one client may change and no record-level $1/b$ fallback is available.

(2) We give the round-indexed certificate recalculation in Proposition 3.3 and state precisely its mechanism and trust boundaries. Corollary 3.6 applies only to the server view that already contains the released transcript.

(3) We introduce SABA as a continuum between one global profile, $T \ll K$ tier profiles, and K per-client profiles. With valid per-round ceilings, Certified-SABA restores the nominal certificate; empirical variants remain diagnostics rather than certificates.

(4) We report an independent one-run membership-inference witness and a five-seed certified MLP sweep. Neither the attack nor the certified columns use the empirical sensitivity estimator that defines the empirical tiers.

Related work. *Heterogeneous FL privacy.* Early heterogeneous-DP formulations Aldaghri et al. [2021], FedHDP Gao et al. [2024], and GDPFed/+ Xu et al. [2026] address heterogeneous privacy preferences by forming privacy groups under standard gradient clipping. Our work is orthogonal: we uncover sensitivity heterogeneity in unclipped split-FL even when a uniform privacy target is intended. UDP-FL Feng et al. [2024] harmonizes different DP mechanisms assuming correct per-round inputs; our results concern how a valid representation-sensitivity input is obtained for the studied mechanism. L-RDP Behnia et al. [2025] provides rigorous per-client accounting for LDP; SABA can supply global, tiered, or per-client calibration profiles for split-FL activations. A fair adaptation of these methods must specify the same valid representation-sensitivity input; we do not claim empirical dominance over them (Appendix 12).

2 Formal Model

We write $[K] = \{1, \ldots, K\}$, $[R] = \{1, \ldots, R\}$, and $(x)_+ = \max(x, 0)$. All logarithms are natural.

Assumptions and roadmap.

A1. Adjacency. We use replace-one-client adjacency over disjoint, fixed-size client datasets (Definition 2.2).

A2. Trust boundary. The honest-but-curious server observes each client-side-noised representation. Withholding releases, secure aggregation, or trusted execution defines a different observed mechanism.

A3. Released mechanism. Each release is an unclipped batch-mean representation from the fixed encoder state for that release plus isotropic Gaussian noise. Certifying a private update immediately followed by release requires a bound for the complete update-and-release map.

A4. Evidence. Empirical SABA takes a maximum over $10{,}000$ public batch pairs, disjoint within each pair; this is a lower approximation to the supremum, not a certificate. Direct and RDP use that empirical input, whereas MI-LB is an independent one-sided attack witness. Empirical and Dynamic Empirical SABA are not certified. Certified-SABA uses valid MLP spectral ceilings recomputed from the fixed state of each release.

A5. Composition. Equation (7) is a standard upper certificate, not the mechanism's minimal privacy parameter.

A6. Certified profiles. Tier membership is architecture-fixed and public, and certified tier t uses $C_{t,r} = \max_{k \in \mathcal{T}_t} U_k^{(r)}$. These ceilings are protocol inputs; communicating additional raw, data-dependent spectral statistics requires separate privacy analysis.

Lemma 3.1 states the local calibration precondition, Proposition 3.3 recalculates the transcript certificate, Corollary 3.6 describes the observed server view, and Proposition 3.7 gives the certified tier and utility properties.

Definition 2.1 (Differential Privacy Dwork et al. [2006])**.** $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-DP if for all adjacent $D \sim D'$ and measurable $S \subseteq \mathcal{Y}$, $\mathbb{P}[\mathcal{M}(D) \in S] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(D') \in S] + \delta$.

Definition 2.2 (Replace-One-Client Adjacency)**.** Let client k hold a fixed-size dataset $\mathcal{D}_k \in \mathcal{Z}^{n_k}$ and let $D = (\mathcal{D}_1, \ldots, \mathcal{D}_K)$. Datasets D and $D' = (\mathcal{D}_1, \ldots, \mathcal{D}'_{k^*}, \ldots, \mathcal{D}_K)$ are adjacent, written $D \sim_c D'$, when exactly one client coordinate is replaced. The replacement $\mathcal{D}'_{k^*} \in \mathcal{Z}^{n_{k^*}}$ may differ from $\mathcal{D}_{k^*}$ in one, several, or all records; it is not required to differ in every record.

Definition 2.3 (Fixed-State Batch-Mean Sensitivity)**.** Fix the encoder state $w_k^{(r)}$ used for release (k, r). For equal-size batches $B \subset \mathcal{D}_k$ and $B' \subset \mathcal{D}'_k$ with $|B| = |B'| = b$, let $\bar{h}_{k,r}(B) = b^{-1} \sum_{x \in B} f_k(x; w_k^{(r)})$. The sensitivity is

$$\Delta_k^{(r)} = \sup_{\mathcal{D}_k, \mathcal{D}'_k} \sup_{\substack{B \subset \mathcal{D}_k,\, B' \subset \mathcal{D}'_k \\ |B| = |B'| = b}} \|\bar{h}_{k,r}(B) - \bar{h}_{k,r}(B')\|_2, \tag{2}$$

where $\mathcal{D}_k$ and $\mathcal{D}'_k$ occupy the replaced client coordinate in an adjacent global pair.

Remark 2.4 (Empirical and certified inputs). Empirical SABA approximates (2) using $10{,}000$ disjoint public batch pairs (Appendix 9.1); this sampled maximum is a lower approximation, not a certificate. For an MLP with ReLU activations, we instead use the fixed-state upper bound

$$U_k^{(r)} = L_k^{(r)} \operatorname{diam}(\mathcal{X}), \qquad L_k^{(r)} = \prod_{\ell} \|W_{k,\ell}^{(r)}\|_2, \tag{3}$$

for Certified-SABA [Bartlett et al., 2017]. Since ReLU is 1-Lipschitz, $U_k^{(r)} \geq \Delta_k^{(r)}$ for the fixed encoder state. We do not apply this MLP certificate to ViTs.

Remark 2.5 (No record-level $1/b$ fallback). Under record-level adjacency, at most one selected record changes and a $1/b$ factor can arise. Under replace-one-client adjacency, the worst case permits all b selected records to change. If the fixed encoder is $L_k^{(r)}$-Lipschitz,

$$\Delta_k^{(r)} \leq \sup_{B, B'} \frac{1}{b} \sum_{i=1}^{b} L_k^{(r)} \|x_i - x'_i\|_2 \leq L_k^{(r)} \operatorname{diam}(\mathcal{X}). \tag{4}$$

The average therefore provides no record-level $1/b$ reduction for the client-level claim.

Definition 2.6 (Round-Indexed Release and Server Update)**.** For fixed encoder state $w_k^{(r)}$, client k releases

$$\tilde{h}_{k,r} = \bar{h}_{k,r}(B_{k,r}) + \xi_{k,r}, \qquad \xi_{k,r} \sim \mathcal{N}(0, \sigma_{k,r}^2 \mathbf{I}_{d_k}). \tag{5}$$

The round release is the tuple $T_r = (\tilde{h}_{1,r}, \ldots, \tilde{h}_{K,r})$, and the released transcript is $T = (T_1, \ldots, T_R)$. The update is any measurable map

$$w^{(r)} = \mathrm{Agg}((\tilde{h}_{1,r}, \ldots, \tilde{h}_{K,r}), w^{(r-1)}). \tag{6}$$

No particular aggregation rule is assumed. If private local optimization changes $w_k^{(r)}$ immediately before transmission, Equation (3) certifies only the fixed-state input-to-representation map; the complete update-and-release map needs an additional sensitivity bound.

