Privacy Risk Scales with Effective Dimension in Federated Learning

Md Mamunur Rashid[1] **Md Palash Uddin**[1] **Yong Xiang**[1] **Keshav Sood**[1]
Longxiang Gao[2]

[1]School of Information Technology, Deakin University, Melbourne, Australia
[2]Shandong Computer Science Center, Qilu University of Technology, China; and
Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing,
Shandong Fundamental Research Center for Computer Science, Jinan, China
{md.m.rashid, m.uddin, yong.xiang, keshav.sood}@deakin.edu.au
gaolx@sdas.org

Abstract

Privacy mechanisms in federated learning (FL) are often calibrated without explicit regard to model scale, implicitly assuming that privacy risk remains stable as federated models grow. We challenge this assumption by introducing *federated leakage*, a signal-level mutual-information measure of client-specific information surviving in privatized update trajectories, and by identifying *effective model dimension*, rather than raw parameter count, as the operative scaling variable. Under a Gaussian signal-channel model with spectral growth of client-informative directions, we establish a regime-conditional scaling law: leakage grows as $\Theta(d/\log d)$ with effective dimension d. This result yields a leakage-ratio notion of *privacy debt*, where fixed-noise deployment can expose increasing client-informative signal even when the accountant-reported DP configuration is unchanged. We therefore propose a scale-aware Gaussian calibration rule that preserves the target leakage regime up to constant factors, and show across vision and language settings that it substantially reduces cross-scale privacy drift relative to fixed-noise baselines.

1 Introduction

Model scaling is a central feature of modern machine learning (ML) [1]. Larger models improve performance, transferability, and data efficiency, motivating their deployment across a wide range of applications. At the same time, federated learning (FL) has become a standard paradigm for privacy-sensitive settings, enabling collaborative training without centralizing raw data [2]. As these trends converge, one basic question remains poorly understood: *how does privacy risk evolve as federated models scale?* The FL privacy literature studies membership inference, reconstruction, secure aggregation, and differential privacy (DP), but typically at fixed model scale [2, 3, 4]. Scaling-law studies characterize how utility changes with model size, data, and compute, but focus on predictive performance rather than privacy [5, 6]. Consequently, federated pipelines often calibrate privacy mechanisms independently of model scale, implicitly assuming that protection remains adequate as models grow.

There are strong reasons to question this assumption [7]. In heterogeneous FL, local updates can encode client-specific patterns that become more expressive as model capacity grows. A mechanism that is adequate at one scale may therefore become under-protective at a larger scale even when the protocol is unchanged. Recent studies report stronger memorization, extraction, and membership-inference behavior in larger models [8, 9], but do not explain how privacy risk should scale in

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

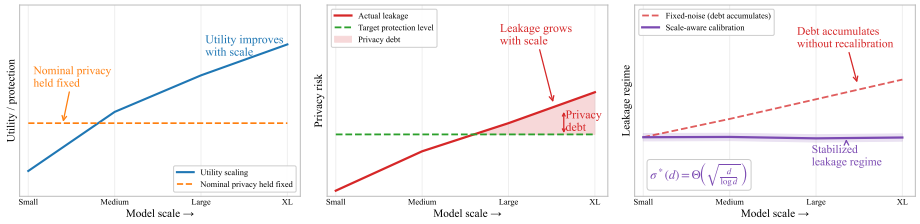


Figure 1: Privacy risk is not scale-neutral in FL. (A) Utility improves with model scale under fixed privacy settings. (B) Leakage grows with effective dimension, creating privacy debt. (C) The scale-aware calibration rule $\sigma_{\mathrm{SA}}(d) = \Theta\left(\sqrt{d/\log d}\right)$ stabilizes privacy risk across scales.

federated settings. This creates what we term *privacy debt*: a growing mismatch between intended protection and effective risk as models scale, as illustrated in Figure 1.

In this paper, we treat privacy as a scale-dependent property of federated optimization. We introduce *federated leakage*, a signal-level mutual-information measure of client-specific information in privatized update trajectories, and identify *effective model dimension* as the operative scaling variable. Under a Gaussian signal-channel model with spectral growth of client-informative directions, leakage grows as $\Theta(d/\log d)$, motivating a leakage-ratio view of *privacy debt*: fixed-noise deployment can expose increasing client-specific signal even when the accountant-reported DP configuration is unchanged. We then propose a scale-aware Gaussian calibration rule that preserves the target leakage regime up to constant factors. Our main contributions are as follows:

- We introduce *federated leakage* as a signal-level formalization of scaling-dependent client privacy risk, identify effective dimension as the operative scaling variable, and derive a regime-conditional $\Theta(d/\log d)$ leakage law.
- We define *privacy debt* as a leakage-ratio index quantifying how fixed-noise deployment exposes increasing client-informative signal as effective dimension grows; this is an analysis-side leakage measure, not an accountant-reported DP budget.
- We propose a scale-aware Gaussian calibration rule and show empirically, across heterogeneous federated settings, model scales, and privacy mechanisms, that it substantially reduces cross-scale privacy drift relative to fixed-noise deployment.

2 Related Work

Privacy-preserving Federated Optimization. A substantial literature integrates DP into FL through clipping, client-level perturbation, and privacy accounting [2, 10, 3, 4, 11, 12]. More recent work studies adaptive privacy spending, improved privacy–utility trade-offs, partial participation, benchmark-driven DP-FL systems, and decentralized privacy-preserving FL [13, 14, 15, 16, 17]. These works improve mechanisms and accounting, but they largely analyze privacy at fixed model scale or under fixed calibration assumptions.

Leakage from Gradients and Update Trajectories. Model updates are a real privacy surface. Gradient inversion and reconstruction attacks can recover training examples from shared updates [18, 19, 20, 21], while related work studies membership inference, attribute inference, and other update-based privacy risks [22, 23, 24]. However, this line of work establishes the existence of leakage without characterizing how client-specific leakage in privatized update trajectories changes as federated models scale.

Scaling, Memorization, and Model Capacity. Scaling-law studies show regular utility trends with model size, data, and compute [5, 6, 25], and larger models are known to memorize more content and exhibit stronger extraction behavior [26, 8, 9]. However, memorization and federated leakage are distinct: the former concerns what a trained model retains about its corpus, while the latter concerns what an observer infers about a specific client from collaborative updates. Most memorization analyses are centralized and do not identify the operative scaling variable for client-level privacy risk.

Closest Connections and Our Distinction. Closest work improves DP-FL mechanisms, communication-efficient privatization, private LoRA-FL, and privacy accounting under fixed budgets [27, 14, 28, 29, 17, 30, 31]. Our focus is different: whether the client-informative signal exposed by a fixed-noise update channel remains stable as effective dimension grows. Standard RDP accounting tracks sensitivity, sampling, composition, and noise scale [11, 32], but not the scale-dependent covariance spectrum of client-specific update signals. Our contribution is a signal-level scaling analysis and empirical validation of privacy drift, not a replacement for DP accounting.

3 Analytical Framework and Problem Formulation

3.1 FL Setting

We consider standard cross-device FL with K clients and a central coordinator. Client $k \in \{1, \dots, K\}$ holds a private dataset $\mathcal{D}_k$ drawn from client-specific distribution P_k. Let $\mathbf{w}^{(t)} \in \mathbb{R}^p$ denote the global model at round t, where p is the raw parameter count. At each round, a subset $\mathcal{S}_t$ receives the current model, performs local optimization, and returns an update to the server. Client k minimizes $F_k(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_k}[\ell(\mathbf{w}; \xi)]$, and the global objective is

$$F(\mathbf{w}) = \sum_{k=1}^{K} p_k F_k(\mathbf{w}), \tag{1}$$

where $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$ are aggregation weights, typically proportional to client data sizes. After local training, client k transmits an update $\Delta_k^{(t)}$ or an equivalent model increment.

3.2 Effective Model Dimension

Raw parameter count is not the right privacy-scaling variable. In overparameterized models, many coordinates contribute little to the client-specific structure of local updates. We therefore define an *effective model dimension* from the covariance of client gradients. Let $\Sigma_k(\mathbf{w}) = \mathbb{E}_{\xi \sim \mathcal{D}_k}\big[\nabla_{\mathbf{w}} \ell(\mathbf{w}; \xi) \nabla_{\mathbf{w}} \ell(\mathbf{w}; \xi)^\top\big]$ denote client k's local gradient covariance, and let $\bar{\Sigma}(\mathbf{w}) = \sum_{k=1}^{K} p_k \Sigma_k(\mathbf{w})$ denote the aggregated covariance. Let $\lambda_1(\bar{\Sigma}(\mathbf{w})) \geq \lambda_2(\bar{\Sigma}(\mathbf{w})) \geq \cdots \geq 0$ be its eigenvalues in descending order. For a retained-variance threshold $\tau \in (0, 1)$, define

$$d_\tau \triangleq \min \left\{ r \in \{1, \dots, p\} : \frac{\sum_{i=1}^{r} \lambda_i\big(\bar{\Sigma}(\mathbf{w})\big)}{\sum_{i=1}^{p} \lambda_i\big(\bar{\Sigma}(\mathbf{w})\big)} \geq \tau \right\}. \tag{2}$$

Thus d_τ is the smallest number of principal update-space directions explaining at least fraction τ of aggregated covariance energy; we fix $\tau = 0.99$ and write $d \equiv d_{0.99}$ throughout. In the theory, d indexes the privacy-relevant signal covariance $\mathbf{A}_k(d)$. In experiments, we estimate d from the empirical update-spectrum induced by client-round update observations, implemented with randomized SVD [33, 34]. We use $\tau = 0.99$ to retain nearly all client-informative update energy while excluding small residual directions; Appendix D shows nearby thresholds preserve the conclusions.

3.3 Heterogeneity and the Privatized Update Channel

Client-level privacy risk is meaningful only under heterogeneity, since identical client distributions would remove systematic client-specific signal. We therefore assume a standard non-IID regime.

Assumption 1 (Client heterogeneity)**.** *The client distributions $P_1, \dots, P_K$ are non-identical. In particular, there exist $j \neq k$ such that $\mathrm{TV}(P_j, P_k) \geq \gamma > 0$, where $\mathrm{TV}(\cdot, \cdot)$ denotes total variation distance and γ is a heterogeneity constant.*

Under Assumption 1, local updates may contain non-vanishing client-specific components [35]. In the implemented DP-FL protocols, updates are clipped and perturbed before aggregation, while the theory abstracts this process as a Gaussian signal channel over the client-informative trajectory component. This abstraction is intentionally signal-level: it does not claim that fixed globally L_2-clipped updates have growing covariance trace, since fixed clipping imposes a fixed trace bound.

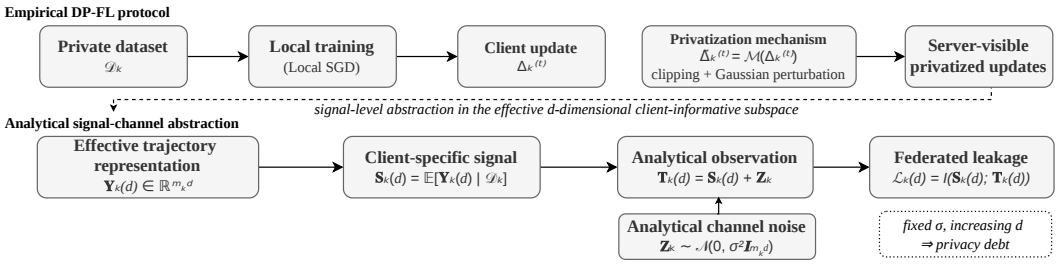


Figure 2: Signal-level privatized update channel. The empirical DP-FL protocol produces server-visible privatized updates, which are abstracted into an effective trajectory used to define the analytical signal channel and federated leakage.

