FINAL-MODEL PRIVACY FOR PROXIMAL-ANCHORED DP-SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard Rényi differential privacy (RDP) composition charges a record at every access, so its bound grows with the number of fine-tuning updates even when only the final model is released. We study this hidden-state setting for clipped, projected proximal-anchored DP-SGD. Our shifted-Rényi PABI certificate separates the discrepancy created by an accessed record from the state difference propagated by later updates. Exact Gaussian smoothing gives dimension-free propagation factors, while a bounded correction at a fixed public reference state reduces the smoothing bias there without changing those factors. For any public access pattern, causal shift allocation computes the certificate in $O(T)$ time. If the certified factors contract on every contiguous interval, the final-model RDP bound is independent of T and has no separate domain-diameter term. We also give an adaptive anchor rule with a joint bound for its privatized monitor bits and final model. At 500 epochs, the certificate is 2.731, compared with 10.956 under step-wise composition.

1 INTRODUCTION

Fine-tuning pretrained models adapts them to downstream tasks but can leak downstream training data. For example, membership inference can determine whether a record was used for fine-tuning, and training-data extraction can recover memorized examples (Shokri et al., 2017; Carlini et al., 2021). Private fine-tuning must therefore preserve downstream utility while limiting what the released model reveals about individual records. Differential privacy (DP) formalizes this with an (ε, δ) guarantee. Gradient-based private fine-tuning commonly uses DP-SGD (Abadi et al., 2016; Yu et al., 2022), with an RDP accountant accumulating privacy cost across updates. If every update incurs a fixed positive cost, the RDP upper bound grows as $O(T)$ (Mironov, 2017). This creates a dilemma: more updates improve utility but linearly increase privacy loss. Under a fixed (ε, δ) budget, capping the $O(T)$ loss limits updates and utility; allowing more updates improves utility only by accepting a weaker privacy guarantee.

Privacy amplification by iteration (PABI) mitigates the $O(T)$ accumulation by hiding intermediate iterates and analyzing only the released final model (Feldman et al., 2018). Originally developed for contractive iterations, its one-epoch PNSGD analysis yields an $O(1/n)$ RDP bound that decreases with the number of remaining updates. For smooth convex losses on a bounded domain, Altschuler and Talwar obtain $O(n^{-2}\min\{T, \Theta(n)\})$, making the bound independent of T after a data-dependent burn-in (Altschuler & Talwar, 2022). Shifted interpolation further gives convergent f-DP bounds under strong convexity or bounded-domain convexity in full, cyclic, and stochastic batch settings (Bok et al., 2024). Later work extends convergent RDP bounds to bounded-domain nonconvex nonsmooth losses with Hölder-continuous gradients and modulus-contractive updates (Chien & Li, 2025; Bravo et al., 2025), while related hidden-state extensions do so for smooth zeroth-order optimization (Chien et al., 2026).

PABI fits private fine-tuning because only the final model is released. Its shifted-Rényi transport theorem requires a public bound L_t on how an update responds to a difference in hidden states. For full-model transformers, obtaining this quantity from a global Hessian envelope can multiply loose bounds across attention, normalization, and residual blocks. To make this requirement operational, we construct a public L_t-certification algorithm. The algorithm runs before private training, computes the bound from the declared update configuration, checks the propagation condition re-

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

quired by the PABI accountant, and records the values used by the subsequent privacy analysis. This public certification check separates privacy certification from private training trajectories and avoids placing an architecture-specific Hessian calculation inside the training loop.

For the clipped, projected proximal update, we study when repeated accesses to one record admit a horizon-independent bound. Existing analyses obtain such bounds using a bounded parameter domain, strong convexity, or other contraction conditions (Altschuler & Talwar, 2022; Bok et al., 2024; Chien & Li, 2025; Bravo et al., 2025). For nonconvex composite cyclic DP-SGD, Kong & Ribero (2025) give a domain-dependent bound on a bounded domain, while their unbounded-domain bound grows with the number of epochs. Here the public values L_t determine the hidden-state response, and a public update profile provides a sufficient interval-decay condition that permits individual expansive steps without an explicit diameter term in the final formula.

In this paper, we focus on the following question:

In multi-epoch fine-tuning, can we certify a T-independent privacy loss without an explicit domain-diameter term or global strong convexity?

Prior empirical studies compare leakage at different training iterates and report that a privacy attack can become stronger and later weaker as training continues (Carlini et al., 2019; Jagielski et al., 2023; Duan et al., 2024; Leybzon & Kervadec, 2024). This motivates a final-model analysis that keeps access order and subsequent state propagation visible instead of replacing the whole run by one average factor.

▷ For a fixed public batch schedule, each access to the changed record creates a per-step map discrepancy. We use shifted-Rényi PABI to propagate the remaining state difference and optimize the causal shift allocation before θ_T is released. The resulting certificate is no larger than immediate per-access charging and is computable in $O(T)$ time per public participation pattern, including profiles with zero propagation factors.

▷ We make the hidden-state condition verifiable for full-model fine-tuning with a public L_t-certification algorithm. The public certification check computes the bound before any private record is accessed and checks the propagation condition used by the PABI accountant, without requiring a global architecture-specific Hessian envelope. The resulting RDP expression has no explicit domain-diameter term, although the public update factors may depend on the declared parameter region.

▷ We test whether the public propagation factors decay over every training interval at a target rate $\gamma > 0$, up to a certified exponent B. This permits individual expansive updates while requiring net interval decay. With fixed noise calibration, the resulting RDP certificate is independent of T when $B = O(1)$ and γ stays positive.

▷ We implement the interval-decay test as a one-state online controller that raises the anchor strength before a violating update. Privatized Gaussian threshold tests of the clipped batch gradient at the public anchor provide data-dependent candidate strengths. A joint privacy bound accounts separately for the monitor bits and the final-model release and remains independent of T when the number of checkpoints is bounded.

Our contributions are as follows.

- We specialize shifted-Rényi PABI to clipped, projected proximal-anchored DP-SGD and optimize its causal Gaussian shift allocation. The resulting final-model certificate is no larger than immediate per-access charging and is computable in $O(T)$ time per public participation pattern, including zero propagation factors.
- We construct a public L_t-certification algorithm that computes the hidden-state response bound before private training and checks the propagation condition required by the PABI accountant. The public certification check takes $O(T)$ time and uses no private trajectory information.
- We show that a contiguous-interval product condition with exponent B and rate γ makes the RDP certificate independent of T when B and γ^{-1} are uniform over the run.

- We give an online anchor-selection rule that maintains the interval certificate and admits private threshold decisions at the public anchor. A uniform conditional final-model bound and adaptive RDP composition account jointly for the decisions and released model.

2 PRELIMINARIES

2.1 NOTATION

Let n be a public number of indexed training slots and write $D = (z_1, \dots, z_n) \in (\mathcal{Z} \cup \{\emptyset\})^n$, where $\emptyset$ denotes an absent record and $\ell(\theta; \emptyset) = 0$. We use add/remove adjacency in a fixed-slot representation.