Definition 2.7 (Gaussian Upper Certificate)**.** For a valid ℓ_2-sensitivity upper bound U, noise scale σ, and $\delta \in (0,1)$, $\epsilon_G(U,\sigma,\delta)$ denotes a valid upper certificate for the Gaussian mechanism. We use analytic Gaussian calibration [Balle and Wang, 2018]. When the classical sufficient rule $\sigma = C\sqrt{2\ln(1.25/\delta)}/\epsilon_0$ is used in its stated regime, substitution gives the sufficient recalculation $\epsilon_0 U/C$. Neither quantity is asserted to be the mechanism's minimal DP parameter.

Proposition 2.8 (Basic Transcript Certificate)**.** *Suppose $U_k^{(r)} \geq \Delta_k^{(r)}$ is valid for each fixed-state release. Under replace-one-client adjacency, parallel composition within each round and adaptive sequential composition across rounds give*

$$\varepsilon_{\mathrm{cert}} = \sum_{r=1}^{R} \max_{k\in[K]} \epsilon_G(U_k^{(r)}, \sigma_{k,r}, \delta), \qquad \delta_{\mathrm{total}} = R\delta. \tag{7}$$

Thus the transcript is $(\varepsilon_{\mathrm{cert}}, R\delta)$-DP. In our experiments, $R = 50$ and $\delta = 10^{-5}$, hence $\delta_{\mathrm{total}} = 5 \times 10^{-4}$.

3 The Sensitivity Gap: Main Results

3.1 Local Calibration and Transcript Recalculation

Lemma 3.1 (Local Calibration Precondition)**.** *Let $U_k^{(r)} \geq \Delta_k^{(r)}$ be a valid sensitivity ceiling for release (k,r). With Gaussian scale $\sigma_{k,r}$, that release has the upper certificate $\epsilon_G(U_k^{(r)}, \sigma_{k,r}, \delta)$. Consequently, a shared surrogate C supports the claimed certificate for the studied unclipped mechanism only when $C \geq U_k^{(r)}$ for every release to which the shared scale is applied.*

Proof. The first statement is the Gaussian mechanism applied with a valid sensitivity upper bound. The second follows by setting $U = C$: without a proof that C upper-bounds the release sensitivity, the sufficient calibration premise is absent. This statement does not characterize the minimal DP parameter. □

Lemma 3.2 (Parallel Then Sequential Certificate)**.** *For replace-one-client adjacency, the K fixed-state releases in round r operate on disjoint client coordinates, so their certificates compose in parallel as*

$$\epsilon_{\mathrm{cert}}^{(r)} = \max_{k\in[K]} \epsilon_G(U_k^{(r)}, \sigma_{k,r}, \delta). \tag{8}$$

Because later encoder states and releases may depend on earlier server updates, basic adaptive composition sums these round certificates and their failure probabilities.

Proof. For an adjacent pair, only one client coordinate is replaced within a round, which gives the maximum. Standard adaptive sequential composition then applies across the R round mechanisms. □

3.2 Round-Indexed Certificate Recalculation

Proposition 3.3 (Round-Indexed Certificate Recalculation)**.** *Let $U_k^{(r)}$ be valid upper bounds for all fixed-state releases and let $\sigma_{k,r}$ be their actual Gaussian scales. Then the transcript has the certificate in* (7)*. If the scale for client k is set by the classical sufficient rule using profile $C_{t(k),r}$ and target ϵ_0, the corresponding sufficient-rule recalculation is*

$$\epsilon_{\mathrm{rule},k,r} = \epsilon_0 \frac{U_k^{(r)}}{C_{t(k),r}}, \qquad \epsilon_{\mathrm{rule},T} = \sum_{r=1}^{R} \max_k \epsilon_{\mathrm{rule},k,r}, \tag{9}$$

within the classical rule's stated parameter regime. Analytic Gaussian calibration is used otherwise.

Proof. Apply Lemma 3.2 and substitute the scale $\sigma_{k,r} = C_{t(k),r}\sqrt{2\ln(1.25/\delta)}/\epsilon_0$ for the classical-rule display. The first equality is a standard composition upper bound. It is neither an equality with nor a lower bound on the mechanism's minimal DP parameter. □

Algorithm 1: SABA: Global, Tiered, or Per-Client Calibration

Input: Clients $[K]$; target ϵ_0, δ; public prespecified profile assignment $t(k) \in [T]$; mode *Empirical* or *Certified*.

Output: Scales $\sigma_{k,r}$ and evidence label.

foreach *round* r **do**
  if *Certified* **then**
    Compute valid fixed-state MLP ceilings $U_k^{(r)} = L_k^{(r)} \operatorname{diam}(\mathcal{X})$
    $q_{k,r} \leftarrow U_k^{(r)}$
  else
    Reuse $q_{k,r} \leftarrow \hat{\Delta}_k$ from 10,000 disjoint public batch pairs, refreshing only on the prespecified dynamic schedule
  foreach *profile* $t \in [T]$ **do**
    $C_{t,r} \leftarrow \max_{k:t(k)=t} q_{k,r}$
    $\sigma_{t,r} \leftarrow C_{t,r}\sqrt{2\ln(1.25/\delta)}/\epsilon_0$
  foreach *client* k **do**
    Transmit $\tilde{h}_{k,r} = \bar{h}_{k,r}(B_{k,r}) + \mathcal{N}(0, \sigma^2_{t(k),r}\mathbf{I})$
if *Certified* **then**
  Report the certificate from (7)
else
  Report Empirical or Dynamic Empirical SABA; do not claim certification

Remark 3.4 (Why the round index matters). The composition rule itself is standard. The relevant split-FL issue is that $U_k^{(r)}$ changes with the encoder state and the client attaining the round maximum may also change. A ceiling checked only at round zero does not certify the full transcript. The certified sweep therefore recomputes Equation (3) from each round-r MLP state.

Remark 3.5 (Empirical diagnostics are separate). Replacing $U_k^{(r)}$ in (9) by a sampled maximum $\hat{\Delta}_k^{(r)}$ produces the *Direct* empirical-input diagnostic. An RDP accountant supplied with the same empirical input is likewise empirical. Neither is a certificate. MI-LB, defined in Section 4, uses neither sensitivity estimates nor tier labels and is an independent one-sided attack witness.

3.3 What Static Server Post-Processing Can and Cannot Change

Corollary 3.6 (Observed-Transcript Corollary)**.** *Let $T(D)$ be the client transcript already disclosed to the honest-but-curious server and let A be any measurable static aggregation map. The actual server view is*

$$V(D) = (T(D), A(T(D))). \tag{10}$$

If T fails a target (ϵ, δ)-DP guarantee, then V also fails that target.

Proof. If V satisfied (ϵ, δ)-DP, applying the projection $(t, a) \mapsto t$ and the post-processing property would imply that T satisfied the same guarantee, a contradiction. □

The corollary says only that appending static aggregation cannot erase a failure already present in the observed transcript. It does not apply to $A(T)$ alone when T is withheld, secure aggregation, trusted execution, rejection or retransmission, or interactive re-noising.

3.4 The SABA Calibration Continuum

Proposition 3.7 (SABA Profile and Utility Properties)**.** *Assume valid ceilings $U_k^{(r)} \in [\Delta_{\min}, \Delta_{\max}]$.*

(1) For Certified-SABA, $C_{t,r} = \max_{k:t(k)=t} U_k^{(r)}$ is valid for every client in profile t. Each classical-rule local certificate is therefore at most ϵ_0, and the basic transcript certificate is $(R\epsilon_0, R\delta)$.