Instead, the theory isolates how client-informative directions can proliferate with effective dimension, and the experiments test whether fixed-noise DP-FL exhibits compatible finite-scale privacy drift. In this abstraction, the server observes $\widetilde{\Delta}_k^{(t)} = \mathcal{M}(\Delta_k^{(t)})$, with $\mathcal{M}(\Delta) = \Delta + \mathbf{z}$ and $\mathbf{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$, where d is the effective dimension and σ is the Gaussian noise scale. The resulting information flow is summarized in Figure 2.

Threat Model. This work considers an honest-but-curious server that has access to the privatized update channel defined by $\widetilde{\Delta}_k^{(t)} = \mathcal{M}(\Delta_k^{(t)})$. The adversary does not observe raw client data or unclipped pre-noise updates, but attempts to infer client-specific information from the observable privatized trajectories, consistent with prior privacy threat models for FL [23]. We instantiate this threat empirically through membership inference across both modalities and canary-based extraction in the language setting [23, 36]. Poisoning, Byzantine behavior, collusion among malicious clients, and attacks that bypass the deployed privatization mechanism are out of the scope.

3.4 Privacy Risk and Federated Leakage

To formalize scaling-dependent privacy risk, let $\mathcal{I}_k \subseteq \{1, \dots, T\}$ denote the participation rounds of client k and let $m_k = |\mathcal{I}_k|$. Let $\mathbf{Y}_k(d) \in \mathbb{R}^{m_k d}$ be the effective update-trajectory representation in the d-dimensional client-informative subspace, and define the client-specific signal as $\mathbf{S}_k(d) \triangleq \mathbb{E}[\mathbf{Y}_k(d) \mid \mathcal{D}_k]$, which removes local optimization stochasticity while retaining the component systematically induced by the client dataset. The analytical observation is modeled as an additive Gaussian channel, $\mathbf{T}_k(d) = \mathbf{S}_k(d) + \mathbf{Z}_k$ with $\mathbf{Z}_k \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{m_k d})$.

Definition 1 (Federated leakage)**.** *The federated leakage of client k at effective dimension d is*

$$\mathcal{L}_k(d) \triangleq I(\mathbf{S}_k(d); \mathbf{T}_k(d)) \,, \tag{3}$$

the mutual information between the client-specific signal and the observable effective trajectory in the analytical channel.

Federated leakage is a tractable signal-level surrogate, not a replacement for dataset-level privacy accounting. Under the Gaussian signal model in Section 4, the log-det expression for $\mathcal{L}_k(d)$ is exact; without Gaussianity, it should be read as a covariance-constrained Gaussian upper bound rather than a fully distribution-free equality [37, 38].

3.5 Privacy Debt

Let d_0 be a reference effective dimension and let σ_0 be the Gaussian noise level calibrated at that reference scale. We define the *leakage-ratio debt* as $D_{\mathcal{L}}(d; d_0) \triangleq \mathcal{L}_k(d; \sigma_0)/\mathcal{L}_k(d_0; \sigma_0)$, which measures how much larger the signal-level federated leakage becomes when the same noise level is reused at effective dimension d instead of at d_0. By Theorem 2, $D_{\mathcal{L}}(d; d_0) = \Theta(d \log d_0/(d_0 \log d))$. For interpretability, we also report the square-root leakage-debt index

$$\varepsilon_{\mathrm{debt}}(d) \triangleq \varepsilon_{\mathrm{target}} \sqrt{D_{\mathcal{L}}(d; d_0)} = \Theta\left(\varepsilon_{\mathrm{target}} \sqrt{\frac{d \log d_0}{d_0 \log d}}\right). \tag{4}$$

The quantity $\varepsilon_{\mathrm{debt}}(d)$ is an analysis-side leakage index, not an accountant-reported DP guarantee or an equivalent formal privacy budget. It summarizes the accumulated *privacy debt*: the increase in client-informative signal exposure caused by reusing a fixed-noise channel as the effective dimension grows. Preserving the leakage regime, therefore, requires a calibration rule that depends on effective dimension.

4 Theoretical Analysis

4.1 Setup and Assumptions

For client k, let $m_k = |\mathcal{I}_k|$ be the number of participation rounds, $\mathbf{S}_k(d) \in \mathbb{R}^{m_k d}$ the client-specific signal in the effective update-trajectory subspace, and $\mathbf{A}_k(d) \triangleq \mathrm{Cov}(\mathbf{S}_k(d))$. We study the Gaussian signal channel

$$\mathbf{T}_k(d) = \mathbf{S}_k(d) + \mathbf{Z}_k, \qquad \mathbf{Z}_k \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{m_k d}), \tag{5}$$

and characterize how $I(\mathbf{S}_k(d); \mathbf{T}_k(d))$ scales with effective dimension d.

Assumption 2 (Regular local objectives)**.** *Each local objective F_k is L-smooth in the neighborhood of the iterates considered in the update-trajectory model.*

Assumption 3 (Spectral growth of informative directions)**.** *For each client k, let $\lambda_{k,1}(d), \ldots, \lambda_{k,r_k(d)}(d)$ be the nonzero eigenvalues of $\mathbf{A}_k(d)$. There exist constants $0 < c_{1,k} \le c_{2,k} < \infty$ and $0 < a_k \le b_k < \infty$ such that, for all sufficiently large d,*

$$c_{1,k} m_k d \le r_k(d) \le c_{2,k} m_k d, \qquad \frac{a_k}{\log d} \le \lambda_{k,i}(d) \le \frac{b_k}{\log d}, \quad i = 1, \ldots, r_k(d). \tag{6}$$

Assumption 3 states that the number of client-informative directions grows linearly with $m_k d$, while their average strength decays as $1/\log d$, a regime supported by the spectral diagnostics in Figure 3. It implies

$$\mathrm{tr}(\mathbf{A}_k(d)) = \Theta\left(\frac{m_k d}{\log d}\right). \tag{7}$$

Assumption 4 (Gaussian additive mechanism)**.** *The effective observation satisfies $\mathbf{T}_k(d) = \mathbf{S}_k(d) + \mathbf{Z}_k$, where $\mathbf{Z}_k \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{m_k d})$ is independent of $\mathbf{S}_k(d)$ [39, 11].*

Assumption 5 (Gaussian signal model)**.** *For each client k, $\mathbf{S}_k(d)$ is modeled as Gaussian with covariance $\mathbf{A}_k(d)$. Without this approximation, the same log-det expression is interpreted as a covariance-constrained Gaussian upper bound on mutual information [37].*

Scope of the signal-channel model. Assumption 3 applies to the analytical signal $\mathbf{S}_k(d)$, not to fixed globally L_2-clipped updates. A fixed bound $\|\Delta_k^{(t)}\|_2 \le C$ on every projected update would imply $\mathrm{tr}(\mathbf{A}_k(d)) = O(m_k C^2)$, contradicting spectral growth. Thus the theory is a conditional Gaussian signal-channel analysis; the experiments test whether practical clipped DP-FL exhibits compatible finite-scale attack behavior.

4.2 Main Guarantees

Proof sketches are brief; full proofs are deferred to Appendix B.

Proposition 1 (Dataset-level dominance in the signal channel)**.** *Let $\mathcal{L}_k^{\mathrm{data}} \triangleq I(\mathcal{D}_k; \mathbf{T}_k(d))$ and $\mathcal{L}_k(d) \triangleq I(\mathbf{S}_k(d); \mathbf{T}_k(d))$. Since $\mathbf{S}_k(d)$ is deterministic given $\mathcal{D}_k$ and $\mathbf{T}_k(d) = \mathbf{S}_k(d) + \mathbf{Z}_k$ with independent noise,*

$$\mathcal{D}_k \to \mathbf{S}_k(d) \to \mathbf{T}_k(d) \tag{8}$$

and therefore $\mathcal{L}_k^{\mathrm{data}} = \mathcal{L}_k(d)$. More generally, if $\mathbf{S}_k(d)$ is a signal-level statistic inside a richer observation channel, then $\mathcal{L}_k(d) \le I(\mathcal{D}_k; \mathbf{O}_k)$ for any observation $\mathbf{O}_k$ from which $\mathbf{T}_k(d)$ is generated.

Proof Sketch. The Markov chain follows from construction and independent Gaussian noise; equality follows from the chain rule, and the general inequality follows from data processing.

Proposition 2 (Illustrative sufficient condition for Assumption 3)**.** *Consider* $\mathbf{S}_k(d) = \mathbf{U}_{k,d}\mathbf{c}_{k,d}$*, where* $\mathbf{U}_{k,d} \in \mathbb{R}^{m_k d \times r_k(d)}$ *has orthonormal columns,* $r_k(d) = \alpha_k m_k d$ *for some* $\alpha_k \in (0, 1]$*, and* $\mathrm{Cov}(\mathbf{c}_{k,d}) = \mathrm{diag}(\mu_{k,1}(d), \ldots, \mu_{k,r_k(d)}(d))$ *with*

$$\frac{a_k}{\log d} \le \mu_{k,i}(d) \le \frac{b_k}{\log d}. \tag{9}$$

Then $\mathbf{A}_k(d) = \mathrm{Cov}(\mathbf{S}_k(d))$ *satisfies Assumption 3.*

Proof Sketch. Orthonormal projection preserves the nonzero spectrum, so the nonzero eigenvalues of $\mathbf{A}_k(d)$ are exactly $\{\mu_{k,i}(d)\}_{i=1}^{r_k(d)}$.

Proposition 3 (Population-level support for Assumption 3)**.** *Let* Π *be a population distribution over clients, define* $\mathbf{G}_P(\mathbf{w}) \triangleq \nabla F_P(\mathbf{w}) - \nabla F(\mathbf{w})$ *for* $P \sim \Pi$*, and let*

$$\mathbf{R}(d) \triangleq \mathrm{Cov}_{P\sim\Pi}[\mathbf{G}_P(\mathbf{w})] \tag{10}$$

have nonzero eigenvalues $\nu_1(d), \ldots, \nu_{\rho(d)}(d)$*. Suppose that* $\rho(d) = \Theta(d)$*,* $\nu_j(d) = \Theta(1/\log d)$ *for* $j = 1, \ldots, \rho(d)$*, and the nonzero eigenvalues of* $\mathbf{A}_k(d)$ *can be indexed as* $\lambda_{k,\ell,j}(d)$ *with* $\ell = 1, \ldots, m_k$ *and* $j = 1, \ldots, \rho(d)$ *such that* $\lambda_{k,\ell,j}(d) = \Theta(\nu_j(d))$ *up to constant-factor truncation. Then* $\mathbf{A}_k(d)$ *satisfies Assumption 3.*

Proof Sketch. The conditions imply $r_k(d) = \Theta(m_k\rho(d)) = \Theta(m_k d)$ and nonzero eigenvalues $\Theta(1/\log d)$, exactly matching Assumption 3. This is a population-level decomposition, not a finite-K rank argument.

Theorem 1 (Conditional signal-channel leakage scaling)**.** *Under Assumptions 3, 4, and 5, for any fixed* $\sigma > 0$*,*

$$\mathcal{L}_k(d) = \Theta\left(\frac{m_k d}{\sigma^2 \log d}\right), \qquad \frac{\mathcal{L}_k(d)}{m_k} = \Theta\left(\frac{d}{\sigma^2 \log d}\right). \tag{11}$$

Proof Sketch. The Gaussian channel gives

$$\mathcal{L}_k(d) = \frac{1}{2}\log\det\left(\mathbf{I} + \frac{\mathbf{A}_k(d)}{\sigma^2}\right) = \frac{1}{2}\sum_{i=1}^{r_k(d)} \log\left(1 + \frac{\lambda_{k,i}(d)}{\sigma^2}\right). \tag{12}$$

By Assumption 3, $r_k(d) = \Theta(m_k d)$ and $\lambda_{k,i}(d) = \Theta(1/\log d)$. For fixed σ, $\lambda_{k,i}(d)/\sigma^2 \le 1$ eventually, so $\log(1 + x) = \Theta(x)$ and $\mathcal{L}_k(d) = \Theta(\mathrm{tr}(\mathbf{A}_k(d))/\sigma^2)$. At finite scales, this uses the linear log-det regime for $x = \lambda_{k,i}(d)/\sigma^2$; Appendix D.6 reports a 6.8% mean relative error on the empirical spectra.