Definition 1 (Neighboring datasets)**.** *Two datasets D, D' are neighboring, written $D \sim D'$, if they have the same public slots and differ at one index $i^\star$, where one entry is a record in $\mathcal{Z}$ and the other is $\emptyset$.*

The number of nonempty records may differ, but the slot count n does not. A model with parameters $\theta \in \mathbb{R}^d$ has per-example loss $\ell(\theta; z)$ and fixed-denominator empirical loss $L_D(\theta) = n^{-1} \sum_{i=1}^n \ell(\theta; z_i)$. We use $\| \cdot \|$ for the Euclidean norm. Training has $T \geq 1$ updates and public batches $S_t \subseteq [n]$ of size $b \in \mathbb{N}_{>0}$. Batch indices and the denominator b do not depend on which slots are empty. The public pretrained initialization is θ_0, and only the final model is released. For $R \geq 0$, vector clipping is $\mathrm{clip}_R(v) = v \min\{1, R/\|v\|\}$ with $\mathrm{clip}_R(0) = 0$.

Throughout the paper, a quantity is public when it is fixed before private training and is independent of the private records and the training noise. This includes the batch schedule, initialization, fixed rescaling, smoothing scales, reference states, and correction caps.

2.2 DIFFERENTIAL PRIVACY

A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if, for every neighboring pair $D \sim D'$ under the fixed-slot relation and every measurable set $\mathcal{A}$,

$$\Pr[\mathcal{M}(D) \in \mathcal{A}] \leq e^\varepsilon \Pr[\mathcal{M}(D') \in \mathcal{A}] + \delta. \tag{1}$$

For $\alpha > 1$, the Rényi divergence of P from Q is

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int \left(\frac{dP}{dQ}\right)^\alpha dQ. \tag{2}$$

A mechanism is (α, ρ)-RDP if $D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \leq \rho$ for every neighboring pair. RDP composes additively at a fixed order, and implies $\varepsilon \leq \rho + \log(1/\delta)/(\alpha - 1)$ for (ε, δ)-DP. DP-SGD clips per-example gradients and adds Gaussian noise to each update (Abadi et al., 2016; Yu et al., 2022).

3 RELATED WORK

3.1 PRIVACY AMPLIFICATION BY ITERATION

Privacy amplification by iteration (PABI) studies the setting in which the intermediate iterates remain hidden and only the final model θ_T is released. Later noisy updates can attenuate the effect of an earlier record access, so the final-model cost can be smaller than the sum of all per-step costs (Feldman et al., 2018). Repeated access changes the question: the same record can create a new map discrepancy at several times, and contraction before a later access cannot protect that later access.

Existing work gives convergent or horizon-independent bounds under different conditions. Smooth convex noisy SGD has bounded-domain and strongly convex results (Altschuler & Talwar, 2022); shifted interpolation covers full, cyclic, and stochastic batches (Bok et al., 2024). For nonconvex or nonsmooth losses, Chien and Li use Hölder continuity and forward Wasserstein tracking (Chien & Li, 2025). Bravo et al. study projected Langevin mixing under a modulus of continuity (Bravo et al., 2025), while Kong and Ribero analyze cyclic clipped DP-SGD and obtain an unbounded-domain

bound that grows with the number of epochs (Kong & Ribero, 2025). Hidden-state amplification has also been studied for smooth zeroth-order optimization (Chien et al., 2026).

Our contribution is to keep the repeated-access times explicit for a clipped, projected proximal update. We then make the Lipschitz premise for state propagation verifiable with exact Gaussian smoothing and derive a contiguous-interval product condition that is sufficient for a horizon-independent certificate.

3.2 Proximal-Anchored Optimization for Fine-Tuning

Proximal anchoring adds a quadratic penalty that keeps the adapted parameters near a public reference model. In our notation, the penalty is $\lambda_t\|\phi-\phi_0\|^2/2$, where ϕ_0 represents the public pretrained initialization. The term has a direct optimization role: it limits parameter drift when the downstream objective is noisy or repeatedly revisited, while leaving the batch schedule and the per-record clipping rule unchanged.

This idea is well established in related optimization settings. FedProx uses a proximal term to limit client drift in federated optimization (Li et al., 2020). In foundation-model fine-tuning, related regularization and reference-preserving strategies have been used to improve safety or reduce forgetting (Huang et al., 2024; Tian et al., 2024; Qiu et al., 2026). These works motivate the use of a public reference model, but they do not provide the final-model privacy certificate studied here. Our analysis uses the anchor for a second purpose: it changes the Lipschitz factor of the hidden update, which determines how much of an earlier record effect can be attenuated by later noisy updates.

The resulting connection is between optimization and privacy accounting. A stronger anchor can reduce the propagation factor, but the useful choice depends on the entire schedule rather than on one update in isolation. Section 4 makes this dependence explicit through the causal shift-allocation recursion, and Section 5 chooses the anchor strength from privatized pre-training measurements while preserving the same final-model analysis.

4 Certification Framework

This section turns the comparison of two neighboring runs into a computable update bound. The runs share the public initialization and batch order. When the changed slot is used, the two updates can differ directly. At later steps, the two hidden model states can still differ even if the current batch is the same. We call the first quantity the per-step map discrepancy and the second effect state propagation. PABI keeps them separate.

We write $\theta = \theta_0 + \mathsf{S}\phi$, where the fixed rescaling map S is computed from public initialization and architecture. All distances, clipping, projection, and Gaussian noise are then defined in ϕ. The Gaussian-smoothed field controls state propagation; the reference correction contributes only to the per-step map discrepancy. Sections 4.2–4.3 turn these quantities into the shifted-Rényi recursion and interval check.

PABI papers often denote the per-step map discrepancy by s_t, the allocated shift by a_t, and the remaining shift by z_t. We use $a_t I_t$, u_t, and h_t for these three roles, respectively, because $s_t = \eta_t\sigma_t$ denotes our Gaussian noise scale.

4.1 DP Proximal-Anchored SGD

We use add/remove adjacency in a fixed-slot representation. For public batches S_t of size b, write $I_t = \mathbf{1}\{i^\star \in S_t\}$ for access to the changed slot. The denominator remains b, and $\ell(\theta;\emptyset) = 0$. For each parameter block, let $r_j = \max\{\|\theta_0^{(j)}\|_F, 1\}$ and $\mathsf{S} = \mathrm{diag}((r_j/\sqrt{d_j})I_{d_j})$. This fixed public rescaling makes distances across blocks comparable and introduces no data-dependent state. We set $\theta = \theta_0 + \mathsf{S}\phi$, $\phi_0 = 0$, and $\widetilde{\ell}(\phi; z) = \ell(\theta_0 + \mathsf{S}\phi; z)$. Assume the public domain $\mathcal{K}_\theta$ is closed convex and contains θ_0 (we use a product of tensor Frobenius balls in the intended implementation). Then the induced set $\mathcal{K}_\phi = \{\phi : \theta_0 + \mathsf{S}\phi \in \mathcal{K}_\theta\}$ is also nonempty and closed convex, and its Euclidean projection $\Pi_{\mathcal{K}_\phi}$ is 1-Lipschitz. Clipping, projection, anchoring, and noise are applied in ϕ; the released model is the post-processing $\theta_T = \theta_0 + \mathsf{S}\phi_T$.