(2) With T equal-width profiles, for any client assigned ceiling C_t,

$$\frac{C_t^2 - U_k^2}{U_k^2} \leq \frac{2\Delta_{\max}(\Delta_{\max} - \Delta_{\min})}{T\Delta_{\min}^2} = O(1/T). \tag{11}$$

(3) $T = 1$ is global calibration and $T = K$ is the per-client endpoint. Per-client calibration may be preferable when K certified profiles can be validated and deployed. Tiering quantizes the K assignments into $T \ll K$ noise configurations; it is a profile-count/utility trade-off, not a universal improvement over per-client calibration.

Proof. Part (1) follows from $U_k^{(r)} \leq C_{t,r}$ and Proposition 3.3. For part (2), equal-width profiles give $C_t - U_k \leq (\Delta_{\max} - \Delta_{\min})/T$. Hence $(C_t^2 - U_k^2)/U_k^2 = (C_t - U_k)(C_t + U_k)/U_k^2$, and bounding $C_t + U_k \leq 2\Delta_{\max}$ and $U_k \geq \Delta_{\min}$ yields the display. Part (3) follows from the number of distinct scale configurations. □

4 Experiments

We use CIFAR-10 and CIFAR-100 as controlled proof-of-concept validation, not as production validation. Data are partitioned across $K \in \{10, 50, 100\}$ clients with Dirichlet $\beta = 0.5$; batch size is 32. Client encoders are heterogeneous MLPs with depths $\{2, 4, 6, 8\}$ or ViT-Tiny/ViT-Small. Training uses $R = 50$ releases, $\epsilon_0 = 0.1$, $\delta = 10^{-5}$ per release, and five seeds. Empirical sensitivity uses 10,000 disjoint public batch pairs. Certified ceilings are recomputed from round-r MLP states and are never applied to ViTs. The displayed W_1 is measured after training and is not an input optimized or supplied to SABA. Appendix 10 gives the remaining training and compute details.

4.1 Operational Inputs and Profile Construction

At a fixed evaluation state, the empirical input for client k is

$$\widehat{\Delta}_k^{(r)} = \max_{j \in [10,000]} \left\| \bar{h}_{k,r}(B_j) - \bar{h}_{k,r}(B_j') \right\|_2, \qquad B_j \cap B_j' = \varnothing, \quad |B_j| = |B_j'| = 32. \tag{12}$$

The pairs use public data and are disjoint within a pair; pairs may reuse public records. Equation (12) is a sampled maximum and therefore a lower approximation to (2), not a valid upper bound. Static Empirical SABA caches the public setup value, while Dynamic Empirical SABA recomputes it on its prespecified ten-round schedule.

The empirical implementation uses $T = \lceil \log_2 K \rceil$ log-spaced bins over the positive public reference values. A client's assignment is fixed before the private releases covered by the reported transcript, and a nonempty profile uses its largest reference value as its scale input. For Certified-SABA, membership is instead a public design input associated with the declared encoder configuration, and the actual round-r scale uses the valid maximum $C_{t,r} = \max_{k:t(k)=t} U_k^{(r)}$. Hence grouping itself does not create validity: a tier maximum helps only when every member input U_k is already a valid bound.

To make the horizontal axis in Figure 1 operational, let $P_\Delta = K^{-1} \sum_k \delta_{\widehat{\Delta}_k}$ and let δ_C be a point mass at the shared surrogate. Then

$$W_1(P_\Delta, \delta_C) = \frac{1}{K} \sum_{k=1}^{K} |\widehat{\Delta}_k - C|. \tag{13}$$

Architecture allocation, optimization, and the resulting encoder states induce this value. We index conditions by the measured post-training value; we neither optimize it nor use it to choose a SABA profile.

Methods and evidence labels. *Naive* uses one shared C. *Empirical SABA* uses a fixed sampled public-data estimate; *Dynamic Empirical SABA* refreshes that estimate on a prespecified schedule; neither is certified. *Certified-SABA* uses Equation (3). *Oracle* and empirical per-client calibration use K sampled profiles and are utility references. *Rep-Clip* clips each representation to radius C; because its output diameter is at most $2C$, it is an empirical utility/attack baseline rather than a certified sensitivity-C baseline. Direct is the sampled-input version of (9); RDP is also empirical unless supplied valid sensitivity bounds. MI-LB is the independent witness below.

One-run MI-LB protocol. For each seed, $m = 5{,}000$ audit slots are assigned in advance to fixed clients and scheduled batches. Independent $S_i \in \{-1, +1\}$ select a canary or fixed non-canary. Flipping one slot changes one record in one client while all other data remain fixed, a restricted pair contained in replace-one-client adjacency. The audit uses only the scheduled $R = 50$ transcript: $\tilde{h}_i$ is the saved noisy batch-mean release containing the fixed slot and other $b - 1$ records. The final server head scores that saved release; no client is requeried and no new representation is generated. The tested null is

$$(\epsilon_{\text{naive}}, \delta_{\text{total}}) = (5, 5 \times 10^{-4}).$$

The prespecified score is $s_i = -L_{\text{CE}}(g_\phi(\tilde{h}_i), y_i) = \log p_\phi(y_i \mid \tilde{h}_i)$. The 1,000 largest scores are guessed *in*, the 1,000 smallest *out*, and 3,000 abstain, identically for every method. With $T_i \in \{-1, 0, +1\}$, $r_a = 2{,}000$, $W = \sum_i \max\{0, T_i S_i\}$, $p_\epsilon = e^\epsilon/(1 + e^\epsilon)$, and $Z \sim \text{Binomial}(r_a, p_\epsilon)$, we implement Steinke et al. [2023]:

$$p\text{-value}(\epsilon) = \Pr[Z \geq W] + \delta_{\text{total}} m\, \alpha(\epsilon, W, r_a), \quad \alpha = \max_{j \in [m]} 2^j \Pr[W > Z \geq W - j]. \tag{14}$$

We invert at significance 0.05. The largest rejected null is ϵ_{LB} and the reported quantity is MI-LB gap $= [\epsilon_{\text{LB}} - \epsilon_{\text{naive}}]^+$, mean $\pm$ SE over five seeds. A positive gap rejects the common nominal claim under this audit; a near-zero gap does not certify privacy. SE is 0.02–0.07, compared with a Naive-to-SABA difference of about 1.06 at $W_1 = .30$.

Gap computation and uncertainty. For a profile ceiling $C_{t(k),r}$, Direct replaces the unknown worst-case sensitivity by the sampled value in the same sufficient-rule calculation:

$$\epsilon_{\text{Direct}} = \sum_{r=1}^{R} \max_k \epsilon_0 \frac{\widehat{\Delta}_k^{(r)}}{C_{t(k),r}}, \qquad \text{Direct gap} = [\epsilon_{\text{Direct}} - R\epsilon_0]^+. \tag{15}$$

The RDP gap is $[\epsilon_{\text{RDP}} - R\epsilon_0]^+$ after giving the Opacus accountant the same sampled sensitivity input. MI-LB instead subtracts the same nominal target from a rejected-null lower witness. Thus the three gaps share a numerical origin at $R\epsilon_0 = 5$ but not an evidentiary meaning. All table entries are means over five independent seeds with $\text{SE} = s/\sqrt{5}$, where s is the sample standard deviation. The $\max \widehat{\Delta} - C$ column is the largest sampled overshoot of shared C in that condition.