Theorem 2 (Leakage-ratio privacy debt under fixed-noise deployment)**.** *Let* d_0 *be a reference effective dimension and* $\sigma_0 > 0$ *the noise level calibrated at that scale. Reusing* σ_0 *at larger* d *yields*

$$D_{\mathcal{L}}(d; d_0) \triangleq \frac{\mathcal{L}_k(d; \sigma_0)}{\mathcal{L}_k(d_0; \sigma_0)} = \Theta\left(\frac{d \log d_0}{d_0 \log d}\right). \tag{13}$$

Consequently, the square-root leakage index

$$\varepsilon_{\mathrm{debt}}(d) \triangleq \varepsilon_{\mathrm{target}}\sqrt{D_{\mathcal{L}}(d; d_0)} = \Theta\left(\varepsilon_{\mathrm{target}}\sqrt{\frac{d \log d_0}{d_0 \log d}}\right) \tag{14}$$

is an analysis-side leakage index, not an accountant-reported DP budget.

Proof Sketch. Apply Theorem 1 at d and d_0 under the same noise σ_0 and take the ratio; this avoids matching to a vanishing reference-scale noise level.

Theorem 3 (Order-optimal scale-aware signal calibration)**.** *Let* σ_0 *be the Gaussian noise level at reference dimension* d_0*. The schedule*

$$\sigma_{\mathrm{SA}}(d) = \sigma_0\sqrt{\frac{d \log d_0}{d_0 \log d}} \tag{15}$$

anchors the leading constant at (d_0, σ_0) and preserves the reference leakage regime up to constant factors:

$$\mathcal{L}_k(d; \sigma_{\mathrm{SA}}(d)) = \Theta(\mathcal{L}_k(d_0; \sigma_0)) \,. \tag{16}$$

If $\sigma_0 = \Theta(1/\varepsilon_{\mathrm{target}})$, then $\sigma_{\mathrm{SA}}(d) = \Theta\Big(\sqrt{d/(\varepsilon^2_{\mathrm{target}} \log d)}\Big)$; any $\sigma(d) = o(\sqrt{d/\log d})$ is insufficient to keep the leakage ratio bounded.

Proof Sketch. Substituting $\sigma_{\mathrm{SA}}(d)$ into Theorem 1 keeps $d/(\sigma(d)^2 \log d)$ at the reference order $d_0/(\sigma_0^2 \log d_0)$. Necessity follows from the exact log-det lower bound and $\log(1+x) \geq \min\{x/2, \log 2\}$.

Corollary 1 (Noise-power cost of scale-aware signal calibration)**.** *Under Theorem 3,*

$$\mathbb{E}\big[\|\mathbf{Z}_k\|_2^2\big] = m_k d\, \sigma_{\mathrm{SA}}(d)^2 = \Theta\bigg(\frac{m_k d^2}{\varepsilon^2_{\mathrm{target}} \log d}\bigg) \,, \tag{17}$$

so the per-coordinate noise-power cost scales as $\Theta(d/(\varepsilon^2_{\mathrm{target}} \log d))$.

Proof Sketch. Substitute the calibration law into $\mathbb{E}[\|\mathbf{Z}_k\|_2^2] = m_k d\, \sigma(d)^2$.

Conditional Nature of the Results. These are conditional signal-channel results: if the client-informative covariance has $\Theta(m_k d)$ active directions of strength $\Theta(1/\log d)$, then fixed-noise Gaussian perturbation yields participation-normalized leakage $\Theta(d/\log d)$. They are not theorems that fixed globally clipped DP updates have growing covariance trace; fixed clipping imposes a fixed trace bound. The theory isolates a spectral leakage mechanism, while the experiments test whether practical fixed-noise DP-FL exhibits compatible finite-scale privacy drift.

5 Experiments

5.1 Experimental Setup

We evaluate the framework on four datasets spanning vision and language. CIFAR-100 [40] and a Stack Overflow[1] user subset form the primary analysis, while FEMNIST and Shakespeare from LEAF [41] provide robustness checks. For CIFAR-100, we use $K = 100$ clients with Dirichlet-α partitioning, taking $\alpha = 0.1$ as the primary heterogeneous regime, $\alpha = 1.0$ as a lower-heterogeneity comparison, and the standard pathological two-class-per-client split [2]. Stack Overflow uses the natural per-user split over the top 2,000 users (minimum 50 posts each); FEMNIST and Shakespeare likewise use natural user-level partitions.[2] The vision ladder comprises ResNet-20, 32, 44, 56, 110, and WRN-28-10 [42, 43]; the language ladder comprises GPT-2 Small, Medium, and Large, with LoRA [44] ($r = 8$, $\alpha_{\mathrm{LoRA}} = 16$) applied to Medium and Large. Language-side conclusions are interpreted through effective dimension rather than adaptation pathway alone; supporting matched-effective-dimension evidence appears in Table 8. Throughout, the operative scaling variable is $d = d_{0.99}$ from Section 3, estimated as the smallest retained rank explaining 99% of empirical update-spectrum energy.

Vision experiments use FedAvg [2] with sampling rate $q = 0.1$, $T = 500$ rounds, $E = 5$ local epochs, SGD with momentum 0.9, batch size 32, clipping threshold $C = 1.0$, and learning rate 0.05 (except 0.03 for WRN-28-10). Language experiments use $T = 200$, $E = 1$, Poisson sampling with $q = 0.05$, AdamW, batch size 4, sequence length 128, weight decay 0.01, and learning rates 5×10^{-5} for GPT-2 Small and 10^{-4} for Medium/Large; convergence diagnostics are in Appendix C. We compare four mechanisms: *(i)* non-private FL ($\sigma = 0$); *(ii)* fixed-noise DP-FL, where σ_0 is calibrated at the smallest model to $(\varepsilon, \delta) = (3.0, 10^{-5})$ via RDP [11] and held fixed across scale; *(iii)* adaptive-clipping DP-FL [45] with the same σ_0; and *(iv)* scale-aware DP-FL with $\sigma(d)$ from Theorem 3. Vision results are averaged over five seeds (ResNets) and three seeds (WRN-28-10); language results use three seeds. Privacy is evaluated via membership inference AUC (MIA AUC) and, for language, canary extraction rate [36]; fit candidates are compared by AICc, where lower values indicate a preferred fit. Additional details are in Appendix C.

[1] Used under CC BY-SA 4.0; see `https://stackoverflow.com/help/licensing`

[2] Artifact available at `https://github.com/mamunurrashidete88/PrivScale-FL`.

5.2 Main Results

Privacy risk increases with effective dimension under fixed-noise deployment, the observed trend is compatible with the predicted $d/\log d$ regime, and scale-aware calibration substantially stabilizes privacy risk at measurable utility cost. The evidence combines a six-point vision ladder with a three-point language ladder and should be interpreted as finite-scale. Additional diagnostics are in Appendix D.

Spectral Diagnostics. Figure 3 reports spectral diagnostics across the evaluated ladders. Across these settings, the participation-ratio estimate $\widehat{r}(d)$ grows approximately linearly with effective dimension, the mean informative eigenvalue $\bar{\lambda}(d)$ is consistent with a $1/\log d$ trend, and the total proxy signal power $\mathrm{tr}(\widehat{\mathbf{A}}_{\mathrm{upd}}(d))$ is broadly consistent with $d/\log d$ growth. These diagnostics do not prove Assumption 3, but they place the evaluated settings in a regime empirically compatible with its structural predictions.

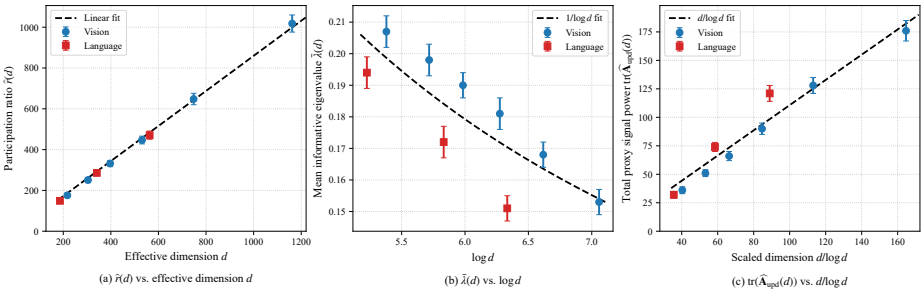


Figure 3: Spectral diagnostics: $\widehat{r}(d)$ grows with d, $\bar{\lambda}(d)$ follows $1/\log d$, and $\mathrm{tr}(\widehat{\mathbf{A}}_{\mathrm{upd}}(d))$ follows $d/\log d$.

Privacy Risk Under Fixed-Noise Deployment. Figure 4 shows that privacy risk rises systematically with effective dimension when the privacy mechanism is held fixed. On the vision ladder, fixed-noise DP-FL increases MIA AUC from 0.547 ± 0.012 at ResNet-20 to 0.704 ± 0.021 at WRN-28-10; on the language ladder, MIA AUC rises from 0.573 ± 0.013 at GPT-2 Small to 0.731 ± 0.019 at GPT-2 Large (LoRA), while extraction increases from 2.4% to 8.2%. The trend is consistent across seeds and heterogeneity settings. The dashed lines are monotone guides rather than fitted scaling laws.

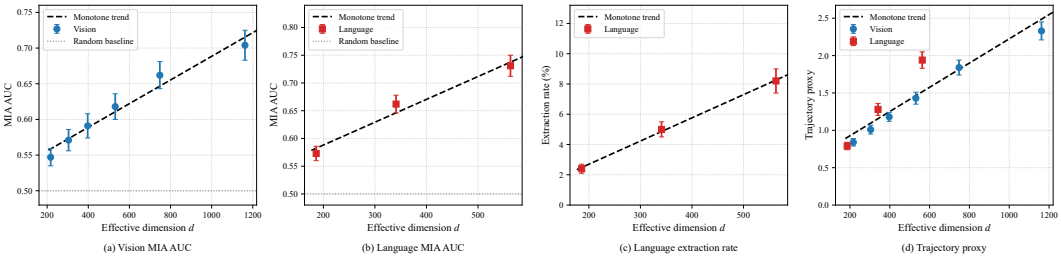
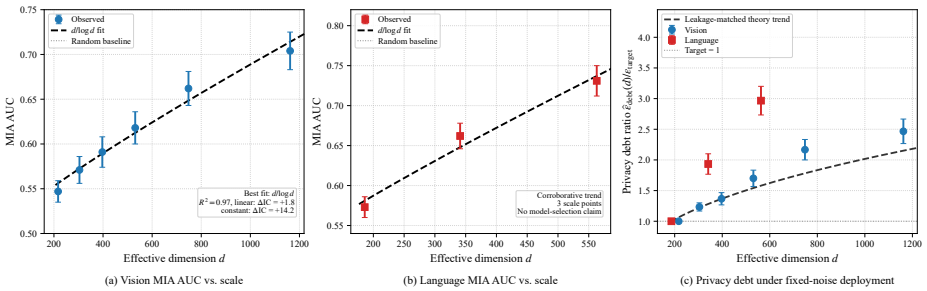


Figure 4: Privacy risk under fixed-noise deployment. Panels (a)–(c) show attack-based observables across vision and language ladders; panel (d) shows the trajectory-level proxy.

Privacy Debt Under Model Growth. Figure 5 connects the observed scaling trend to the leakage-ratio consequence in Theorem 2. The six-point vision ladder provides the primary functional-form evidence: within the tested candidate families, the $d/\log d$ curve fits best and is preferred by AICc over constant and linear alternatives. The three-point language ladder is used only as directionally consistent cross-modality evidence, not for model-selection claims. Panel (c) reports the square-root leakage-debt index $\widehat{\varepsilon_{\mathrm{debt}}}(d)/\varepsilon_{\mathrm{target}}$, which is an analysis-side signal-leakage index rather than an accountant-reported DP budget.