For a clipping norm $C > 0$ and an allowed record z, let $v_z(\phi) = \mathrm{clip}_C(\nabla_\phi \widetilde{\ell}(\phi; z))$. We assume that v_z is a measurable clipped per-record gradient map on $\mathbb{R}^d$. At the public scale $\tau_t > 0$, define the exact Gaussian-smoothed field

$$\bar{v}_{\tau_t,z}(\phi) = \mathbb{E}_{u\sim\mathcal{N}(0,I_d)}\left[v_z(\phi + \tau_t u)\right]. \tag{3}$$

The reference correction uses a public cap $R_t \geq 0$, weight $\kappa_t \in [0,1]$, and a public reference state $\bar{\phi}_t$:

$$\begin{aligned} r^R_{t,z}(\bar{\phi}_t) &= \mathrm{clip}_{R_t}\left(v_z(\bar{\phi}_t) - \bar{v}_{\tau_t,z}(\bar{\phi}_t)\right), \\ \widehat{v}_{t,z}(\phi) &= \bar{v}_{\tau_t,z}(\phi) + \kappa_t r^R_{t,z}(\bar{\phi}_t). \end{aligned} \tag{4}$$

The sequence $\{\bar{\phi}_t\}_{t<T}$ is fixed before private training from public information and is independent of the private data, training noise, and current hidden state. Each $\bar{\phi}_t$ is a public model state, not an iterate from either private run. One may set every $\bar{\phi}_t = \phi_0$ or compute the states once on a fixed public reference dataset. The batch field is

$$\widehat{g}^D_t(\phi) = \frac{1}{b}\sum_{i\in S_t} \widehat{v}_{t,z_i}(\phi). \tag{5}$$

When $\kappa_t = 1$ and the cap is inactive, the correction restores the clipped batch direction at $\bar{\phi}_t$. It is constant in the hidden state, so it changes only the per-step map discrepancy. In particular, when $\kappa_t = 1$ and $R_t = 2C$, the correction is exact at the reference state. Its contribution gives a batch-field discrepancy bound of $3C/b$ and an update-map discrepancy bound of $3\eta_t C/b$. Smaller caps reduce this bound at the cost of a larger correction error at the reference state.

The exact-expectation update is the mechanism covered by the privacy theorem. The finite-probe update below is a Monte Carlo approximation. To certify that approximation, one must add a uniform approximation error to the update-map discrepancy.

For a fixed public probe set $\{u_{t,k}\}_{k=1}^K$, chosen before private training, this implementation is

$$\begin{aligned} \widehat{\bar{v}}^{(K)}_{t,z}(\phi) &= \frac{1}{K}\sum_{k=1}^K v_z(\phi + \tau_t u_{t,k}), \\ \widetilde{v}^{(K)}_{t,z}(\phi) &= \widehat{\bar{v}}^{(K)}_{t,z}(\phi) + \kappa_t\,\mathrm{clip}_{R_t}\left(v_z(\bar{\phi}_t) - \widehat{\bar{v}}^{(K)}_{t,z}(\bar{\phi}_t)\right). \end{aligned} \tag{6}$$

The same probes are used at ϕ and $\bar{\phi}_t$. Thus, when $\kappa_t = 1$ and $R_t = 2C$, the estimator equals $v_z(\bar{\phi}_t)$ at the reference state for every K. The exact expectation in equation 3, rather than this finite-probe approximation, is the object of the privacy theorem.

For public schedules $\eta_t > 0$, $\lambda_t \geq 0$, and $\sigma_t > 0$, use the explicit gradient step for the loss plus the quadratic anchor $\lambda_t\|\phi - \phi_0\|^2/2$. We call it proximal-anchored because the quadratic term pulls the iterate toward ϕ_0.

$$\widetilde{m}^D_t(\phi) = \phi - \eta_t\left[\widehat{g}^D_t(\phi) + \lambda_t(\phi - \phi_0)\right]. \tag{7}$$

Let $\zeta_t \sim \mathcal{N}(0, I_d)$ be independent across updates and set $s_t = \eta_t\sigma_t$. The next hidden state is obtained by adding Gaussian noise and projecting back to $\mathcal{K}_\phi$; its transition kernel is

$$\widetilde{K}^D_t(\phi, \cdot) = \mathrm{Law}\big(\Pi_{\mathcal{K}_\phi}\left[\widetilde{m}^D_t(\phi) + s_t\zeta_t\right]\big). \tag{8}$$

Both runs start at ϕ_0; only θ_T is released. Let $\mathcal{M}_{\mathrm{base}}(D)$ denote the distribution of this released final model.

The Gaussian-smoothed field is Lipschitz with constant L_t. The reference correction is independent of the current hidden state, so it does not change the propagation factor. The map difference therefore has the form

$$\left\|\widetilde{m}^D_t(\phi) - \widetilde{m}^{D'}_t(\phi')\right\| \leq c_t\|\phi - \phi'\| + a_t I_t, \tag{9}$$

with

$$\begin{aligned} c_t &= |1 - \eta_t\lambda_t| + \eta_t L_t, \\ a_t &= \frac{\eta_t(C + \kappa_t R_t)}{b}. \end{aligned} \tag{10}$$

Here c_t bounds the propagation of an existing state difference, and $a_t I_t$ bounds the new difference caused by an access. The correction is constant in the current state, so it contributes to a_t through its record dependence. The next subsection uses these two quantities to bound final-model privacy.

4.2 FINAL-MODEL PRIVACY CERTIFICATE

At update t, neighboring runs can enter with different hidden-state laws even when $I_t = 0$. For $\alpha > 1$ and $h \geq 0$, define the shifted-Rényi divergence

$$\mathcal{D}_\alpha^{(h)}(P\|Q) = \inf_{\widetilde{P}: W_\infty(P,\widetilde{P}) \leq h} D_\alpha(\widetilde{P}\|Q). \tag{11}$$

Here W_∞ is the essential-supremum transport distance and $\mathcal{D}_\alpha^{(0)} = D_\alpha$. Thus h records an unresolved hidden-state difference that can be moved by at most h before measuring RDP. Write $\widetilde{K}_t^D P$ for the law obtained by applying the kernel to P.

Lemma 1 (One-step shifted-Rényi transport)**.** *Under equation 9, let P, Q be supported on $\mathcal{K}_\phi$. For $\alpha > 1$, $h \geq 0$, and $0 \leq u \leq c_t h + a_t I_t$, set $h^+ = c_t h + a_t I_t - u$. Then*

$$\mathcal{D}_\alpha^{(h^+)}(\widetilde{K}_t^D P\|\widetilde{K}_t^{D'} Q) \leq \mathcal{D}_\alpha^{(h)}(P\|Q) + \frac{\alpha u^2}{2s_t^2}. \tag{12}$$

The update maps create a discrepancy bounded by $c_t h + a_t I_t$. The Gaussian step accounts for a chosen amount u of this discrepancy and incurs RDP cost $\alpha u^2/(2s_t^2)$; the remaining shift enters the next hidden state. Projection onto $\mathcal{K}_\phi$ is nonexpansive. This is the one-step PABI reduction: the update map propagates the remaining shift, and the Gaussian shift-reduction step accounts for the amount allocated at that update. The reference correction is already included in a_t.