Table 1: Empirical evidence on CIFAR-10 ($K = 50$, $R = 50$). Direct and RDP use sampled sensitivity inputs; MI-LB is an independent one-sided witness. Values are mean $\pm$ SE over five seeds.

Method	W_1	$\max \hat{\Delta} - C$	Direct gap	MI-LB gap	RDP gap	Acc. (%)	Δ_{acc}
Naive	.30	.62	2.48±.05	1.27±.04	1.31±.03	80.9±.2	—
Naive	.60	1.24	4.96±.09	2.53±.07	2.61±.05	80.1±.3	—
Naive	.90	1.86	7.44±.14	3.79±.10	3.88±.08	79.7±.3	—
ORACLE†	.30	—	0.00	0.02±.01	0.01±.01	80.8±.2	−0.1
ORACLE†	.60	—	0.00	0.03±.01	0.02±.01	80.0±.2	−0.1
ORACLE†	.90	—	0.00	0.04±.01	0.03±.01	79.6±.3	−0.1
Empirical SABA	.30	.62	0.42±.04	0.21±.02	0.22±.02	80.6±.2	−0.3
Empirical SABA	.60	1.24	0.83±.06	0.43±.03	0.44±.03	80.0±.2	−0.1
Empirical SABA	.90	1.86	1.25±.08	0.65±.04	0.66±.04	79.4±.3	−0.3
Rep-Clip‡	.30	—	0.00	0.01±.01	0.00±.01	76.6±.3	−4.3

†Empirical per-client utility reference, not a certificate. ‡Empirical clipping baseline; radius C has diameter at most $2C$.

Independent empirical agreement. At $W_1 = .30$, Naive→Empirical SABA changes from $2.48 \pm .05 \to 0.42 \pm .04$ under Direct, $1.31 \pm .03 \to 0.22 \pm .02$ under empirical-input RDP, and $1.27 \pm .04 \to 0.21 \pm .02$ under MI-LB, with a 0.3-point accuracy change. Across five CIFAR-10/100 conditions, Direct decreases by 83.1%–83.3% and MI-LB by 82.8%–84.3%, with 0.1–0.6 accuracy-point changes. Agreement is not circular because MI-LB takes neither the sensitivity estimates nor tier labels as input.

Figure 1 retains the original heterogeneity and shift visualizations. Both panels show empirical diagnostics: the dashed line is a linear reference rather than a privacy theorem, and W_1 is measured after training rather than supplied to SABA.

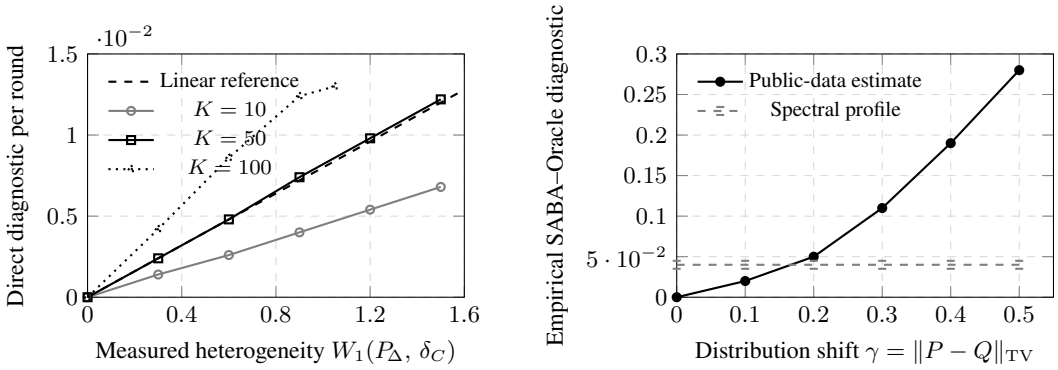

Figure 1: Original empirical diagnostic plots, retained with corrected interpretation. *Left:* Direct diagnostic per round versus measured post-training heterogeneity. *Right:* Empirical SABA–Oracle diagnostic versus public/private shift. Points are means over five seeds; neither panel is a DP certificate.

Table 2: Completed certified MLP sweep. For each round, the largest local client certificate is selected and the displayed transcript certificate sums those R round maxima; the largest local value also determines whether the classical or analytic Gaussian calculation is used. Bounds are recomputed from the round-r states, and neither certificate column uses Monte Carlo sensitivity estimates. Accuracy is for Certified-SABA.

Setting	Shared-C cert.	Certified-SABA	Acc. (%)
Original Table 4 configuration	7.44	5.00	78.2±.4
C10, $K = 50$, $W_1 = .30/.60/.90$	7.48/9.96/12.44	5.00/5.00/5.00	78.2±.4/76.9±.4/75.8±.5
C10, $W_1 = .60$, $K = 10/100$	7.78/13.06	5.00/5.00	77.4±.4/76.3±.5
C100, $K = 50$, $W_1 = .30/.60$	7.48/9.96	5.00/5.00	57.8±.5/56.1±.6

These are upper certificates, not estimates of the mechanism's minimal DP parameter. The table uses the classical-rule recalculation for releases in its stated regime and analytic Gaussian calibration otherwise. Certified tier ceilings are public protocol inputs; raw data-dependent spectral statistics are not included in the certified transcript.

Certified sweep audit. For each fixed-state MLP release we form

$$e_{k,r} = \epsilon_G(U_k^{(r)}, \sigma_{k,r}, \delta), \qquad e_r = \max_k e_{k,r}, \qquad e_T = \sum_{r=1}^{R} e_r. \tag{16}$$

The largest local value $\max_r e_r$ determines whether the classical-rule display is in its stated regime; Equation (20) is used otherwise. The shared-C column reports e_T under the shared scales. Certified-SABA checks $U_k^{(r)} \le C_{t(k),r}$ before each release, which gives $e_r \le \epsilon_0$ and the nominal transcript ceiling 5.00. This is a certificate for the fixed-state representation mechanism. It does not certify the private procedure that may have trained or selected the state, and it does not include communication of raw spectral norms.

The repeated numbers in the submission are intentionally separated. In Table 1, 7.44 is the $W_1 = .90$ Direct *gap*, so the corresponding empirical-input total is 12.44. In Table 2, the $W_1 = .30$ shared-C total is 7.48, 12.44 is the $W_1 = .90$ total, and 7.44 remains only for the distinct original Table 4 configuration. Appendix 11.1 records the reporting identities and this reconciliation in full.

Table 3: Measured profile and runtime trade-off at $K = 100$. Training takes approximately 12 GPU hours. Accuracy is for the representative CIFAR-10 setting.

Method	One-time setup	Per round	Profiles	Status / Acc. (%)
Naive	none	none	1	valid only if shared C is valid / $80.9\pm.2$
Rep-Clip	none	clipping pass	1	empirical baseline / $76.6\pm.3$
Empirical SABA	80 GPU-s + $<$ 2 ms sort	scalar; negligible	$T = 7$	not certified / $80.6\pm.2$
Empirical per-client	80 GPU-s	negligible	$K = 100$	sampled: not certified / $80.8\pm.2$

For ViT-Tiny, Phase 1 costs 0.8 GPU-s per client, or 80 GPU-s at $K = 100$; fixed-profile overhead is negligible. Ten-round empirical refresh adds below 5% and changes the drift diagnostic from 0.05 to 0.01. That refresh remains empirical.