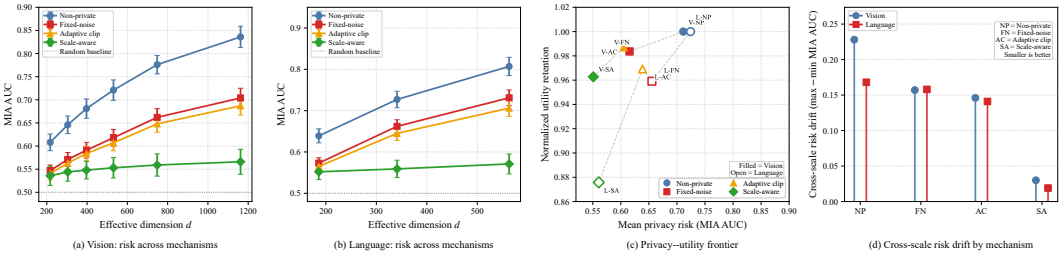
Scale-Aware Mechanism Comparison. Figure 6 compares non-private FL, fixed-noise DP-FL, adaptive-clipping DP-FL, and the proposed scale-aware calibration rule. Fixed-noise DP-FL shows clear cross-scale drift, while adaptive clipping offers only modest reduction; at the largest vision scale, adaptive clipping reaches 0.687 ± 0.020 MIA AUC and is not significantly different from



(a) Vision MIA AUC vs. scale (b) Language MIA AUC vs. scale (c) Privacy debt under fixed-noise deployment

Figure 5: Privacy debt under fixed noise: vision supports the $d/\log d$ fit, language gives directional corroboration, and $\widehat{\varepsilon}_{\mathrm{debt}}/\varepsilon_{\mathrm{target}}$ increases with d.

fixed-noise DP-FL ($p = 0.18$, three seeds). In contrast, the scale-aware rule remains much flatter, with vision MIA ranging from 0.536 ± 0.021 to 0.566 ± 0.027 and language MIA from 0.552 ± 0.019 to 0.571 ± 0.024. Thus, scale-aware calibration stabilizes attack-based privacy risk within DP-FL protocols, but should not be interpreted as a replacement for formal DP accounting.



(a) Vision: risk across mechanisms (b) Language: risk across mechanisms (c) Privacy–utility frontier (d) Cross-scale risk drift by mechanism

Figure 6: Scale-aware calibration stabilizes MIA AUC across vision and language, reducing cross-scale privacy drift relative to non-private, fixed-noise, and adaptive-clipping baselines.

Cross-Scale Mechanism Comparison. Table 1 reports privacy and utility outcomes across scales for all four mechanisms. Three patterns emerge: privacy risk rises with scale under non-private and fixed-noise deployment; adaptive clipping modestly reduces risk but does not remove the scaling trend; and the scale-aware mechanism keeps risk substantially flatter across the full ladder, at progressively larger utility cost, consistent with the $\Theta(d^2/\log d)$ noise-power bound in Corollary 1.

Table 1: Cross-scale privacy and utility comparison across mechanisms.

Mod.	Dataset	Model	d	Non-private		Fixed-noise DP			Adaptive-clip DP			Scale-aware DP		
				Util.	MIA	Util.	MIA	Ext.	Util.	MIA	Ext.	Util.	MIA	Ext.
Vision	CIFAR-100	ResNet-20	217	68.4	0.608±0.018	68.0	0.547±0.012	—	68.2	0.541±0.014	—	67.3	**0.536±0.021**	—
		ResNet-32	304	69.7	0.646±0.019	68.9	0.571±0.015	—	69.1	0.563±0.013	—	67.9	**0.544±0.020**	—
		ResNet-44	397	71.0	0.681±0.021	70.1	0.591±0.017	—	70.3	0.584±0.016	—	68.8	**0.548±0.019**	—
		ResNet-56	531	72.7	0.721±0.022	71.7	0.618±0.018	—	71.9	0.607±0.017	—	70.4	**0.553±0.022**	—
		ResNet-110	748	74.5	0.776±0.020	72.9	0.662±0.019	—	73.2	0.648±0.018	—	71.0	**0.559±0.024**	—
		WRN-28-10	1162	79.1	0.836±0.023	76.5	0.704±0.021	—	76.9	0.687±0.020	—	73.4	**0.566±0.027**	—
Lang.	SO subset (2K)	GPT-2 Small	186	34.3	0.639±0.017	35.0	0.573±0.013	2.4	34.7	0.565±0.014	2.2	36.2	**0.552±0.019**	**1.6**
		GPT-2 Medium (LoRA)	341	30.3	0.727±0.020	31.5	0.662±0.016	5.0	31.1	0.645±0.017	4.8	33.8	**0.559±0.021**	**1.9**
		GPT-2 Large (LoRA)	563	27.5	0.807±0.022	29.4	0.731±0.019	8.2	29.1	0.706±0.020	7.7	35.1	**0.571±0.024**	**2.2**

Note. d: effective dimension. Vision Util.: top-1 accuracy (%, ↑); Language Util.: perplexity (↓). Ext.: canary extraction (%, ↓). MIA: membership inference AUC (↓). Mean ± std over 5 seeds (ResNets) or 3 seeds (WRN, language).

Effective Dimension as Privacy-Scaling Variable. To test whether effective dimension is the more appropriate privacy-scaling variable, we refit the primary-ladder observables under two views: raw parameter count p and retained-variance effective dimension d. Table 2 reports the resulting fit-quality comparison. Across all observables, the effective-dimension view yields stronger fit, with ΔR^2 ranging from $+0.067$ to $+0.097$. The largest gains appear for the trajectory-level proxy ($+0.093$ in vision and $+0.097$ in language), while even attack-based observables improve consistently; for example, MIA AUC improves from 0.842 to 0.913 in vision and from 0.861 to 0.928 in language. These results support the interpretation that privacy debt is driven less by nominal model growth than by the expansion of the active client-informative subspace captured by d.

Table 2: Fit quality on the primary ladders under raw parameter count and effective dimension.

Observable	**Modality**	R^2 **under** p	R^2 **under** d	ΔR^2
MIA AUC	Vision	0.842	0.913	+0.071
MIA AUC	Language	0.861	0.928	+0.067
Extraction rate	Language	0.826	0.901	+0.075
Trajectory proxy	Vision	0.794	0.887	+0.093
Trajectory proxy	Language	0.817	0.914	+0.097

Note. Both views use the same AICc-based pipeline. The reported comparisons describe finite-scale behavior on the primary vision and language ladders and are not intended as asymptotic model identification. Positive ΔR^2 favors the effective-dimension view.

Low-FPR and per-client risk audit. In the additional vision audit, we use 10,000 non-members to estimate TPR at 1% and 0.1% FPR. The per-client attack advantage is the maximum separation between the attack's true-positive and false-positive rates for that client under the evaluation protocol; client P95 summarizes the upper tail of this advantage across clients. Table 3 records the selected observations. Bootstrap 95% confidence-interval half-widths over 2,000 replicates were 0.4–0.9 percentage points for TPR@1% and 0.2–0.5 points for TPR@0.1%; the latter should be read more cautiously.

Table 3: Additional low-FPR vision audit for selected models. Client P95 is the 95th percentile of per-client attack advantage.

Model	Mechanism	TPR@1%	TPR@0.1%	Client P95
ResNet-20	Fixed	2.4%	0.6%	0.118
	Scale-aware	2.2%	0.5%	0.110
ResNet-56	Fixed	5.1%	1.5%	0.257
	Scale-aware	2.6%	0.7%	0.129
WRN-28-10	Fixed	8.7%	3.1%	0.401
	Scale-aware	3.0%	0.8%	0.148

5.3 Limitations and Deployment Scope

The vision ladder provides the main functional-form evidence, while the language ladder offers directionally consistent cross-modality evidence. Our empirical claim concerns privacy drift with measured effective dimension, not proof of the exact asymptotic law. The $\Theta(d/\log d)$ result assumes a particular latent-signal spectrum and does not apply asymptotically to fixed globally clipped updates; the experiments instead use post-clipping spectra to predict finite-scale attack drift and guide calibration. Absolute MIA levels depend on heterogeneity and task-specific spectra, and greater memorization may also contribute to attack success. Six vision and three language scales cannot identify a unique asymptotic law or causally separate these mechanisms. The main experiments estimate d using a preliminary run; a private-data run without protected estimation does not provide an end-to-end DP guarantee for that estimator. A separately evaluated DP-accounted pilot supports feasible multiplier selection (Appendix C.2), but does not retroactively make the main experiments end-to-end private. The isotropic rule also incurs utility cost: at WRN-28-10 it uses $2.02\sigma_0$ and reduces accuracy from 76.5% to 73.4% relative to fixed noise. Its cost and effectiveness beyond the tested scales remain open.

6 Conclusion

This paper shows that privacy risk in FL can drift with effective model dimension rather than remain scale-neutral. Through a conditional signal-channel analysis, we formalized how fixed-noise deployment can accumulate leakage-ratio privacy debt as client-informative directions expand. The proposed scale-aware calibration rule reduces this cross-scale drift in vision and language experiments, with the expected utility cost from stronger noise. These results should be interpreted as signal-level evidence and calibration guidance, not as a replacement for formal DP accounting. Future work should develop private effective-dimension estimators, layerwise noise schedules, and larger-scale validations.

Acknowledgments and Disclosures

This work was supported in part by the Australian Research Council under Grants DP220100983 and DP250102634. The authors declare no competing interests.

References

[1] S. Subramanian, P. Harrington, K. Keutzer, W. Bhimji, D. Morozov, M. W. Mahoney, and A. Gholami, "Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior," *Advances in Neural Information Processing Systems*, vol. 36, pp. 71242–71262, 2023.

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

[3] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2018.

[4] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *International Conference on Learning Representations*, 2018.

[5] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[6] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, *et al.*, "Training compute-optimal large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 30016–30030, 2022.

[7] K. Wei, J. Li, M. Ding, C. Ma, H. Su, B. Zhang, and H. V. Poor, "User-level privacy-preserving federated learning: Analysis and performance optimization," *IEEE Transactions on Mobile Computing*, vol. 21, no. 9, pp. 3388–3401, 2021.

[8] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[9] W. Yu, T. Pang, Q. Liu, C. Du, B. Kang, Y. Huang, M. Lin, and S. Yan, "Bag of tricks for training data extraction from language models," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, pp. 40806–40827, PMLR, 2023.

[10] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[11] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275, IEEE, 2017.

[12] N. Agarwal, P. Kairouz, and Z. Liu, "The skellam mechanism for differentially private federated learning," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 5052–5064, 2021.

[13] S. Kiani, N. Kulkarni, A. Dziedzic, S. Draper, and F. Boenisch, "Differentially private federated learning with time-adaptive privacy spending," in *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Y. Li, L. Fu, T. Wang, J. Lou, B. Chen, L. Yang, J. Shen, Z. Zheng, and C. Chen, "Clients collaborate: Flexible differentially private federated learning with guaranteed improvement of utility-privacy trade-off," in *International Conference on Machine Learning*, pp. 34333–34354, PMLR, 2025.

[15] R. Reshef and K. Y. Levy, "Privacy-preserving federated convex optimization: Balancing partial-participation and efficiency via noise cancellation," in *Forty-second International Conference on Machine Learning*, 2025.

[16] M. Pelikan, S. S. Azam, V. Feldman, J. Silovsky, K. Talwar, C. Brinton, and T. Likhomanenko, "Enabling differentially private federated learning for speech recognition: Benchmarks, adaptive optimizers, and gradient clipping," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[17] X. Li, C. Wang, B. Su, Q. Long, and W. J. Su, "Mitigating the privacy–utility trade-off in decentralized federated learning via f-differential privacy," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[18] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," *Advances in neural information processing systems*, vol. 32, 2019.

[19] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients-how easy is it to break privacy in federated learning?," *Advances in neural information processing systems*, vol. 33, pp. 16937–16947, 2020.