Let h_t be the remaining shift before update t and $u_t \geq 0$ the shift allocated at that update. Set $h_0 = 0$ and require $h_T = 0$ at release. A feasible schedule obeys

$$h_{t+1} = c_t h_t + a_t I_t - u_t \geq 0. \tag{13}$$

The causal final-model privacy certificate is

$$\mathcal{C}_\alpha(I) = \min_{(u,h)\ \text{feasible}} \frac{\alpha}{2} \sum_{t=0}^{T-1} \frac{u_t^2}{s_t^2}. \tag{14}$$

Theorem 1 (Causal final-model privacy certificate)**.** *Under equation 9, for every $\alpha > 1$ and every neighboring pair $D \sim D'$ with access pattern I,*

$$\begin{aligned} D_\alpha\big(\mathcal{M}_{\text{base}}(D)\|\mathcal{M}_{\text{base}}(D')\big) &\leq \mathcal{C}_\alpha(I) \\ &\leq \frac{\alpha}{2} \sum_{t=0}^{T-1} \left(\frac{a_t I_t}{s_t}\right)^2. \end{aligned} \tag{15}$$

The first inequality is the PABI final-model privacy guarantee. Immediate allocation, $u_t = a_t I_t$, means charging each access entirely at the update where it occurs and gives the second inequality. The optimized program can allocate part of a discrepancy to later updates and be strictly smaller. Its feasible schedules form a one-dimensional convex quadratic program with ordered cumulative constraints, so the monotone lower-hull scan used here takes $O(T)$ arithmetic operations; if a propagation factor is zero, the scan restarts because no earlier shift can cross that step. For the fixed-slot DP guarantee, the pattern-specific value is maximized over the possible changed slots; the uniform envelope below performs this maximization through $e_\star$.

4.3 VERIFIABLE CONDITIONS FOR AN EPOCH-INDEPENDENT CERTIFICATE

Smoothing averages each clipped per-record gradient around the current hidden state. Because the gradients are clipped, this average changes at a bounded rate as the state moves. This gives the Lipschitz premise needed for the PABI propagation factor. Gaussian integration by parts gives

$$\begin{aligned} p^\top D\bar{v}_{\tau,z}(\phi) q &= \frac{1}{\tau}\mathbb{E}\left[p^\top v_z(\phi + \tau u)\, q^\top u\right], \\ \|D\bar{v}_{\tau,z}(\phi)\|_{\text{op}} &\leq \frac{C}{\tau}\sqrt{\frac{2}{\pi}}. \end{aligned} \tag{16}$$

Algorithm 1 Public certification of the reference-corrected field

Require: Public $C, b, \{\eta_t, \lambda_t, \tau_t, R_t, \kappa_t, \bar{\phi}_t\}_{t<T}$, target rate γ, and allowance $B \geq 0$
Ensure: Public $\{L_t, a_t, c_t\}_{t<T}$ and an accepted certification check, or infeasible
for $t = 0, \ldots, T-1$ **do**
  $L_t \leftarrow C\sqrt{2/\pi}/\tau_t$
  $a_t \leftarrow \eta_t(C + \kappa_t R_t)/b$
  $c_t \leftarrow |1 - \eta_t\lambda_t| + \eta_t L_t$
  if $\eta_t L_t > e^{-\gamma}$ **then**
    return infeasible
  end if
end for
Check the interval maximum by a prefix scan and require
  $\max_{0 \leq j < k \leq T} \{\log \prod_{t=j}^{k-1} c_t + \gamma(k-j)\} \leq B$
Treat zero-product intervals through the causal scan
return the public record and accepted

Lemma 2 (Gaussian smoothing certificate)**.** *For every $\tau > 0$, every allowed record z, and all $\phi, \phi' \in \mathbb{R}^d$,*

$$\|\bar{v}_{\tau,z}(\phi) - \bar{v}_{\tau,z}(\phi')\| \leq L_\tau \|\phi - \phi'\|,$$
$$L_\tau = \frac{C}{\tau}\sqrt{\frac{2}{\pi}}. \tag{17}$$

The same constant bounds every fixed-denominator batch average.

The reference correction does not change L_t; it changes only the per-step map-discrepancy coefficient in equation 10. The public certification check sets

$$L_t = L_{\tau_t} = \frac{C}{\tau_t}\sqrt{\frac{2}{\pi}}. \tag{18}$$

The test $\eta_t L_t \leq e^{-\gamma}$ keeps the anchor controller feasible for every monitor transcript. The stronger condition is

$$|1 - \eta_t\lambda_t| + \eta_t L_t \leq e^{-\gamma}. \tag{19}$$

It controls state propagation; the correction sensitivity affects the allocated shift in the final-model certificate.

Algorithm 1 is the public certification check for a fixed anchor schedule. It computes L_t and a_t before private training; the adaptive controller computes c_t after choosing λ_t. The correction caps do not reduce L_t.

The public certification check costs $O(T)$ time and constant memory. Let $e_t(I_t) = a_t I_t$ and $e_\star = \max_t a_t$. The causal scan allows $c_t = 0$; the closed-form envelope below assumes $c_t > 0$, which holds when $\tau_t, C > 0$ and $0 \leq \lambda_t \leq 1/\eta_t$. If $\eta_t = \eta$, $\sigma_t = \sigma$, and the positive c_t satisfy

$$\prod_{t=j}^{k-1} c_t \leq \exp\big(B - \gamma(k-j)\big) \text{for every } 0 \leq j < k \leq T, \tag{20}$$

then the following epoch-independent certificate holds, provided B, γ, $e_\star$, η, and σ have public bounds independent of T.

Theorem 2 (Epoch-independent final-model privacy certificate)**.** *For every $\alpha > 1$,*

$$D_\alpha\big(\mathcal{M}_{\text{base}}(D) \| \mathcal{M}_{\text{base}}(D')\big) \leq \frac{\alpha e_\star^2}{2\eta^2\sigma^2} \min\left\{T, \frac{4e^{2B}}{1 - e^{-\gamma}}\right\}. \tag{21}$$

The causal program in equation 14 is the pattern-specific certificate; for fixed-slot DP its value is maximized over the possible changed slots. Equation equation 21 is the uniform envelope, independent of the number of updates when its public parameters are uniform. The reference correction enters the per-step map-discrepancy coefficient and leaves the PABI propagation argument unchanged.

5 Certification-Adaptive Algorithm

For the adaptive result, $\eta_t = \eta$ and $\sigma_t = \sigma$ are public; privatized threshold bits choose the anchor before training, and the interval condition is enforced for every transcript.