5 Certificate Geometry and Profile Trade-Off

Explicit recalculation. For scales set by shared C under the classical sufficient rule, valid ceilings yield

$$\epsilon_{\mathrm{rule},T} = \sum_{r=1}^{R} \max_k \epsilon_0 \frac{U_k^{(r)}}{C}. \tag{17}$$

The nominal $R\epsilon_0$ certificate follows only if C is a valid ceiling for every release. Equation (17) is a sufficient-rule recalculation, not an exact privacy cost. For empirical rows, replacing U by $\hat{\Delta}$ defines Direct and cannot establish a worst-case guarantee.

Batch size and certified ceilings. Larger batches can reduce the sampled Direct diagnostic because averaging suppresses typical variation. Under replace-one-client adjacency, however, all b selected elements may change, so the valid worst-case MLP ceiling in (3) has no record-level $1/b$ factor.

Tiers versus per-client calibration. The representative accuracy ordering is Oracle/Empirical SABA/tuned per-client clipping/global clipping $= 80.8/80.6/79.1/76.6\%$. This supports tiering as a useful empirical profile-count/utility compromise, not as universally superior to per-client calibration. The tier-count ablation in Appendix 11.5 is consistent with the explicit $O(1/T)$ bound in (11); no $O(1/T^2)$ rate is claimed.

Choosing the profile resolution. The number of profiles is a public systems parameter, not an extra privacy budget. $T = 1$ minimizes configuration and validation cost but inherits the largest valid global ceiling. $T = K$ minimizes within-profile over-noising but requires validating, distributing, and maintaining K scale configurations. An intermediate T is useful when a deployment supports only a small catalog of noise profiles. The appropriate choice is therefore the largest profile count whose bounds and assignments can be validated under the stated trust model, rather than the value that happens to minimize an empirical gap. The profile-count curve in Appendix 11.5 reports utility only; it is not used to select or validate a privacy certificate.

Certificate contract. The certified result takes four inputs: replace-one-client adjacency, the client-side-observed release tuple, public fixed profile membership, and a valid $U_k^{(r)}$ for the fixed state used by each release. Its output is the upper transcript certificate in (7). Changing the adjacency relation, hiding individual releases, selecting profiles from private statistics, or including a private optimizer in the query changes the mechanism and invalidates direct reuse of that certificate. This contract is why empirical and certified SABA can share an implementation structure while retaining different evidence labels.

Residual empirical gap. The $0.42\pm.04$ Direct gap for Empirical SABA reflects the sampled maximum, state drift, and public/private coverage. Dynamic refresh reduces the measured drift diagnostic but remains empirical. Only the MLP rows using valid per-round spectral ceilings support the certificate in Table 2.

6 Discussion

What is and is not new. Parallel-within-round and sequential-across-round composition, Gaussian rescaling, and RDP accounting are standard. The contribution is the mechanism-specific calibration precondition: an unclipped heterogeneous split-FL implementation does not enforce its shared surrogate as a valid sensitivity ceiling, and that ceiling can change by round.

Server scope. Corollary 3.6 concerns the actual view $(T, A(T))$ after T has already been disclosed. It makes no claim about an aggregate-only release, secure aggregation, enclaves, rejection or retransmission, or interactive re-noising.

Comparison scope. FedHDP Gao et al. [2024], GDPFed Xu et al. [2026], UDP-FL Feng et al. [2024], and L-RDP Behnia et al. [2025] address different privacy preferences, mechanisms, or accounting layers. Adapting them fairly requires specifying a valid representation-sensitivity input for the same unclipped release. Our experiments do not establish dominance over those methods.

Operational audit workflow. For a new split-FL deployment, the calibration check is: (i) write the exact server-visible transcript and adjacency relation; (ii) bound the complete query producing each disclosed representation; (iii) declare every global, tiered, or per-client ceiling and assignment that is a public protocol input; (iv) calibrate the actual noise scales and compose the resulting local upper certificates; and (v) report sampled maxima and attack tests separately from the certificate. A failure at step (ii) cannot be repaired by choosing a more advanced accountant at step (iv), because the accountant still requires valid local inputs. Conversely, a valid bound may be used with a tighter accountant than basic composition; our displayed sums use the elementary rule so the calibration dependence remains transparent.

Limitations. The certificate covers only fixed-state MLP input-to-representation maps with valid public protocol ceilings. If private local optimization changes the encoder immediately before release, the complete update-and-release map requires another bound. Empirical and Dynamic Empirical SABA, ViT experiments, Direct, and empirical-input RDP are diagnostic. MI-LB is one-sided: failure to reject does not certify privacy. Communicating data-dependent spectral statistics would enlarge the transcript and needs separate analysis. The results also do not automatically transfer to DP-SGD, the shuffle model, other mechanisms, or production deployments.

External validity. The evaluated heterogeneity is controlled through CIFAR-10/100 partitions and MLP/ViT encoders, not a production hierarchical split-FL stack. The empirical ordering may change with other modalities, optimizers, or public reference distributions. The certificate is less dataset-specific but remains tied to the declared adjacency, fixed-state query, and input-domain diameter; an end-to-end train-and-release claim must also bound the private optimizer and any state-dependent selection or communication.

Audit power and reporting. Five seeds quantify run-to-run variability but do not exhaust architecture or data heterogeneity. MI-LB tests $5{,}000$ prespecified slots with one score and abstention rule; another attack may have different power. We therefore keep the attack, sampled Direct/RDP diagnostics, and spectral certificate in separate columns rather than treating agreement as equality. The appendix retains their full constructions, accounting, ablations, and runtime.

7 Conclusion

A shared Gaussian scale certifies the studied unclipped heterogeneous split-FL mechanism only when its calibration constant is a valid ceiling for every observed release. We give the corresponding round-indexed upper certificate, delimit what static processing of an already observed transcript can change, and introduce SABA's global/tiered/per-client profile continuum. Independent attack evidence supports the empirical ordering, while a completed MLP sweep using valid per-round spectral ceilings restores the nominal certificate. These claims concern calibration of this release mechanism, not new composition theory or the minimal DP parameter.

References

N. Aldaghri, H. Mahdavifar, and A. Beirami. Federated learning with heterogeneous differential privacy. *arXiv preprint arXiv:2110.15252*, 2021.

B. Balle and Y.-X. Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR, 2018.

P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

R. Behnia, J. Birrell, A. Riasi, R. Ebrahimi, K. Dutta, and T. Hoang. Local differential privacy for federated learning with fixed memory usage and per-client privacy. *arXiv preprint arXiv:2510.12908*, 2025.

R. Deng, X. Du, Z. Lu, Q. Duan, S.-C. Huang, and J. Wu. Hsfl: Efficient and privacy-preserving offloading for split and federated learning in iot services. In *2023 IEEE International conference on web services (ICWS)*, pages 658–668. IEEE, 2023.

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

S. Feng, M. Mohammady, H. Hong, S. Yan, A. Kundu, B. Wang, and Y. Hong. Harmonizing differential privacy mechanisms for federated learning: Boosting accuracy and convergence. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy*, pages 60–71, 2024.

C. Gao, A. Lowy, X. Zhou, and S. Wright. Private heterogeneous federated learning without a trusted server revisited: Error-optimal and communication-efficient algorithms for convex losses. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=sSAEhcdB9N`.