[20] J. Deng, Y. Wang, J. Li, C. Wang, C. Shang, H. Liu, S. Rajasekaran, and C. Ding, "Tag: Gradient attack on transformer-based language models," in *Findings of the association for computational linguistics: EMNLP 2021*, pp. 3600–3610, 2021.

[21] M. Balunovic, D. Dimitrov, N. Jovanović, and M. Vechev, "Lamp: Extracting text from gradients with language model priors," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7641–7654, 2022.

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

[23] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753, IEEE, 2019.

[24] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706, IEEE, 2019.

[25] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish, "Scaling laws for transfer," *arXiv preprint arXiv:2102.01293*, 2021.

[26] K. Tirumala, A. Markosyan, L. Zettlemoyer, and A. Aghajanyan, "Memorization without overfitting: Analyzing the training dynamics of large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38274–38290, 2022.

[27] F. Tramer and D. Boneh, "Differentially private learning needs better features (or much more data)," in *International Conference on Learning Representations*, 2021.

[28] R. Mckenna, Y. Huang, A. Sinha, B. Balle, Z. Charles, C. A. Choquette-Choo, B. Ghazi, G. Kaissis, R. Kumar, R. Liu, D. Yu, and C. Zhang, "Scaling laws for differentially private language models," in *Proceedings of the 42nd International Conference on Machine Learning*, pp. 43375–43398, 2025.

[29] Q. Li, Z. Chen, and A. Banerjee, "Sketched gaussian mechanism for private federated learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[30] S. Lee, S. Park, D. B. Lee, D. Wagner, H. Seong, T. Bocklet, J. Lee, and S. J. Hwang, "FedSVD: Adaptive orthogonalization for private federated learning with loRA," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[31] J. Liu, Y. Miao, N. Xi, and J. Liu, "Rethinking loRA for privacy-preserving federated learning in large models," in *The Fourteenth International Conference on Learning Representations*, 2026.

[32] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato, "Hypothesis testing interpretations and renyi differential privacy," in *International Conference on Artificial Intelligence and Statistics*, pp. 2496–2506, PMLR, 2020.

[33] A. P. Shikhaliev, L. C. Potter, and Y. Chi, "Low-rank structured covariance matrix estimation," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 700–704, 2019.

[34] S. Fort and A. Scherlis, "The goldilocks zone: Towards better understanding of neural network loss landscapes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3574–3581, 2019.

[35] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*.

[36] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

[37] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[38] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[39] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and trends® in theoretical computer science*, vol. 9, no. 3-4, pp. 211–487, 2014.

[40] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[41] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[43] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[44] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[45] G. Andrew, O. Thakkar, B. McMahan, and S. Ramaswamy, "Differentially private learning with adaptive clipping," *Advances in neural information processing systems*, vol. 34, pp. 17455–17466, 2021.

[46] Y.-X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled rényi differential privacy and analytical moments accountant," in *The 22nd international conference on artificial intelligence and statistics*, pp. 1226–1235, PMLR, 2019.

[47] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

A Notation

Table 4 summarizes the key notation used throughout this work.

Table 4: Summary of key symbols.

Symbol	Meaning
K	Number of clients.
$\mathcal{D}_k$	Private dataset of client k.
$\mathbf{w}^{(t)}$	Global model at communication round t.
p	Raw model parameter count.
d_τ, d	Retained-variance effective dimension; throughout, $d = d_{0.99}$.
$\Delta_k^{(t)}$	Client update or model increment at round t.
$\widetilde{\Delta}_k^{(t)}$	Privatized update observed by the server.
$\mathcal{M}$	Privatization mechanism applied to client updates.
C	Clipping threshold used in empirical DP-FL protocols.
σ	Gaussian noise scale in the analytical channel or empirical mechanism.
σ_0	Reference noise level calibrated at dimension d_0.
$\sigma_{\mathrm{SA}}(d)$	Scale-aware noise schedule at effective dimension d.
$\mathcal{I}_k$, m_k	Participation rounds of client k and their count, $m_k = \lvert\mathcal{I}_k\rvert$.
$\mathbf{Y}_k(d)$	Effective update-trajectory representation of client k.
$\mathbf{S}_k(d)$	Client-specific signal, $\mathbb{E}[\mathbf{Y}_k(d) \mid \mathcal{D}_k]$.
$\mathbf{Z}_k$	Gaussian channel noise, $\mathbf{Z}_k \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{m_k d})$.
$\mathbf{T}_k(d)$	Analytical observation, $\mathbf{S}_k(d) + \mathbf{Z}_k$.
$\mathbf{A}_k(d)$	Signal covariance, $\mathrm{Cov}(\mathbf{S}_k(d))$.
$r_k(d)$, $\lambda_{k,i}(d)$	Rank and nonzero eigenvalues of $\mathbf{A}_k(d)$.
$\mathcal{L}_k(d)$	Federated leakage, $I(\mathbf{S}_k(d); \mathbf{T}_k(d))$.
d_0	Reference effective dimension for fixed-noise calibration.
$D_{\mathcal{L}}(d; d_0)$	Leakage-ratio debt, $\mathcal{L}_k(d; \sigma_0)/\mathcal{L}_k(d_0; \sigma_0)$.
$\varepsilon_{\mathrm{target}}$	Reference target privacy budget.
$\varepsilon_{\mathrm{debt}}(d)$	Square-root leakage-debt index, $\varepsilon_{\mathrm{target}}\sqrt{D_{\mathcal{L}}(d; d_0)}$.
$\widehat{\mathbf{A}}_{\mathrm{upd}}(d)$, $\widehat{r}(d)$, $\bar{\lambda}(d)$	Empirical update-spectrum proxy, participation-ratio estimate, and mean informative eigenvalue.

B Full Proofs of the Proposition, Theorem, and Corollary

This appendix provides full proofs of the main results stated in Section 4. The notations follow the signal-channel formulation in the main paper. In particular,

$$\mathbf{Y}_k(d) \in \mathbb{R}^{m_k d}, \qquad \mathbf{S}_k(d) \triangleq \mathbb{E}[\mathbf{Y}_k(d) \mid \mathcal{D}_k], \qquad \mathbf{T}_k(d) = \mathbf{S}_k(d) + \mathbf{Z}_k, \tag{18}$$

where

$$\mathbf{Z}_k \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{m_k d}), \qquad \mathbf{A}_k(d) \triangleq \mathrm{Cov}(\mathbf{S}_k(d)), \tag{19}$$

and

$$\mathcal{L}_k(d) \triangleq I(\mathbf{S}_k(d); \mathbf{T}_k(d)). \tag{20}$$

Asymptotic Convention. Unless stated otherwise, all asymptotic notation is with respect to $d \to \infty$, while client-specific quantities such as m_k, $c_{1,k}$, $c_{2,k}$, a_k, b_k, and calibration constants at the reference scale are treated as fixed. We write $f(d) \asymp g(d)$ to denote equality up to positive d-independent multiplicative constants.

Elementary Logarithmic Bound. For $x \in [0, 1]$,

$$\frac{x}{2} \le \log(1 + x) \le x. \tag{21}$$

B.1 Proof of Proposition 1

Proof. By definition,

$$\mathbf{S}_k(d) = \mathbb{E}[\mathbf{Y}_k(d) \mid \mathcal{D}_k], \tag{22}$$

so $\mathbf{S}_k(d)$ is a deterministic function of $\mathcal{D}_k$. Under the analytical Gaussian channel,

$$\mathbf{T}_k(d) = \mathbf{S}_k(d) + \mathbf{Z}_k, \tag{23}$$

where $\mathbf{Z}_k$ is independent of $\mathcal{D}_k$ and $\mathbf{S}_k(d)$. Therefore,

$$\mathcal{D}_k \to \mathbf{S}_k(d) \to \mathbf{T}_k(d) \tag{24}$$

is a Markov chain. By the chain rule for mutual information,

$$I(\mathcal{D}_k; \mathbf{T}_k(d)) = I(\mathbf{S}_k(d); \mathbf{T}_k(d)) + I(\mathcal{D}_k; \mathbf{T}_k(d) \mid \mathbf{S}_k(d)). \tag{25}$$

The conditional mutual-information term vanishes by the Markov property, giving

$$I(\mathcal{D}_k; \mathbf{T}_k(d)) = I(\mathbf{S}_k(d); \mathbf{T}_k(d)). \tag{26}$$

For the more general statement, let $\mathbf{O}_k$ be a richer observation from which $\mathbf{T}_k(d)$ is generated. Then

$$\mathcal{D}_k \to \mathbf{O}_k \to \mathbf{T}_k(d) \tag{27}$$

and data processing gives

$$I(\mathcal{D}_k; \mathbf{O}_k) \geq I(\mathcal{D}_k; \mathbf{T}_k(d)) \geq I(\mathbf{S}_k(d); \mathbf{T}_k(d)), \tag{28}$$

where the last inequality follows because $\mathbf{S}_k(d)$ is a deterministic function of $\mathcal{D}_k$. □

B.2 Proof of Proposition 2

Proof. By assumption,

$$\mathbf{S}_k(d) = \mathbf{U}_{k,d}\mathbf{c}_{k,d}, \tag{29}$$

where $\mathbf{U}_{k,d}$ has orthonormal columns and

$$\mathrm{Cov}(\mathbf{c}_{k,d}) = \mathrm{diag}\big(\mu_{k,1}(d), \ldots, \mu_{k,r_k(d)}(d)\big). \tag{30}$$

Therefore,

$$\mathbf{A}_k(d) = \mathrm{Cov}(\mathbf{S}_k(d)) = \mathbf{U}_{k,d}\,\mathrm{Cov}(\mathbf{c}_{k,d})\,\mathbf{U}_{k,d}^\top. \tag{31}$$

Because multiplication by a matrix with orthonormal columns preserves the nonzero spectrum, the nonzero eigenvalues of $\mathbf{A}_k(d)$ are exactly

$$\mu_{k,1}(d), \ldots, \mu_{k,r_k(d)}(d). \tag{32}$$

Since

$$r_k(d) = \alpha_k m_k d \tag{33}$$

and

$$\frac{a_k}{\log d} \leq \mu_{k,i}(d) \leq \frac{b_k}{\log d}, \tag{34}$$

Assumption 3 holds, for example with $c_{1,k} = c_{2,k} = \alpha_k$. □

B.3 Proof of Proposition 3

Proof. By condition (i), the population residual covariance $\mathbf{R}(d)$ has rank

$$\rho(d) = \Theta(d). \tag{35}$$

Thus there exist constants $0 < \underline{c} \leq \bar{c} < \infty$ such that

$$\underline{c}\, d \leq \rho(d) \leq \bar{c}\, d \tag{36}$$

for all sufficiently large d.

By condition (iii), for each client k, the nonzero spectrum of $\mathbf{A}_k(d)$ consists, up to constant-factor truncation, of m_k trajectory-level copies of the population residual spectrum. Therefore,

$$r_k(d) = \Theta(m_k \rho(d)) = \Theta(m_k d). \tag{37}$$

This gives the required rank-growth component of Assumption 3.