Algorithm 2 Certification-adaptive reference-corrected DP training

Require: Dataset D; public $\theta_0, \mathsf{S}, \mathcal{K}_\phi, \{S_t, \tau_t, R_t, \kappa_t, \bar{\phi}_t, L_t\}_{t<T}, C, b, \eta, \sigma$
Require: Public $\gamma, B_\star, \mathcal{Q}, \omega, \sigma_{\mathrm{m}}, \lambda_{\mathrm{lo}}, \lambda_{\mathrm{hi}}$
for $t \in \mathcal{Q}$ **do**
  Sample $\xi_t \sim \mathcal{N}(0, \sigma_{\mathrm{m}}^2)$ and set $Z_t \leftarrow \mathbf{1}\{\|\widehat{g}_t^D(\phi_0)\| + \xi_t \geq \omega\}$
end for
$\phi \leftarrow \phi_0$, $H \leftarrow 0$, $f \leftarrow 0$
for $t = 0, \ldots, T-1$ **do**
  if $\eta L_t > \exp(B_\star - H - \gamma)$ **then**
    break
  end if
  if $t \in \mathcal{Q}$ and $Z_t = 1$ **then**
    $f \leftarrow 1$
  end if
  Compute $\lambda_t^{\mathrm{safe}}$ from equation 23 with $H_t = H$
  $\lambda_t \leftarrow \max\{\lambda_t^{\mathrm{safe}}, (1-f)\lambda_{\mathrm{lo}} + f\lambda_{\mathrm{hi}}\}$
  $c_t \leftarrow |1 - \eta\lambda_t| + \eta L_t$
  Sample $\zeta_t \sim \mathcal{N}(0, I_d)$ and set $\phi \leftarrow \Pi_{\mathcal{K}_\phi}[\widetilde{m}_t^D(\phi) + \eta\sigma\zeta_t]$
  $H \leftarrow \max\{0, H + \log c_t + \gamma\}$
end for
return $\theta_0 + \mathsf{S}\phi$

5.1 Pre-training Private Monitoring and Anchor Selection

Fix a target rate $\gamma > 0$ and public allowance $B_\star \geq 0$. For completed updates let $H_0 = 0$ and

$$H_{t+1} = \max\{0, H_t + \log c_t + \gamma\}. \tag{22}$$

The safe anchor enforces the contiguous-interval condition from Section 4.3.

$$\lambda_t^{\mathrm{safe}} = \frac{[1 + \eta L_t - \exp(B_\star - H_t - \gamma)]_+}{\eta}. \tag{23}$$

An update is accepted when $\eta L_t \leq \exp(B_\star - H_t - \gamma)$ with $0 \leq \lambda_t \leq 1/\eta$; correction caps enter through the per-step map-discrepancy bound. If $\eta L_t \leq e^{-\gamma}$ for all t, every monitor outcome completes all T updates.

5.2 Private Monitoring and Adaptive Updates

Let $\mathcal{Q} \subseteq \{0, \ldots, T-1\}$ be finite and let $\sigma_{\mathrm{m}} > 0$. At each checkpoint $t \in \mathcal{Q}$, threshold $q_t(D) = \|\widehat{g}_t^D(\phi_0)\|$ with independent $\xi_t \sim \mathcal{N}(0, \sigma_{\mathrm{m}}^2)$ and set $Z_t = \mathbf{1}\{q_t(D) + \xi_t \geq \omega\}$. The fixed-slot sensitivity is

$$|q_t(D) - q_t(D')| \leq \frac{C + \kappa_t R_t}{b} \mathbf{1}\{i^\star \in S_t\}. \tag{24}$$

Assume $0 \leq \lambda_{\mathrm{lo}} \leq \lambda_{\mathrm{hi}} \leq 1/\eta$. A positive bit selects λ_{hi}; the actual strength is the maximum of this candidate and $\lambda_t^{\mathrm{safe}}$. The bits are fixed inputs.

5.3 Joint Privacy Guarantee

Let T be the declared maximum horizon, $\bar{a}_t = (C + \kappa_t R_t)/b$, $\bar{e}_\star = \max_t \bar{a}_t$, and $e_t = \eta \bar{a}_t I_t$. The bound uses this public horizon even when a transcript stops early.

Theorem 3 (Privacy of certification-adaptive training)**.** *For every $\alpha > 1$, the auxiliary joint output containing the monitor bits and final model satisfies*

$$D_\alpha\big(\mathcal{M}_{\mathrm{adapt}}^{\mathrm{joint}}(D) \| \mathcal{M}_{\mathrm{adapt}}^{\mathrm{joint}}(D')\big) \leq \frac{\alpha|\mathcal{Q}|}{2\sigma_{\mathrm{m}}^2} \max_{t \in \mathcal{Q}} \left(\frac{C + \kappa_t R_t}{b}\right)^2 + \frac{\alpha \bar{e}_\star^2}{2\sigma^2} \min\left\{T, \frac{4e^{2B_\star}}{1 - e^{-\gamma}}\right\}. \tag{25}$$

When $\mathcal{Q} = \varnothing$, the first term is defined to be zero. The released final model inherits this guarantee by post-processing.

The first term pays for the privatized bits; the second is uniform over all transcripts. The controller depends only on c_t, so the interval condition holds on every branch, while the PABI recursion accounts for attenuation by later noisy updates.

6 EXPERIMENTS

We evaluate the final-model certificate, its computation time, and utility. Privacy curves use the exact certificate; utility uses one antithetic probe pair and therefore evaluates a finite-probe approximation.

6.1 EXPERIMENTAL SETTING

All utility methods start from the same public task-fine-tuned initialization, whose test accuracies are 0.7523, 0.8632, 0.8777, and 0.5436 on SST-2, AG News, CIFAR-10, and CIFAR-100. We use $n = b = 8192$ fixed slots, fixed-slot add/remove adjacency, zero gradient for an empty slot, and $\delta = 10^{-5}$, without subsampling amplification. The smoothed-anchor run uses Gaussian smoothing at the clipping scale and a proximal strength accepted by the public certification check; DP-SGD and DP-Adam use step-wise RDP accounting.

6.2 SUMMARY OF RESULTS

The exact certificate remains bounded as the horizon grows: at 500 epochs it is 2.73, compared with 10.96 for step-wise RDP. The public certification check and causal accountant take less than 1 ms on one CPU thread. The finite-probe utility results depend on the initialization and dataset: the smoothed anchor is highest on SST-2 after public task fine-tuning, whereas DP-SGD or DP-Adam is higher on the other task-fine-tuned datasets and on the original pretrained checkpoints. The complete privacy and runtime curves are shown in Figure 1; the four public-initialization utility curves are in Figure 2, and the two original-pretrained curves are in Figure 3. Appendix C contains every experiment and its full tables; Appendix C.2 gives the complete privacy and runtime sweeps, while Appendices C.3 and C.4 give the two utility settings.