Z. Lin, W. Wei, Z. Chen, C.-T. Lam, X. Chen, Y. Gao, and J. Luo. Hierarchical split federated learning: Convergence analysis and system optimization. *IEEE Transactions on Mobile Computing*, 2025.

I. Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

F. J. Piran, Z. Chen, Y. Zhang, Q. Zhou, J. Tang, and F. Imani. Privacy-preserving decentralized federated learning via explainable adaptive differential privacy. *arXiv preprint arXiv:2509.10691*, 2025.

T. Steinke, M. Nasr, and M. Jagielski. Privacy auditing with one (1) training run. In *Advances in Neural Information Processing Systems*, volume 36, pages 49268–49280, 2023.

J. Xu, R. Hu, and O. Kotevska. Optimal client sampling in federated learning with client-level heterogeneous differential privacy. *IEEE Internet of Things Journal*, 2026.

8 Certificate Derivations and Mechanism Boundaries

8.1 Gaussian sufficient rules

For a query with ℓ_2-sensitivity at most U, the classical Gaussian mechanism result gives the sufficient condition

$$\sigma \geq \frac{U\sqrt{2\ln(1.25/\delta)}}{\epsilon}, \qquad \epsilon \in (0,1). \tag{18}$$

This condition is sufficient, not necessary. Our implementation uses analytic Gaussian calibration [Balle and Wang, 2018]; the classical expression is retained only to make transparent the recalculation for scales that were set using that rule.

For completeness, write $\mu = U/\sigma$ and let Φ denote the standard normal cumulative distribution function. The exact Gaussian privacy profile used by the analytic calculation is

$$\delta_G(\epsilon; U, \sigma) = \Phi\left(\frac{\mu}{2} - \frac{\epsilon}{\mu}\right) - e^{\epsilon}\Phi\left(-\frac{\mu}{2} - \frac{\epsilon}{\mu}\right). \tag{19}$$

We define

$$\epsilon_G(U, \sigma, \delta) = \inf\{\epsilon \geq 0 : \delta_G(\epsilon; U, \sigma) \leq \delta\}, \tag{20}$$

and solve this monotone one-dimensional inversion numerically. For every release, the implementation first checks the local value implied by the classical rule. Equation (9) is used only in its stated $\epsilon \in (0,1)$ regime; otherwise Equation (20) supplies the local upper certificate before round-wise maximization and transcript composition.

8.2 Proof details for Proposition 3.3

Fix a round r and an adjacent pair differing in client k^*. All other client inputs agree, so parallel composition bounds the tuple $T_r = (\tilde{h}_{1,r}, \ldots, \tilde{h}_{K,r})$ by the largest client certificate:

$$\epsilon_{\text{cert}}^{(r)} = \max_k \epsilon_G(U_k^{(r)}, \sigma_{k,r}, \delta).$$

The server update and future encoders may depend on earlier release tuples, so basic adaptive composition sums this quantity over rounds and sums one δ per round. This proves (7). If the actual scale is $\sigma_{k,r} = C_{t(k),r}\sqrt{2\ln(1.25/\delta)}/\epsilon_0$, substitution in the same sufficient rule yields $\epsilon_0 U_k^{(r)}/C_{t(k),r}$. No lower-bound direction is used.

8.3 Fixed-state MLP spectral ceiling

For an MLP $f_k = W_{k,L}\rho(W_{k,L-1}\rho(\cdots\rho(W_{k,1}x)))$ with ReLU ρ, submultiplicativity and 1-Lipschitzness give

$$\|f_k(x) - f_k(x')\|_2 \leq \left(\prod_{\ell=1}^{L}\|W_{k,\ell}^{(r)}\|_2\right)\|x - x'\|_2.$$

Applying this to every paired summand in two batches and using replace-one-client adjacency gives $U_k^{(r)} = L_k^{(r)}\operatorname{diam}(\mathcal{X})$ as in Equation (3). The state $w_k^{(r)}$ is held fixed in this argument. Bias terms, when present, cancel in the difference. If the public preprocessing contract bounds coordinate j to $[a_j, b_j]$, the diameter used here is the valid deterministic bound

$$\operatorname{diam}(\mathcal{X}) \leq \left(\sum_j (b_j - a_j)^2\right)^{1/2}. \tag{21}$$

For two batches, the triangle inequality gives

$$\left\|\frac{1}{b}\sum_{i=1}^{b} f_k(x_i) - \frac{1}{b}\sum_{i=1}^{b} f_k(x_i')\right\|_2 \leq \frac{L_k^{(r)}}{b}\sum_{i=1}^{b}\|x_i - x_i'\|_2 \tag{22}$$

$$\leq L_k^{(r)}\operatorname{diam}(\mathcal{X}), \tag{23}$$

which makes explicit why client replacement cancels the apparent $1/b$ factor. A private optimizer that changes the state immediately before release is part of a different query and needs a bound for the complete map.

8.4 Observed transcript

The corollary does not follow by reversing a data-processing inequality. Instead, it uses projection. If the joint server view $V(D) = (T(D), A(T(D)))$ were (ϵ, δ)-DP, its first-coordinate projection would make $T(D)$ (ϵ, δ)-DP by post-processing. Therefore a target already failed by the observed T is also failed by V. This says nothing about $A(T)$ alone when T is hidden.

8.5 Trust boundary and protocol inputs

Gaussian noise is sampled client-side before transmission, since the honest-but-curious server is the adversary that observes the release. Secure aggregation or a trusted enclave that withholds individual representations changes this observation model. In the certified experiments, architecture-fixed tier assignments and tier ceilings are treated as public protocol inputs. Raw data-dependent spectral statistics are not added to the certified transcript; any protocol that communicates them through another channel must analyze that channel separately.

8.6 Failure-probability accounting

Within a round, replace-one-client adjacency changes at most one component of the release tuple, so parallel composition uses one δ. Across adaptive rounds, basic composition yields $\delta_{\mathrm{total}} = \sum_{r=1}^{R} \delta = R\delta$. Thus the experimental setting has $50 \times 10^{-5} = 5 \times 10^{-4}$.

9 SABA Details

9.1 Empirical approximation

At a fixed evaluation state $w_k^{(r)}$, Empirical SABA samples $J = 10{,}000$ public batch pairs of size $b = 32$. The two batches are disjoint within a pair, $B_j \cap B'_j = \varnothing$; different pairs may reuse public records. The estimator is

$$\widehat{\Delta}_k^{(r)} = \max_{j \in [J]} \left\| \bar{h}_{k,r}(B_j) - \bar{h}_{k,r}(B'_j) \right\|_2 . \tag{24}$$

Exhaustive maximization over datasets and batches is combinatorial, so this finite maximum is a lower approximation to the worst-case supremum. Increasing J can improve public-sample coverage but cannot turn the maximum into an upper bound. The estimator is used to compare and group profiles, not to make a certified claim.

9.2 Public profile construction

The profile map $t : [K] \to [T]$ is fixed before the private releases covered by the reported transcript. The empirical implementation uses $T = \lceil \log_2 K \rceil$ log-spaced bins over the positive public reference values $q_k = \widehat{\Delta}_k$ and assigns each client to the bin containing its reference value. The scale for a nonempty profile is set from its largest q_k. Dynamic Empirical SABA changes the public profile ceiling on its prespecified refresh rounds but does not use the displayed post-training W_1 value as an input.