Next, condition (ii) states that the nonzero eigenvalues of $\mathbf{R}(d)$ satisfy

$$\nu_j(d) = \Theta(1/\log d), \qquad j = 1, \ldots, \rho(d). \tag{38}$$

By condition (iii), the corresponding eigenvalues of $\mathbf{A}_k(d)$ satisfy

$$\lambda_{k,\ell,j}(d) = \Theta(\nu_j(d)) = \Theta(1/\log d). \tag{39}$$

Hence there exist client-specific constants $0 < a_k \leq b_k < \infty$ such that

$$\frac{a_k}{\log d} \leq \lambda_{k,i}(d) \leq \frac{b_k}{\log d} \tag{40}$$

for all nonzero eigenvalues of $\mathbf{A}_k(d)$ and all sufficiently large d. Combining the rank and eigenvalue bounds proves Assumption 3. □

B.4 Proof of Theorem 1

Proof. Under Assumptions 4 and 5,

$$\mathbf{T}_k(d) = \mathbf{S}_k(d) + \mathbf{Z}_k, \qquad \mathbf{Z}_k \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{m_k d}), \tag{41}$$

with $\mathbf{Z}_k$ independent of $\mathbf{S}_k(d)$. Since $\mathbf{S}_k(d)$ is Gaussian with covariance $\mathbf{A}_k(d)$,

$$I(\mathbf{S}_k(d); \mathbf{T}_k(d)) = \frac{1}{2} \log \det \left(\mathbf{I} + \frac{\mathbf{A}_k(d)}{\sigma^2} \right). \tag{42}$$

Let the nonzero eigenvalues of $\mathbf{A}_k(d)$ be

$$\lambda_{k,1}(d), \ldots, \lambda_{k,r_k(d)}(d). \tag{43}$$

Then

$$\mathcal{L}_k(d) = \frac{1}{2} \sum_{i=1}^{r_k(d)} \log \left(1 + \frac{\lambda_{k,i}(d)}{\sigma^2} \right). \tag{44}$$

By Assumption 3,

$$\frac{a_k}{\log d} \le \lambda_{k,i}(d) \le \frac{b_k}{\log d}, \qquad i = 1, \ldots, r_k(d). \tag{45}$$

Hence

$$0 \le \frac{\lambda_{k,i}(d)}{\sigma^2} \le \frac{b_k}{\sigma^2 \log d}, \tag{46}$$

so for sufficiently large d we have $\lambda_{k,i}(d)/\sigma^2 \le 1$ for all i. Applying (21) termwise with $x = \lambda_{k,i}(d)/\sigma^2$ gives

$$\frac{1}{4\sigma^2} \sum_{i=1}^{r_k(d)} \lambda_{k,i}(d) \le \mathcal{L}_k(d) \le \frac{1}{2\sigma^2} \sum_{i=1}^{r_k(d)} \lambda_{k,i}(d). \tag{47}$$

By Assumption 3,

$$c_{1,k} m_k d \le r_k(d) \le c_{2,k} m_k d, \qquad \frac{a_k}{\log d} \le \lambda_{k,i}(d) \le \frac{b_k}{\log d}. \tag{48}$$

Therefore,

$$\frac{a_k c_{1,k} m_k d}{\log d} \le \sum_{i=1}^{r_k(d)} \lambda_{k,i}(d) \le \frac{b_k c_{2,k} m_k d}{\log d}, \tag{49}$$

which implies

$$\sum_{i=1}^{r_k(d)} \lambda_{k,i}(d) = \Theta\left(\frac{m_k d}{\log d} \right). \tag{50}$$

Substituting into the mutual-information bounds yields

$$\mathcal{L}_k(d) = \Theta\left(\frac{m_k d}{\sigma^2 \log d} \right). \tag{51}$$

Dividing by m_k gives

$$\frac{\mathcal{L}_k(d)}{m_k} = \Theta\left(\frac{d}{\sigma^2 \log d} \right). \tag{52}$$

□

Remark on local stochasticity. The signal variable $\mathbf{S}_k(d) = \mathbb{E}[\mathbf{Y}_k(d) \mid \mathcal{D}_k]$ integrates out mini-batch sampling and other local optimization randomness. Therefore, the covariance $\mathbf{A}_k(d) = \mathrm{Cov}(\mathbf{S}_k(d))$ should not be interpreted as being directly contracted by batch size. Batch-size effects are evaluated empirically as protocol-level effects on attack observables rather than as a theorem-level contraction of the client-specific signal covariance.

B.5 Proof of Theorem 2

Proof. Let the deployed noise level be fixed at $\sigma_0 > 0$. By Theorem 1,

$$\mathcal{L}_k(d;\sigma_0) = \Theta\left(\frac{m_k d}{\sigma_0^2 \log d}\right), \tag{53}$$

and at the reference scale,

$$\mathcal{L}_k(d_0;\sigma_0) = \Theta\left(\frac{m_k d_0}{\sigma_0^2 \log d_0}\right). \tag{54}$$

Taking the ratio gives

$$D_{\mathcal{L}}(d;d_0) = \frac{\mathcal{L}_k(d;\sigma_0)}{\mathcal{L}_k(d_0;\sigma_0)} = \Theta\left(\frac{d\log d_0}{d_0 \log d}\right). \tag{55}$$

The square-root leakage index is defined as

$$\varepsilon_{\mathrm{debt}}(d) \triangleq \varepsilon_{\mathrm{target}}\sqrt{D_{\mathcal{L}}(d;d_0)}. \tag{56}$$

Substituting the previous display yields

$$\varepsilon_{\mathrm{debt}}(d) = \Theta\left(\varepsilon_{\mathrm{target}}\sqrt{\frac{d\log d_0}{d_0 \log d}}\right). \tag{57}$$

This construction does not equate the large-scale leakage to a reference-scale channel with vanishing noise. It is only a leakage-ratio index and therefore avoids the high-SNR mismatch that would arise from interpreting $\varepsilon_{\mathrm{debt}}(d)$ as an exact DP-equivalent budget. □

B.6 Proof of Theorem 3

Proof. To preserve the reference leakage regime up to constant factors, it is sufficient to require

$$\frac{m_k d}{\sigma_{\mathrm{SA}}(d)^2 \log d} \asymp \frac{m_k d_0}{\sigma_0^2 \log d_0}. \tag{58}$$

Canceling m_k and solving for $\sigma_{\mathrm{SA}}(d)$ gives

$$\sigma_{\mathrm{SA}}(d)^2 = \sigma_0^2 \frac{d\log d_0}{d_0 \log d}, \tag{59}$$

or

$$\sigma_{\mathrm{SA}}(d) = \sigma_0\sqrt{\frac{d\log d_0}{d_0 \log d}}. \tag{60}$$

Applying Theorem 1 with this schedule gives

$$\mathcal{L}_k(d;\sigma_{\mathrm{SA}}(d)) = \Theta(\mathcal{L}_k(d_0;\sigma_0)). \tag{61}$$

If $\sigma_0 = \Theta(1/\varepsilon_{\mathrm{target}})$ under the fixed reference calibration, then

$$\sigma_{\mathrm{SA}}(d) = \Theta\left(\sqrt{\frac{d}{\varepsilon_{\mathrm{target}}^2 \log d}}\right). \tag{62}$$

It remains to show the insufficiency of any asymptotically smaller schedule. Suppose

$$\sigma(d) = o\left(\sqrt{\frac{d}{\log d}}\right). \tag{63}$$

Let $x_d = a_k/(\sigma(d)^2 \log d)$, where a_k is the lower spectral constant from Assumption 3. Since $\lambda_{k,i}(d) \geq a_k/\log d$ and $r_k(d) \geq c_{1,k} m_k d$, the exact log-det expression gives

$$\mathcal{L}_k(d;\sigma(d)) \geq \frac{c_{1,k} m_k d}{2}\log(1+x_d). \tag{64}$$

Using the elementary bound

$$\log(1+x) \geq \min\{x/2, \log 2\}, \qquad x \geq 0, \tag{65}$$

there are two cases. If $x_d \leq 1$ eventually, then

$$\mathcal{L}_k(d; \sigma(d)) \geq \frac{c_{1,k} m_k d}{4} \frac{a_k}{\sigma(d)^2 \log d} \to \infty \tag{66}$$

because $d/(\sigma(d)^2 \log d) \to \infty$. If $x_d > 1$ infinitely often, then along that subsequence

$$\mathcal{L}_k(d; \sigma(d)) \geq \frac{c_{1,k} m_k d}{2} \log 2 \to \infty. \tag{67}$$

Thus, any $\sigma(d) = o(\sqrt{d/\log d})$ fails to keep the leakage bounded relative to the fixed reference leakage regime. □

B.7 Proof of Corollary 1

Proof. Under the Gaussian mechanism,

$$\mathbf{Z}_k \sim \mathcal{N}(0, \sigma(d)^2 \mathbf{I}_{m_k d}), \tag{68}$$

so

$$\mathbb{E}\big[\|\mathbf{Z}_k\|_2^2\big] = m_k d\, \sigma(d)^2. \tag{69}$$

Applying Theorem 3,

$$\sigma(d)^2 = \Theta\left(\frac{d}{\varepsilon_{\text{target}}^2 \log d}\right), \tag{70}$$

which yields

$$\mathbb{E}\big[\|\mathbf{Z}_k\|_2^2\big] = \Theta\left(\frac{m_k d^2}{\varepsilon_{\text{target}}^2 \log d}\right). \tag{71}$$

Dividing by $m_k d$ gives the per-dimension scaling

$$\Theta\left(\frac{d}{\varepsilon_{\text{target}}^2 \log d}\right). \tag{72}$$

□

C Experimental and Reproducibility Details

This appendix documents the implementation details needed to reproduce the experimental study. It covers dataset and model configurations, training protocols, privacy mechanisms, privacy accounting, scale-aware noise instantiation, evaluation metrics, attack protocols, effective-dimension estimation, spectral diagnostics, statistical procedures, and compute resources. The organization follows the experimental pipeline: we first describe the datasets and model ladders, then the privacy mechanisms and accountant, then the evaluation metrics and attacks, and finally the signal-level spectral estimators used to connect the experiments to the theory.

C.1 Datasets, Model Ladders, and Training Protocols

The main paper reports the benchmark roles and core optimization settings needed to interpret the results. Table 5 summarizes the dataset and model ladders, and Table 6 reports the training and privacy settings for the evaluated mechanisms. Effective dimension d is the operative scaling variable throughout.

Modality-specific training horizons. The language ladder uses $T = 200$ rounds with $E = 1$ local epoch, whereas vision uses $T = 500$ rounds with $E = 5$. This reflects faster per-round convergence of pretrained language models under fine-tuning: in pilot runs, GPT-2 models approached near-final perplexity within roughly 100–150 rounds, while CIFAR-100 ResNets trained from scratch required more than 400 rounds for comparable stabilization. The spectral diagnostics also stabilized before round 150 in the language setting, with negligible change in the estimated effective dimension between rounds 150 and 200. These checks suggest that the shorter language horizon does not confound the scaling study.

Table 5: Model configurations.

Modality	Dataset role	Model	Approx. Params.	Eff. dim. d	Architecture notes
Vision	Primary: CIFAR-100 Supporting: FEMNIST	ResNet-20	≈ 0.27M	217	$n=3$ blocks/stage; He et al. (2016), option-A shortcuts
		ResNet-32	≈ 0.47M	304	$n=5$ blocks/stage
		ResNet-44	≈ 0.66M	397	$n=7$ blocks/stage
		ResNet-56	≈ 0.86M	531	$n=9$ blocks/stage
		ResNet-110	≈ 1.73M	748	$n=18$ blocks/stage
		WRN-28-10	≈ 36.5M	1162	Zagoruyko & Komodakis (2016), dropout 0.3
Language	Primary: Stack Overflow (top 2,000 users) Supporting: Shakespeare	GPT-2 Small	≈ 117M	186	Base causal LM; full fine-tuning
		GPT-2 Medium	≈ 345M	341	LoRA ($r=8$, $\alpha_{\mathrm{LoRA}}=16$)
		GPT-2 Large	≈ 774M	563	LoRA ($r=8$, $\alpha_{\mathrm{LoRA}}=16$)

Note. Parameter counts are approximate. Effective dimensions are measured using the retained-variance estimator at the default threshold $d = d_{0.99}$. The six-point vision ladder provides the primary resolution for distinguishing constant, linear, and $d/\log d$ trends, while the three-point language ladder provides corroborative cross-modality evidence.

Table 6: Training and privacy hyperparameters for the evaluated mechanisms. A dash indicates that the corresponding parameter is not applicable.