7 CONCLUSION

We studied final-model privacy for clipped, projected proximal-anchored DP-SGD when one record can be accessed across many updates. The shifted-Rényi PABI analysis separates the discrepancy created at an access from the state difference propagated by later hidden updates. Exact Gaussian smoothing gives a dimension-free propagation bound, and a bounded correction at a fixed public reference state is charged to the access term. Causal shift allocation computes the certificate in linear time, while a contiguous-interval contraction condition makes the bound independent of the training horizon. An adaptive anchor rule extends the certificate to privatized monitor bits and the final model. Experiments show the resulting certificate remains bounded as the horizon grows; finite-probe utility is evaluated separately from the exact-expectation theorem.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing. All AI-assisted polishing were reviewed against the original text and revised by the authors to ensure accuracy. We take full responsibility for the final content of this work, including any text produced with the assistance of generative AI.

ETHICS STATEMENT

This work focuses on the theory and does not involve human subjects, private or personally identifiable information, or interactions with production systems. All experiments are conducted in controlled settings using publicly available models and datasets. The proposed watermarking method is intended to support content provenance and the identification of machine-generated text. Although watermarking research may have dual-use implications, this study does not collect or release sensitive data and explicitly presents its assumptions and limitations. We identify no additional ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

Codes, all experimental results, and all hyperparameter settings are submitted to the Supplementary Material on OpenReview.

REFERENCES

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Jason M. Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. In *Advances in Neural Information Processing Systems*, 2022.

Jinho Bok, Weijie J. Su, and Jason M. Altschuler. Shifted interpolation for differential privacy. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4230–4266, 2024.

Mario Bravo, Juan P. Flores-Mella, and Cristóbal Guzmán. Mixing times and privacy analysis for the projected langevin algorithm under a modulus of continuity. *Journal of Machine Learning Research*, 26(304):1–33, 2025.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284. USENIX Association, 2019.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, pp. 2633–2650, 2021.

Eli Chien and Pan Li. Convergent privacy loss of Noisy-SGD without convexity and smoothness. In *International Conference on Learning Representations*, 2025.

Eli Chien, Wei-Ning Chen, and Pan Li. Privacy amplification in differentially private zeroth-order optimization with hidden states. In *International Conference on Machine Learning*, 2026.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? In *Conference on Language Modeling (COLM)*, 2024.

Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *IEEE Symposium on Foundations of Computer Science*, 2018.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. Lisa: Lazy safety alignment for large language models against harmful fine-tuning attack. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-3320.

Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In *International Conference on Learning Representations*, 2023.

Weiwei Kong and Mónica Ribero. Privacy of the last iterate in cyclically-traversed DP-SGD on nonconvex composite losses, 2025. Version 3.

Danny D. Leybzon and Corentin Kervadec. Learning, forgetting, remembering: Insights from tracking llm memorization during training. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 43–57, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.4.

Tian Li, Anit Kumar Sahu, Maziar Zaheer, Manik Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, 2020.

Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium*, pp. 263–275, 2017. doi: 10.1109/CSF.2017.11.

Xihe Qiu, Leijun Cheng, Teqi Hao, and Xiaoyu Tan. ArMA: Mitigating catastrophic forgetting using attention-regularized model averaging in continual fine-tuning large language models. *IEEE Transactions on Artificial Intelligence*, 2026. doi: 10.1109/TAI.2025.3630623.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017.

Junjiao Tian, Chengyue Huang, and Zsolt Kira. Rethinking weight decay for robust fine-tuning of foundation models. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-0705.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, André Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022.

A Proofs for the Certification Framework

This appendix proves the main ingredients used by the certification framework. We first establish the dimension-free Lipschitz bound for the exact Gaussian-smoothed field and the one-step shifted-Rényi transport inequality. We then prove the causal final-model privacy certificate, the linear-time evaluation of its scalar optimization problem, the concrete map-discrepancy bound, the epoch-independent envelope, and the joint guarantee for the adaptive controller.

A.1 Proof of Lemma 2

Lemma 3 (Gaussian smoothing certificate)**.** *For every $\tau > 0$, every allowed record z, and all $\phi, \phi' \in \mathbb{R}^d$,*

$$\|\bar{v}_{\tau,z}(\phi) - \bar{v}_{\tau,z}(\phi')\| \le \frac{C}{\tau}\sqrt{\frac{2}{\pi}}\|\phi - \phi'\|. \tag{26}$$

The same constant bounds every fixed-denominator batch average.

Proof. Let $v_z(\phi) = \mathrm{clip}_C(\nabla_\phi \widetilde{\ell}(\phi; z))$. The proof uses only $\|v_z(\phi)\| \le C$. A bounded measurable function convolved with a non-degenerate Gaussian density is continuously differentiable. For unit vectors p, q, differentiating the Gaussian density in direction q and integrating by parts gives

$$p^\top D\bar{v}_{\tau,z}(\phi)q = \frac{1}{\tau}\mathbb{E}\left[p^\top v_z(\phi + \tau U)\; q^\top U\right], \qquad U \sim \mathcal{N}(0, I_d). \tag{27}$$

Indeed, the directional derivative of the Gaussian density contributes $(q^\top U)/\tau$, and the boundedness of v_z makes the resulting expectation integrable. Since $|p^\top v_z| \le C$ and $q^\top U \sim \mathcal{N}(0,1)$,

$$\left|p^\top D\bar{v}_{\tau,z}(\phi)q\right| \le \frac{C}{\tau}\mathbb{E}|q^\top U| = \frac{C}{\tau}\sqrt{\frac{2}{\pi}}. \tag{28}$$

Taking the supremum over unit p, q gives

$$\|D\bar{v}_{\tau,z}(\phi)\|_{\mathrm{op}} \le \frac{C}{\tau}\sqrt{\frac{2}{\pi}}. \tag{29}$$

For the segment $\phi_r = \phi + r(\phi' - \phi)$, the fundamental theorem of calculus therefore yields

$$\bar{v}_{\tau,z}(\phi') - \bar{v}_{\tau,z}(\phi) = \int_0^1 D\bar{v}_{\tau,z}(\phi_r)(\phi' - \phi)\, dr, \tag{30}$$

$$\|\bar{v}_{\tau,z}(\phi') - \bar{v}_{\tau,z}(\phi)\| \le \frac{C}{\tau}\sqrt{\frac{2}{\pi}}\|\phi' - \phi\|. \tag{31}$$

A fixed-denominator batch field is the average of the corresponding per-record fields. Applying the triangle inequality to this average gives the same Lipschitz constant. The empty-slot field is identically zero. □

A.2 Proof of Lemma 1

Lemma 4 (One-step shifted-Rényi transport)**.** *Under equation 9, let P, Q be supported on $\mathcal{K}_\phi$. For every $\alpha > 1$, $h \geq 0$, and $0 \leq u \leq c_t h + a_t I_t$, set $h^+ = c_t h + a_t I_t - u$. Then*

$$\mathcal{D}_\alpha^{(h^+)}(\widetilde{K}_t^D P \| \widetilde{K}_t^{D'} Q) \leq \mathcal{D}_\alpha^{(h)}(P\|Q) + \frac{\alpha u^2}{2s_t^2}. \tag{32}$$