For Certified-SABA, the assignment is instead a public design input tied to the declared encoder configuration. The validity argument does not depend on how clients are grouped: it depends only on taking the valid round-wise maximum $C_{t,r} = \max_{k:t(k)=t} U_k^{(r)}$. The endpoints are $T = 1$ (one global configuration) and $T = K$ (one configuration per client). A profile maximum cannot repair an invalid member bound or a public/private coverage failure.

9.3 Dynamic empirical recalibration

A prespecified dynamic variant repeats the public-batch computation every ten rounds and updates the empirical profile scales. This can track fixed-state drift in the diagnostic, but a finite sampled maximum remains empirical. Computation on disjoint public data does not itself access raw private records. Nevertheless, transmitting any new data-dependent statistic beyond the analyzed release transcript requires separate privacy analysis.

9.4 Certified-SABA

For fixed public membership $\mathcal{T}_t$, Certified-SABA sets $C_{t,r} = \max_{k \in \mathcal{T}_t} U_k^{(r)}$ before the corresponding round-r releases. Thus $U_k^{(r)} \le C_{t,r}$ for all members and the classical-rule local certificate is at most ϵ_0. The transcript certificate is therefore $(R\epsilon_0, R\delta)$. This conclusion is conditional on the public/protocol status of the ceilings and on the fixed-state mechanism boundary.

More precisely, the certified object is the representation mechanism conditional on a fixed encoder state and a declared ceiling. The sweep evaluates $U_k^{(r)}$ on each saved round-r MLP state and checks that the ceiling used to set noise is no smaller. It does not claim that the procedure which trained or selected that state is itself private. Nor does it release the individual spectral norms: adding those values, an updated tier map, or a privately selected ceiling to the server view would enlarge the mechanism and would require its own sensitivity and privacy analysis.

9.5 Tier utility bound

For equal-width profiles over $[\Delta_{\min}, \Delta_{\max}]$, $C_t - U_k \le (\Delta_{\max} - \Delta_{\min})/T$. Consequently,

$$\frac{C_t^2 - U_k^2}{U_k^2} = \frac{(C_t - U_k)(C_t + U_k)}{U_k^2} \le \frac{2\Delta_{\max}(\Delta_{\max} - \Delta_{\min})}{T\Delta_{\min}^2}.$$

This is the $O(1/T)$ claim used in the paper. We make no $O(1/T^2)$ claim.

10 Experimental Setup Details

Architecture and training. MLP client encoders have depths in $\{2, 4, 6, 8\}$, hidden dimension 256, and ReLU activations. ViT experiments use ViT-Tiny and ViT-Small with patch size 16 and the `[CLS]` representation. Batch size is 32. Optimization uses AdamW with learning rate 10^{-3}, weight decay 10^{-2}, and five local epochs. CIFAR-10/100 are partitioned across $K \in \{10, 50, 100\}$ clients with Dirichlet $\beta = .5$. Training has $R = 50$ scheduled releases and five seeds. Code will be released upon publication.

The displayed heterogeneity coordinate is computed after training. With

$$P_\Delta = \frac{1}{K}\sum_{k=1}^{K} \delta_{\widehat{\Delta}_k} \quad \text{and} \quad \delta_C \text{ the point mass at } C, \tag{25}$$

its operational value is

$$W_1(P_\Delta, \delta_C) = \frac{1}{K}\sum_{k=1}^{K} |\widehat{\Delta}_k - C|. \tag{26}$$

The architecture allocation and trained encoder states induce P_Δ; the plotted points are indexed by the value measured after training. No target W_1 value, tier label derived from it, or Wasserstein statistic is supplied to SABA.

Evidence and accounting. The nominal basic-composition certificate is $R\epsilon_0 = 50(0.1) = 5$ with $\delta_{\text{total}} = 5 \times 10^{-4}$. Direct substitutes the sampled maximum into the classical-rule expression. The displayed RDP diagnostic supplies the same empirical sensitivity to the Opacus accountant. MI-LB is computed from saved transcript releases using the prespecified test in Equation (14). Direct and empirical-input RDP are not upper certificates; MI-LB is a one-sided witness. The certified columns instead use per-round MLP ceilings from (3) and analytic Gaussian calibration when the classical sufficient rule is inapplicable.

More explicitly, for the empirical profile ceiling $C_{t(k),r}$, the Direct quantity is

$$\epsilon_{\text{Direct}} = \sum_{r=1}^{R} \max_k \epsilon_0 \frac{\widehat{\Delta}_k^{(r)}}{C_{t(k),r}}, \qquad \text{Direct gap} = [\epsilon_{\text{Direct}} - R\epsilon_0]^+. \tag{27}$$

The table's $\max \widehat{\Delta} - C$ column is the largest sampled overshoot of the shared C in that condition. The RDP gap is analogously $[\epsilon_{\text{RDP}} - R\epsilon_0]^+$ after giving the accountant the same sampled sensitivity

input. A zero Direct or RDP gap therefore reports only agreement under an empirical input; it is not a worst-case certificate. Every table reports the mean over five independent seeds and $\mathrm{SE} = s/\sqrt{5}$, where s is their sample standard deviation.

10.1 MI-LB audit construction and interpretation

Before training, the audit fixes $m = 5{,}000$ client, round, batch, and slot locations. For each location, $S_i = +1$ inserts the canary and $S_i = -1$ inserts its fixed non-canary counterpart; all other records and the schedule are identical. Each sign flip is therefore a record-change pair contained in replace-one-client adjacency. It is useful for refuting a client-level claim, but it does not search the worst client replacement.

After the single training run, the audit reads only the saved noisy batch-mean release for each location. The final head produces the prespecified log-probability score; the top and bottom $1{,}000$ scores give the $2{,}000$ non-abstaining decisions. The binomial tail in Equation (14) is evaluated at each candidate ϵ, with δ_{total} fixed to the transcript value rather than a per-round value. Inversion uses a fixed 0.05 significance threshold. Neither the score, ranking, abstention rule, nor null grid is tuned separately by method.

The reported ϵ_{LB} is the largest rejected null on that grid; the table then subtracts the common nominal value and truncates at zero. Thus a positive MI-LB gap is evidence against the advertised transcript target for the restricted neighboring pairs tested. A zero gap can mean that the target holds, that this score lacks power, or that the finite audit is too small; it cannot certify privacy. MI-LB consumes neither $\widehat{\Delta}_k^{(r)}$ nor $t(k)$, which is the precise sense in which its agreement with Direct is independent.

Compute and profiles. Experiments ran on four A100 80 GB GPUs and used approximately 12 GPU-hours. On one A100, the $10{,}000$-pair ViT-Tiny phase costs approximately 0.8 GPU-s per client, or 80 GPU-s at $K = 100$. Sorting 100 scalars is below 2 ms on CPU. A fixed profile requires only a scalar multiplier per round; ten-round empirical refresh adds below 5%.

11 Supplementary Experiments and Ablations

11.1 Certified sweep reporting and numerical reconciliation

For each fixed-state MLP release, define

$$e_{k,r} = \epsilon_G(U_k^{(r)}, \sigma_{k,r}, \delta), \qquad e_r = \max_k e_{k,r}, \qquad e_{\mathrm{local,max}} = \max_r e_r. \tag{28}$$

The sweep checks $e_{\mathrm{local,max}}$ when deciding whether every local value is in the classical rule's regime. The transcript value reported in Table 2 is

$$e_T = \sum_{r=1}^{R} e_r, \tag{29}$$

so it uses the largest local client value in each round rather than an average over clients. Certified-SABA has $e_r \le \epsilon_0$ for every round and hence $e_T \le 5.00$. The displayed value 5.00 is the nominal certified ceiling, not an estimate of the mechanism's minimal privacy parameter.