Parameter	Non-private FL	Fixed-noise DP-FL	Adaptive-clipping DP-FL	Scale-aware DP-FL
Target privacy budget (ε, δ)	–	$(3.0, 10^{-5})$	$(3.0, 10^{-5})$	$(3.0, 10^{-5})$ reference target
Noise scale	–	σ_0	σ_0	$\sigma_{\mathrm{SA}}(d)$
Noise schedule	–	constant across scale	constant across scale	from Theorem 3
Clipping threshold C	–	1.0 (fixed)	adaptive (median quantile)	1.0 (fixed)
Privacy accountant	–	Rényi DP	Rényi DP	Rényi DP
Vision (CIFAR-100, FEMNIST)				
Communication rounds	500	500	500	500
Client sampling rate q	0.1	0.1	0.1	0.1
Local epochs E	5	5	5	5
Optimizer	SGD	SGD	SGD	SGD
Learning rate	0.05 (0.03 for WRN)	0.05 (0.03 for WRN)	0.05 (0.03 for WRN)	0.05 (0.03 for WRN)
Momentum	0.9	0.9	0.9	0.9
Batch size	32	32	32	32
Language (Stack Overflow subset, Shakespeare)				
Communication rounds	200	200	200	200
Client sampling rate q	0.05	0.05	0.05	0.05
Client sampling method	Poisson	Poisson	Poisson	Poisson
Local epochs E	1	1	1	1
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	5×10^{-5} (S); 10^{-4} (M/L)	5×10^{-5} (S); 10^{-4} (M/L)	5×10^{-5} (S); 10^{-4} (M/L)	5×10^{-5} (S); 10^{-4} (M/L)
Weight decay	0.01	0.01	0.01	0.01
Batch size	4	4	4	4
Sequence length	128	128	128	128
Shared evaluation settings				
Effective-dimension estimator	smallest retained rank explaining 99% of empirical update-spectrum energy			
Vision utility metric	top-1 test accuracy			
Language utility metric	perplexity			
Primary privacy metric	membership inference AUC			
Auxiliary privacy metric	client-specific canary extraction rate (language only)			

Note. Vision results use 5 seeds for all ResNet variants (20, 32, 44, 56, 110), 3 seeds for WRN-28-10, and 3 seeds for all language models. The scale-aware mechanism uses the same reference accountant target but selects a dimension-dependent noise multiplier according to the signal-channel calibration rule.

Controls for scale-ladder confounds. The primary functional-form evidence comes from the six-point vision ladder. Within this ladder, ResNet-20 through ResNet-110 provide the most homogeneous architectural sequence, while WRN-28-10 is included as a larger stress point with the standard WRN dropout configuration and a slightly different learning rate. Therefore, conclusions about monotone privacy-risk growth do not rely solely on WRN-28-10; the ResNet-only trend is reported and interpreted separately from the full ladder where appropriate. On the language side, GPT-2 Small uses full fine-tuning, whereas GPT-2 Medium and Large use LoRA. We therefore interpret language results primarily through measured effective dimension d, with supporting matched-effective-dimension FT-vs.-LoRA evidence reported in Table 8. The FT-vs.-LoRA check compares configurations with comparable measured $d_{0.99}$ rather than comparable raw trainable parameter count.

C.2 Privacy Mechanisms, Accounting, and Scale-Aware Instantiation

Privacy composition and scale-aware noise instantiation. The scale-aware schedule is instantiated by anchoring the constant at the smallest reference model. Let d_0 denote the reference effective dimension and let σ_0 be the noise multiplier calibrated at that scale for the target accountant setting. For every larger scale, we use

$$\sigma_{\mathrm{SA}}(d) = \sigma_0 \sqrt{\frac{d \log d_0}{d_0 \log d}}. \tag{73}$$

The schedule is anchored at (d_0, σ_0); in the evaluated regime $d \geq d_0 > e$, it satisfies $\sigma_{\mathrm{SA}}(d) \geq \sigma_0$. Under the same (q, T, C, δ) accountant configuration, the corresponding accountant-reported ε is therefore no larger than the reference target. Thus, the scale-aware comparison should be interpreted as signal-level stabilization, not as an equal-noise comparison. Within each training run, $\sigma_{\mathrm{SA}}(d)$ is fixed before training begins, so the RDP accountant composes T applications of a subsampled Gaussian mechanism with a single fixed noise multiplier, as in standard DP-FedAvg. The schedule changes the selected noise multiplier across model scales, but not across rounds within a given run.

In the reported experiments, d is estimated from a short preliminary run. If this preliminary run is performed on the private federated training data without privacy protection, then the resulting end-to-end procedure should be interpreted as an empirical validation of scale-aware calibration rather than as a strict end-to-end DP guarantee including the estimator. A fully private deployment can instead estimate d using a disjoint public proxy dataset or a DP-accounted pilot phase, after which the selected $\sigma_{\mathrm{SA}}(d)$ can be composed with the remaining training rounds using standard RDP accounting. In our pilot runs, the effective dimension measured at round 50 differed from the full-training estimate by less than 8% across the evaluated models, supporting the use of a fixed pre-training estimate within each run.

Protected pilot feasibility check. Separately from the preliminary estimates used in the main study, the experiment estimated d from privatized client-round outputs using σ_0. In vision, 150 pilot rounds and 1,500 client-round observations produced dimension and selected-multiplier errors of 5.6% and 2.5%; in language, 20 rounds and an expected 2,000 observations gave 3.2% and 1.5%. The reported estimator overhead was 41.8 s / 1.3 GB and 27.6 s / 0.9 GB, respectively. The pilot rounds must be included in the privacy composition; selecting a multiplier from the privatized output is post-processing. Since the selected multipliers are floored at σ_0, accounting for all rounds at σ_0 is conservative under the stated mechanism. A coordinator observing only secure aggregates would require a secure or otherwise authorized covariance estimator. These results are a feasibility check, not evidence that the main-study preliminary run was DP-accounted.

RDP accountant validation. We implement an RDP accountant following [11] with the subsampled-Gaussian analysis of [46]. As a validation check, we cross-reference the calibrated noise multiplier σ_0 against the `dp-accounting` library[3] for the same (q, σ, T, δ) configuration. In the validation runs, the two implementations agreed to within 0.5% on the reported ε, supporting the correctness of the accountant used in the experiments.

C.3 Evaluation Metrics, Attack Protocols, and Statistical Testing

Privacy is evaluated primarily using membership inference AUC (MIA AUC) and, for language, canary extraction rate [36]. Attack metrics measure downstream exploitability, while the trajectory-level proxy in Appendix C.4 provides a signal-level diagnostic aligned with the theoretical channel model. We also report attack advantage and exposure as supplementary diagnostics. Where used, pairwise comparisons rely on two-sided paired t-tests across matched seeds, and candidate-family comparisons use a shared AICc-based pipeline. Given the finite ladders in the primary study, these fits are intended to characterize finite-scale behavior within the tested candidate families rather than establish asymptotic identification.

Membership inference. For vision, the evaluation pool is split into member examples assigned to at least one client during FL training and non-member examples that are either unused by any client or, when the partition covers all training examples, drawn from a consistently carved-out held-out subset.

[3] `https://github.com/google/differential-privacy/tree/main/python/dp_accounting`

For language, training posts from participating users serve as members, and held-out test posts from non-overlapping users serve as non-members. The attack uses per-example loss under the trained model as the membership score, with lower loss indicating higher membership likelihood. We use MIA AUC as the primary membership-inference metric and additionally compute attack advantage $\max |TPR - FPR|$, each over 2,048 member and 2,048 non-member examples per evaluation.

Canary-based extraction. For the language setting, extraction risk is measured following the secret-sharer protocol of [36]. We insert 100 canary sequences, each consisting of a fixed prefix ("The secret key is") and a six-token random suffix, into a small subset of clients' training data, with each canary repeated four times within its assigned client. After training, extraction rate is measured as the fraction of canaries whose suffix is reproduced verbatim by greedy decoding given the prefix. We also compute exposure on a random subset of 20 canaries against a candidate set of 1,024 random alternatives.

Utility retention normalization. When plotting normalized utility retention across modalities, we map each utility metric so that larger values always indicate better utility. For accuracy-based vision tasks, the retention of mechanism m at scale s is

$$R_{\text{util}}^{\text{vision}}(m, s) = \frac{\text{Acc}(m, s)}{\text{Acc}(\text{NonPrivate}, s)}. \tag{74}$$

For perplexity-based language tasks, where lower is better, retention is

$$R_{\text{util}}^{\text{lang}}(m, s) = \frac{\text{PPL}(\text{NonPrivate}, s)}{\text{PPL}(m, s)}. \tag{75}$$

Thus $R_{\text{util}} = 1$ matches non-private utility at the same scale, values below one indicate degradation, and values above one indicate improvement relative to the non-private baseline.

C.4 Effective-Dimension and Spectral Estimation

Effective-dimension estimator. Effective dimension is estimated from the retained energy of the empirical update spectrum. For a retained-variance threshold τ, d_τ is the smallest rank whose leading empirical update-spectrum directions explain at least fraction τ of the total spectrum energy. Unless otherwise stated, we use $\tau = 0.99$ and write $d = d_{0.99}$. Sensitivity to nearby thresholds is reported in Table 8.

Spectral estimation. The spectral validation experiments test whether the evaluated model ladders lie in a regime compatible with Assumption 3. For each scale point, we collect effective update vectors from client-round observations under the same federated protocol as in the main experiments. Let

$$\mathcal{E} \triangleq \{(k, t) : k \in \mathcal{S}_t,\ t = 1, \ldots, T\} \tag{76}$$

denote the set of collected client-round update samples, and let $N = |\mathcal{E}|$. We estimate an empirical update-spectrum proxy by

$$\widehat{\mathbf{A}}_{\text{upd}}(d) \triangleq \frac{1}{N-1} \sum_{(k,t)\in\mathcal{E}} \left(\Delta_k^{(t)} - \bar{\Delta}\right)\left(\Delta_k^{(t)} - \bar{\Delta}\right)^\top, \tag{77}$$

where

$$\bar{\Delta} \triangleq \frac{1}{N} \sum_{(k,t)\in\mathcal{E}} \Delta_k^{(t)}. \tag{78}$$

This estimator is not a literal finite-K estimator of a single-client covariance $\mathbf{A}_k(d)$; rather, it is an empirical client-round update-spectrum proxy used to assess whether the observed update space exhibits rank growth, eigenvalue decay, and trace behavior compatible with the analytical spectral regime. Because the covariance is computed over client-round samples rather than only over K client means, its rank is bounded by $N - 1$ rather than by $K - 1$, avoiding the finite-client rank limitation.

At large scales, we use Johnson–Lindenstrauss projection into a 4,096–16,384 dimensional subspace followed by randomized SVD [47]; updates are stored using reservoir sampling with capacity 256–512

to bound memory. We summarize the estimated spectrum through

$$\widehat{r}(d) \triangleq \frac{\mathrm{tr}(\widehat{\mathbf{A}}_{\mathrm{upd}}(d))^2}{\mathrm{tr}(\widehat{\mathbf{A}}_{\mathrm{upd}}(d)^2)}, \tag{79}$$

the mean informative eigenvalue

$$\bar{\lambda}(d) \triangleq \frac{\mathrm{tr}(\widehat{\mathbf{A}}_{\mathrm{upd}}(d))}{\widehat{r}(d)}, \tag{80}$$

and total signal power $\mathrm{tr}(\widehat{\mathbf{A}}_{\mathrm{upd}}(d))$. Candidate-form comparisons use the same least-squares and AICc pipeline across modalities and settings.

Trajectory-level proxy. To complement attack-based metrics, we use the Gaussian-channel mutual-information proxy

$$\widehat{\mathcal{L}}_{\mathrm{proxy}}(d) = \frac{1}{2}\sum_i \log\left(1 + \frac{\widehat{\lambda}_i(d)}{\sigma^2}\right), \tag{81}$$

where $\widehat{\lambda}_i(d)$ are the eigenvalues of the empirical client-round update-spectrum proxy $\widehat{\mathbf{A}}_{\mathrm{upd}}(d)$ in Eq. (77). This trajectory-level proxy is a signal-level observable aligned with Section 4; it is not a replacement for attack-based metrics and is not an accountant-reported DP quantity.