Proof. If $\mathcal{D}_\alpha^{(h)}(P\|Q) = +\infty$, the claim is immediate. Otherwise, fix $\varepsilon > 0$ and choose a witness $\widetilde{P}_\varepsilon$ such that

$$W_\infty(P, \widetilde{P}_\varepsilon) \leq h, \qquad D_\alpha(\widetilde{P}_\varepsilon \| Q) \leq \mathcal{D}_\alpha^{(h)}(P\|Q) + \varepsilon. \tag{33}$$

Let $\widehat{P}_\varepsilon = (\Pi_{\mathcal{K}_\phi})_\# \widetilde{P}_\varepsilon$. Because P and Q are supported on $\mathcal{K}_\phi$, the projection leaves both unchanged. Applying the projection to a coupling in the definition of W_∞ and using its 1-Lipschitz property gives

$$W_\infty(P, \widehat{P}_\varepsilon) \leq W_\infty(P, \widetilde{P}_\varepsilon) \leq h. \tag{34}$$

The same projection in both arguments and Rényi data processing give

$$D_\alpha(\widehat{P}_\varepsilon \| Q) \leq D_\alpha(\widetilde{P}_\varepsilon \| Q) \leq \mathcal{D}_\alpha^{(h)}(P\|Q) + \varepsilon. \tag{35}$$

The coupling characterization of W_∞ gives a coupling (X, Y) with $X \sim P$, $Y \sim \widehat{P}_\varepsilon$, and $\|X - Y\| \leq h$ almost surely. Put $v = c_t h + a_t I_t$. Applying the map bound pointwise gives

$$\|\widetilde{m}_t^D(X) - \widetilde{m}_t^{D'}(Y)\| \leq c_t \|X - Y\| + a_t I_t \leq v. \tag{36}$$

Consequently $(\widetilde{m}_t^{D'})_\# \widehat{P}_\varepsilon$ is a v-shift witness for $(\widetilde{m}_t^D)_\# P$ relative to $(\widetilde{m}_t^{D'})_\# Q$. Since both witness measures are obtained by the same map $\widetilde{m}_t^{D'}$, Rényi data processing yields

$$\mathcal{D}_\alpha^{(v)}\left((\widetilde{m}_t^D)_\# P \middle\| (\widetilde{m}_t^{D'})_\# Q\right) \leq D_\alpha(\widehat{P}_\varepsilon \| Q). \tag{37}$$

Combining this with equation 35 and then letting $\varepsilon \downarrow 0$ gives the map-level shifted-divergence bound.

It remains to account for the Gaussian update. The Gaussian shift-reduction inequality from Feldman et al. (2018, Lemma 20), in the form used by Altschuler & Talwar (2022), states that for $v \geq u \geq 0$,

$$\mathcal{D}_\alpha^{(v-u)}\left(\mu * \mathcal{N}(0, s_t^2 I_d) \middle\| \nu * \mathcal{N}(0, s_t^2 I_d)\right) \leq \mathcal{D}_\alpha^{(v)}(\mu\|\nu) + \frac{\alpha u^2}{2s_t^2}. \tag{38}$$

The condition $v \geq u$ is exactly the feasibility condition in the lemma. Apply equation 38 to the two unprojected map push-forwards with $v = c_t h + a_t I_t$. Finally, apply the common projection $\Pi_{\mathcal{K}_\phi}$. Projection preserves every W_∞ budget because it is 1-Lipschitz and cannot increase ordinary Rényi divergence by data processing. Thus it cannot increase shifted Rényi divergence, and the resulting projected kernels satisfy equation 32. □

A.3 Proof of Theorem 1

Theorem 4 (Causal final-model privacy certificate)**.** *Under equation 9, for every $\alpha > 1$ and every neighboring pair $D \sim D'$ with access pattern I,*

$$D_\alpha\big(\mathcal{M}_{\text{base}}(D) \| \mathcal{M}_{\text{base}}(D')\big) \leq \mathcal{C}_\alpha(I) \leq \frac{\alpha}{2} \sum_{t<T} \left(\frac{a_t I_t}{s_t}\right)^2. \tag{39}$$

Proof. Let P_t and Q_t be the hidden-state laws after t updates on D and D'. The common public initialization gives $P_0 = Q_0 = \delta_{\phi_0}$ and $\mathcal{D}_\alpha^{(0)}(P_0\|Q_0) = 0$. Since $\phi_0 \in \mathcal{K}_\phi$ and every transition is followed by the projection $\Pi_{\mathcal{K}_\phi}$, all P_t and Q_t are supported on $\mathcal{K}_\phi$. Fix any feasible sequence $(u_t, h_t)_{t<T}$ satisfying equation 13. The bounds in the recursion imply $0 \leq u_t \leq c_t h_t + a_t I_t$, so Lemma 1 gives

$$\mathcal{D}_\alpha^{(h_{t+1})}(P_{t+1}\|Q_{t+1}) \leq \mathcal{D}_\alpha^{(h_t)}(P_t\|Q_t) + \frac{\alpha u_t^2}{2s_t^2}. \tag{40}$$

Applying this inequality for $t = 0, \dots, T-1$ and summing gives

$$\mathcal{D}_\alpha^{(h_T)}(P_T \| Q_T) \le \frac{\alpha}{2} \sum_{t<T} \frac{u_t^2}{s_t^2}. \tag{41}$$

Because feasibility requires $h_T = 0$, $\mathcal{D}_\alpha^{(h_T)}(P_T \| Q_T) = D_\alpha(P_T \| Q_T)$. Taking the infimum over all feasible schedules gives $D_\alpha(P_T \| Q_T) \le \mathcal{C}_\alpha(I)$. The released model is $\theta_T = \theta_0 + \mathsf{S}\phi_T$, a deterministic post-processing of ϕ_T, so Rényi data processing transfers the same bound to $\mathcal{M}_{\mathrm{base}}(D)$ and $\mathcal{M}_{\mathrm{base}}(D')$.

For the second inequality, choose the immediate allocation $u_t = a_t I_t$ and $h_t = 0$ for every t. It satisfies the recursion because $h_{t+1} = c_t \cdot 0 + a_t I_t - a_t I_t = 0$. Its objective value is exactly $\frac{\alpha}{2} \sum_t (a_t I_t / s_t)^2$, and the minimum over all feasible schedules cannot be larger. □

Linear-time evaluation of the causal certificate. We justify the $O(T)$ computation claim following Theorem 1. If $c_j = 0$ for some update, let j be the last such index; if no factor is zero, set $j = 0$. The prefix before j cannot affect the recursion after update j, because $c_j h_j = 0$. Hence the optimum equals the optimum of the suffix $j, \dots, T-1$ with initial shift $h_j = 0$. One direction follows by retaining the allocations from any feasible full schedule from j onward and resetting the suffix initial shift to zero. At update j, the state entering the next step is still $a_j I_j - u_j$, and all later constraints are unchanged, while earlier allocations were nonnegative. For the reverse direction, extend any feasible suffix by setting $u_t = 0$ for $t < j$ and propagating the shift from $h_0 = 0$; the prefix has zero objective cost and the factor $c_j = 0$ erases the prefix shift. The update j remains in the suffix because it can inject $a_j I_j$ even though it erases all earlier state difference.