The repeated values in the submitted tables have distinct meanings. In the empirical CIFAR-10 table, 7.44 is the $W_1 = .90$ *Direct gap*, so its empirical-input total is $5 + 7.44 = 12.44$. In the certified sweep, the $W_1 = .30$ shared-C total is 7.48, while 12.44 is the $W_1 = .90$ total. The separate "Original Table 4 configuration" retains 7.44 only for that legacy configuration. Neither certificate column in the certified sweep uses the sampled maxima that define Direct.

11.2 Clipping and per-client profiles

Clipping a vector to radius C restricts the distance between two outputs to at most $2C$. Thus neither global nor tuned Rep-Clip is labeled a certified sensitivity-C mechanism here. Table 4 reports its empirical utility role alongside the profile continuum.

Table 4: Representative profile/utility comparison on CIFAR-10. Empirical methods require valid ceilings before they can support a certificate.

Method	Profiles	Evidence status	Acc. (%)
ORACLE (empirical per-client)	K	utility reference	$80.8\pm.2$
Empirical SABA	$T = 7$	not certified	$80.6\pm.2$
Tuned per-client Rep-Clip	K	empirical baseline	$79.1\pm.3$
Global Rep-Clip	1	empirical baseline	$76.6\pm.3$

Tuned per-client clipping is 1.7 accuracy points below the Oracle utility reference in this setting. This comparison concerns measured utility and attack behavior; neither clipping row is relabeled as a sensitivity-C certificate.

11.3 Dynamic empirical sensitivity

Table 5 reports a Direct-style drift diagnostic, not a certified privacy-loss quantity. More frequent refresh improves this sampled diagnostic at additional cost.

Table 5: Dynamic Empirical SABA drift diagnostic ($R = 50$).

Refresh frequency	Added compute	Acc. (%)	Direct drift diagnostic
Static (round 0 only)	0%	$80.6\pm.2$	.05
Every 10 rounds	$< 5\%$	$80.8\pm.2$	.01
Every round	$\approx 50\%$	$80.9\pm.2$	.00

11.4 CIFAR-100 empirical results

Table 6: CIFAR-100 empirical evidence ($K = 50$, $R = 50$). Direct and MI-LB have the same interpretations as in Table 1.

Method	W_1	$\max \hat{\Delta} - C$	Direct gap	MI-LB gap	Acc. (%)	Δ_{acc}
Naive	.30	.62	$2.48\pm.05$	$1.15\pm.06$	$62.1\pm.4$	—
Naive	.60	1.24	$4.96\pm.09$	$2.20\pm.08$	$61.3\pm.5$	—
ORACLE	.30	—	.00	$.02\pm.01$	$61.9\pm.3$	$-.2$
ORACLE	.60	—	.00	$.03\pm.01$	$61.0\pm.4$	$-.3$
Empirical SABA	.30	.62	$.42\pm.04$	$.18\pm.03$	$61.5\pm.4$	$-.6$
Empirical SABA	.60	1.24	$.83\pm.06$	$.35\pm.04$	$60.8\pm.5$	$-.5$
Rep-Clip	.30	—	.00	$.01\pm.01$	$55.3\pm.6$	-6.8

11.5 Number of profiles

Figure 2 is an empirical profile-count/utility curve. It is consistent with the $O(1/T)$ upper bound, and it does not imply that tiering dominates the $T = K$ per-client endpoint.

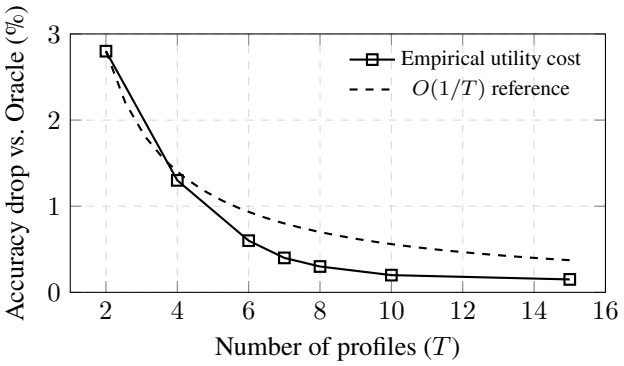

Figure 2: Empirical profile/utility trade-off for $K = 100$ on CIFAR-10.

11.6 Distribution shift

Figure 3 reports accuracy, not a privacy certificate, as public/private shift increases. Empirical SABA degrades from 80.6% to 78.0%. The fixed-state MLP spectral configuration is independent of the public batch sample and stays at 78.2% in this experiment; its certificate remains conditional on the mechanism boundary stated in Definition 2.6.

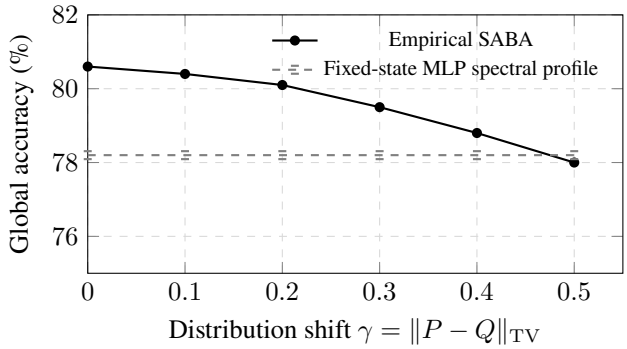

Figure 3: Accuracy under public/private distribution shift.

12 Extended Related Work

Split and hierarchical FL. HSFL [Deng et al., 2023] and hierarchical split-FL [Lin et al., 2025] study system organization, offloading, convergence, and resource allocation when learning is split across clients and edge/cloud servers. The mechanism analyzed here is narrower: the honest-but-curious edge observes each client-side-noised intermediate representation. Secure aggregation, trusted execution, or an aggregate-only interface would expose a different transcript and are not ruled out by the observed-transcript corollary.

Heterogeneous privacy and accounting. Heterogeneous-DP FL and FedHDP [Aldaghri et al., 2021] allocate privacy according to differing client preferences, while private heterogeneous FL without a trusted server [Gao et al., 2024] studies noise-aware aggregation under heterogeneous local privacy requirements. GDPFed [Xu et al., 2026] studies sampling with heterogeneous client privacy; UDP-FL [Feng et al., 2024] harmonizes heterogeneous randomization mechanisms with an accountant; L-RDP [Behnia et al., 2025] gives fixed-memory, per-client local accounting; and PrivateDFL [Piran et al., 2025] addresses cumulative noise in decentralized FL. Those methods start from a specified local randomizer or valid local privacy input. For the unclipped activation release studied here, an adaptation must still supply a valid representation-sensitivity ceiling. We therefore make no empirical dominance claim without implementing each method under the same release and trust boundary.

Calibration and auditing. The classical Gaussian mechanism [Dwork and Roth, 2014] and analytic Gaussian calibration [Balle and Wang, 2018] both require a valid query sensitivity; standard RDP accounting [Mironov, 2017] does not repair an invalid local input. One-run privacy auditing [Steinke et al., 2023] instead supplies a one-sided empirical rejection test. Accordingly, this paper keeps three claims separate: Direct/RDP with sampled sensitivity are diagnostics, MI-LB is an attack witness, and only the fixed-state MLP sweep with valid ceilings is a certificate.