C.5 Compute Resources

All experiments were conducted on a server equipped with 4×NVIDIA A100 (80 GB) GPUs, a 64-core AMD EPYC CPU, and 512 GB RAM. Most vision and language runs used one GPU per run, while the largest language setting used up to two GPUs depending on memory availability. Individual training runs ranged from a few GPU-hours for smaller ResNet and GPT-2 configurations to roughly half a day for the largest WRN and GPT-2 Large settings. Code, configuration files, random seeds, and model-specific hyperparameters are included in the released artifact.

D Extended Empirical Results

D.1 Finite-Scale Candidate-Fit Sensitivity

The comparison uses the six-point vision ladder and a ResNet-only subset to check whether the fit is driven by WRN-28-10. For constant, linear-in-d, $d/\log d$, and fitted d^β candidates, the respective six-point ΔAICc values are 15.1, 3.6, 0, and 1.5; ResNet-only ΔAICc values are 11.4, 3.1, 0, and 1.3. Their leave-one-out RMSE values are 0.046, 0.018, 0.013, and 0.014, and their bootstrap preference shares are 3%, 20%, 49%, and 28%. The fitted power exponent is $\beta = 0.78$ with bootstrap interval $[0.61, 0.96]$. The evidence favors a sublinear finite-scale trend and does not uniquely identify $d/\log d$ over a nearby power law.

D.2 Additional Noise-Matched and Utility-Matched Controls

The uniform $\sigma_{\max}$ comparison in Table 7 shares the same multiplier as scale-aware calibration at WRN-28-10 by construction and gives lower MIA at smaller scales at additional utility cost. In a separate architecture-specific search matching scale-aware accuracy within approximately 0.3 percentage points, ResNet-56 attains 70.5% accuracy, 0.558 MIA AUC, and 2.7% TPR@1% FPR, versus 70.4%, 0.553, and 2.6% for scale-aware. WRN-28-10 attains 73.5%, 0.570, and 3.1%, versus 73.4%, 0.566, and 3.0%. These close outcomes are expected for the same Gaussian mechanism at similar noise levels; the calibration rule's contribution is selecting multipliers across scales without tuning to attack outcomes.

Table 7: Additional vision control: top-1 accuracy (%) / MIA AUC. Uniform $\sigma_{\max}$ is more conservative at smaller scales.

Model	Fixed σ_0	Uniform $\sigma_{\max}$	Scale-aware
ResNet-20	68.0 / 0.547	64.8 / 0.520	67.3 / 0.536
ResNet-56	71.7 / 0.618	68.1 / 0.537	70.4 / 0.553
WRN-28-10	76.5 / 0.704	73.4 / 0.566	73.4 / 0.566

D.3 Clipping Diagnostics

The measured spectrum uses updates after L_2 clipping at $C = 1.0$ and before Gaussian perturbation and aggregation. The diagnostics report raw update-norm P50/P95, clipping rate, and pre/post-clipping retained dimension as follows: ResNet-20, $0.61/1.27$, 17.8%, $236/217$; ResNet-56, $0.77/1.63$, 30.6%, $574/531$; WRN-28-10, $0.94/2.08$, 46.9%, $1284/1162$. The post-clipping dimension still rises over the evaluated range, even though clipping increasingly limits signal energy. This is a finite-scale diagnostic and does not validate Assumption 3 asymptotically.

D.4 Comparison with Per-scale Accountant Recalibration

A natural practical question is whether recalibrating the standard RDP accountant independently at each model scale for the same (ε, δ) target removes the observed privacy drift. In our setting, the accountant-level inputs q, T, δ, and the clipping configuration are held fixed across scales. Therefore, per-scale accountant recalibration selects essentially the same fixed-noise configuration across the ladder and does not introduce explicit dependence on effective dimension d. This preserves the accountant-reported DP target but does not control the signal-level leakage ratio studied in Theorem 2. Empirically, per-scale accountant recalibration exhibits privacy trends similar to fixed-noise deployment on the primary ladders, indicating that accountant-level recalibration alone does not remove the observed cross-scale attack drift. We avoid reducing this comparison to a single closed-form "effective update noise" expression because the precise standard deviation applied to the aggregated model update depends on implementation details such as client normalization, weighting, and whether noise is added before or after averaging. The important point for this paper is that standard accountant recalibration does not by itself introduce the $\sqrt{d/\log d}$ dependence required by the signal-level leakage analysis.

D.5 Robustness, Sensitivity, and Ablation-Style Controls

We report supplementary robustness and sensitivity analyses testing whether the main privacy-scaling effect is stable under heterogeneity, optimization, participation rate, adaptation style, and the effective-dimension estimator. Table 8 collects representative robustness results. Overall, the phenomenon remains stable across these variations, while its magnitude changes in ways consistent with the signal-level interpretation. For heterogeneity, the primary vision-ladder fit remains strong across $\alpha \in \{0.1, 0.5, 1.0\}$, with $R^2 = 0.96$, 0.92, and 0.89, respectively, indicating steeper privacy-risk growth under stronger non-IID partitions. Increasing local batch size reduces attack-based privacy risk in the empirical protocol, but this is reported as an observed protocol-level effect rather than as a theorem-level contraction of $\mathrm{Cov}(\mathbf{S}_k(d))$. Varying participation rate changes magnitude more than functional form, with fitted slope changes below 12%, while the $d/\log d$ family remains preferred within the tested candidate-family set. At GPT-2 Large, full fine-tuning and LoRA differ only marginally at matched measured effective dimension, with $\Delta\mathrm{MIA} = 0.01$, extraction difference $= 0.3$ points, and paired $p > 0.10$. Finally, varying the retained-variance threshold over $\tau \in \{0.95, 0.97, 0.99, 0.995\}$ changes absolute d_τ values but leaves model ordering, fit preference, privacy-debt trends, and mechanism rankings unchanged.

D.6 Diagnostics Linking the Signal-Level Theory to Attack Metrics

This subsection reports supplementary diagnostics linking the signal-level theory in Section 4 to the attack-based observables used in the experiments. Table 9 summarizes four complementary checks. First, the trajectory-level proxy tracks attack metrics closely across the primary ladders, with Spearman correlations of $\rho = 0.88$ for vision MIA, $\rho = 0.91$ for language MIA, and $\rho = 0.89$ for

Table 8: Supplementary robustness summary across optimization and protocol factors.

Ablation factor	Setting	Observed change	Representative evidence
Heterogeneity	CIFAR-100 vision setting, $\alpha \in \{0.1, 0.5, 1.0\}$; pathological split	stronger non-IID partitions yield steeper privacy-risk growth on the vision ladder	$R^2 = 0.96$ ($\alpha = 0.1$), 0.92 ($\alpha = 0.5$), 0.89 ($\alpha = 1.0$)
Batch size	primary vision setting, $B \in \{16, 32, 64, 128, 256\}$	privacy risk decreases with larger local batch size	3.4 ± 0.4 percentage-point reduction in MIA AUC from $B = 16$ to $B = 256$ at ResNet-56
Participation rate	primary vision setting, $q \in \{0.05, 0.10, 0.20\}$	magnitude changes more than functional form	fitted slope changes by $< 12\%$; $d/\log d$ remains the strongest fit within the tested candidate-family set
FT vs. LoRA	GPT-2 Large at matched d	negligible difference in the main language-side privacy observables	ΔMIA $= 0.01$, extraction difference $= 0.3$ points, paired $p > 0.10$
Threshold sensitivity	varying retained-variance threshold τ	absolute d changes but conclusions remain stable	model ordering unchanged; preferred fit family unchanged for $\tau \in \{0.95, 0.97, 0.99, 0.995\}$

Note. Results are setting-specific and supplementary; their role is to confirm qualitative stability of the main conclusions across alternative configurations.

language extraction. Second, the low-SNR approximation remains accurate in the tested regime, with mean relative error 6.8%. Third, proxy-based and attack-based mechanism rankings agree at 5 of the 6 scale points on the primary vision ladder. Taken together, these diagnostics support structural alignment between the signal-level theory and the attack-based observables, without claiming numerical equivalence.

Table 9: Supplementary diagnostics linking the theoretical channel model to attack-based observables on the primary ladders.

Diagnostic	Evaluation focus	Observed outcome
Proxy–MIA association	monotonic agreement between the trajectory-level proxy and MIA across the primary ladders	Spearman $\rho = 0.88$ (vision), 0.91 (language)
Proxy–extraction association	consistency between the trajectory-level proxy and language-side extraction risk	Spearman $\rho = 0.89$
Mechanism-order agreement	agreement between proxy-based and attack-based mechanism ranking on the six-point vision ladder	identical ranking at 5/6 vision scale points
Low-SNR approximation	first-order accuracy of the channel expansion on the empirical spectrum	mean relative error $= 6.8\%$

D.7 Practical viability at very large scale.

The scale-aware rule deliberately increases noise with effective dimension, so it can impose substantial utility cost at very large scales. This is consistent with Corollary 1, which shows that total noise power grows as $\Theta(m_k d^2/(\varepsilon^2_{\text{target}} \log d))$ under the signal-channel calibration. The proposed rule should therefore be viewed as a principled calibration baseline and diagnostic tool rather than a universally optimal deployment recipe. Practical billion-parameter FL systems may require hybrid strategies, such as public-proxy estimation of d, layerwise calibration, structured noise, selective adaptation, or utility-aware privacy budgeting.

E Societal Impacts

This work studies how privacy risk changes as federated models scale and aims to improve the safe deployment of privacy-preserving FL in sensitive domains such as healthcare, finance, mobile systems, and edge intelligence. Its positive societal impact lies in helping practitioners avoid under-protective privacy configurations when larger models are deployed under fixed-noise mechanisms. At the same time, the paper also highlights potential risks: better understanding of privacy leakage mechanisms could inform adversaries about weak points in deployed systems, and stronger privacy stabilization may reduce utility if applied without regard to application context. More broadly, the paper encourages evaluating privacy, utility, and robustness jointly rather than treating privacy as a static configuration choice, thereby contributing to more transparent and responsible FL design.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the paper's core contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion and in the theoretical development, including the finite experimental scale range, the dependence on additive Gaussian mechanisms, and the reliance on a specific spectral-growth regime. These limitations are explicitly acknowledged and scoped.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the full assumptions for the theoretical results in the main text and provides complete proofs in the appendix. The main paper also includes proof sketches to provide intuition and cross-references the formal results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper details the implementations, datasets, and experimental settings sufficient to reproduce the results in Section 5 and in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a github repository link in Section 5 containing the code, evaluation environment, data-access instructions, and step-by-step reproduction instructions for the main results. The supplementary material and appendix also document the key implementation settings and evaluation protocol.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental settings needed to interpret the results, including datasets, model ladders, data partitions, optimization settings, privacy mechanisms, and evaluation protocol. Additional details are summarized in appendix tables and implementation subsections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports mean $\pm$ standard deviation across matched random seeds: five seeds for ResNet vision models and three seeds for WRN-28-10 and all language models. It also reports paired significance testing for selected mechanism comparisons, and

the appendix explains the statistical testing and fit-selection protocol used for the reported trends.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the Experimental section 5 and the Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that the research reported in this paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Please refer to Appendix E.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new high-risk generative model, scraped dataset, or similarly misuse-prone asset requiring special release safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original datasets, model families, and prior implementations used in this work are properly cited, and the relevant external assets are credited in the manuscript and supplemental material. The corresponding references and usage conditions are respected in the experimental study.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released artifact consists of code, configuration files, evaluation scripts, and reproduction instructions. The paper does not introduce a new dataset or pretrained model; all external datasets and model families are cited and documented separately.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, paid data collection, or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects requiring IRB approval or equivalent review.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method and experimental design do not use LLMs as an important, original, or non-standard methodological component. Any incidental use of LLM tools for writing or editing would not affect the scientific contribution and therefore does not require declaration.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.