On the positive suffix, define

$$w_t = \prod_{v=t+1}^{T-1} c_v, \qquad j \le t < T, \tag{42}$$

and, for $j \le m \le T$,

$$\begin{aligned} A_m &= \sum_{t=j}^{m-1} w_t a_t I_t, \\ V_m &= \sum_{t=j}^{m-1} w_t^2 s_t^2, \\ Y_m &= \sum_{t=j}^{m-1} w_t u_t. \end{aligned} \tag{43}$$

Multiplying the causal recursion by $\prod_{v=m}^{T-1} c_v$ and unrolling from j to $m-1$ gives

$$\left(\prod_{v=m}^{T-1} c_v \right) h_m = A_m - Y_m. \tag{44}$$

For $j < m \le T$ the product on the left is positive; at $m = j$ both sides are zero because $c_j = 0$. Therefore feasibility is equivalent to $Y_j = 0$, $Y_T = A_T$, $Y_m \le A_m$ for $j < m < T$, and $Y_{m+1} - Y_m = w_m u_m \ge 0$. Since $V_{m+1} - V_m = w_m^2 s_m^2 > 0$, the objective becomes

$$\mathcal{C}_\alpha(I) = \frac{\alpha}{2} \min_Y \sum_{m=j}^{T-1} \frac{(Y_{m+1} - Y_m)^2}{V_{m+1} - V_m}. \tag{45}$$

Let $\underline{A}_m$ be the greatest convex minorant of the points (V_m, A_m), including the fixed endpoints. Its consecutive slopes are nondecreasing by convexity. They are also nonnegative: the first edge of the lower hull joins $(V_j, A_j) = (0, 0)$ to a later data point (V_k, A_k) with $V_k > 0$ and $A_k \ge 0$, so its slope is nonnegative; every later edge has at least this slope. Thus $Y_m = \underline{A}_m$ satisfies the ordered constraints. For an interior point put $r_m = (Y_{m+1} - Y_m)/(V_{m+1} - V_m)$. The derivative of the

quadratic objective with respect to Y_m is $\alpha(r_{m-1} - r_m)$. The convex minorant has $r_{m-1} \le r_m$, so the multipliers $\lambda_m = \alpha(r_m - r_{m-1})$ are nonnegative. A positive increase of the lower-hull slope can occur only at a contact point with the data; otherwise the corresponding hull vertex could be moved upward slightly while remaining below the data, contradicting maximality of the minorant. Hence $\lambda_m(Y_m - A_m) = 0$. The endpoint constraints and the nonnegative slope constraints are also satisfied. Taking the multipliers of the monotonicity constraints $Y_{m-1} - Y_m \le 0$ to be zero gives stationarity, dual feasibility, and complementarity for the Karush–Kuhn–Tucker system. The strictly convex quadratic program therefore has the convex minorant as its unique optimizer, and

$$\mathcal{C}_\alpha(I) = \frac{\alpha}{2} \sum_{m=j}^{T-1} \frac{(\underline{A}_{m+1} - \underline{A}_m)^2}{V_{m+1} - V_m}. \tag{46}$$

A monotone lower-hull scan adds or removes each suffix point only a constant number of times. The weights, transformed coordinates, and certificate are therefore computed in $O(T)$ arithmetic operations for a fixed access pattern and RDP order.

A.4 Proof of the per-step map-discrepancy bound

Lemma 5 (Reference-correction update bound)**.** *For the field in equation 4, the map bound equation 9 holds with the coefficients in equation 10.*

Proof. Write

$$\begin{aligned}
\widetilde{m}_t^D(\phi) - \widetilde{m}_t^{D'}(\phi') &= (1 - \eta_t\lambda_t)(\phi - \phi') - \eta_t \left[\widehat{g}_t^D(\phi) - \widehat{g}_t^{D'}(\phi')\right] \\
&= (1 - \eta_t\lambda_t)(\phi - \phi') - \eta_t \left[\widehat{g}_t^D(\phi) - \widehat{g}_t^D(\phi')\right] \\
&\quad - \eta_t \left[\widehat{g}_t^D(\phi') - \widehat{g}_t^{D'}(\phi')\right].
\end{aligned} \tag{47}$$

The reference correction for a fixed record z is independent of the current state. Hence it cancels in the same-record state difference $\widehat{v}_{t,z}(\phi) - \widehat{v}_{t,z}(\phi')$. By the Gaussian smoothing lemma, each smoothed per-record field is L_t-Lipschitz, and averaging over the fixed batch with denominator b preserves this constant. Therefore

$$\|\widehat{g}_t^D(\phi) - \widehat{g}_t^D(\phi')\| \le L_t\|\phi - \phi'\|. \tag{48}$$

For the second bracket, the two datasets differ only at slot $i^\star$. If $i^\star \notin S_t$, the batch fields are identical. If $i^\star \in S_t$, one dataset contains the empty slot and the other contains a record z. The empty slot has $v_\emptyset = \bar{v}_{\tau_t,\emptyset} = 0$ and zero reference correction. For the nonempty slot, $\|\bar{v}_{\tau_t,z}(\phi')\| \le C$ because it is an expectation of vectors of norm at most C, while

$$\left\|\kappa_t r_{t,z}^R(\bar{\phi}_t)\right\| \le \kappa_t R_t.$$

The fixed denominator therefore gives

$$\|\widehat{g}_t^D(\phi') - \widehat{g}_t^{D'}(\phi')\| \le \frac{C + \kappa_t R_t}{b} I_t. \tag{49}$$

Combining equation 47, equation 48, and equation 49 with the triangle inequality gives

$$\|\widetilde{m}_t^D(\phi) - \widetilde{m}_t^{D'}(\phi')\| \le \left(|1 - \eta_t\lambda_t| + \eta_t L_t\right)\|\phi - \phi'\| + \frac{\eta_t(C + \kappa_t R_t)}{b} I_t. \tag{50}$$

This is equation 9 with the coefficients in equation 10. The projection is handled later by the shifted-transport lemma and does not change this deterministic map bound. □

A.5 Proof of the epoch-independent envelope

Theorem 5 (Epoch-independent final-model privacy certificate)**.** *Assume $T \ge 1$, fixed $\eta_t = \eta > 0$, fixed $\sigma_t = \sigma > 0$, positive c_t, and $B \ge 0$, $\gamma > 0$ such that*

$$\prod_{t=j}^{k-1} c_t \le \exp(B - \gamma(k - j)) \quad \textit{for every } 0 \le j < k \le T. \tag{51}$$

Let $e_\star = \max_{t<T} a_t$. Then, for every $\alpha > 1$ and every access pattern $I \in \{0,1\}^T$,

$$D_\alpha\big(\mathcal{M}_{\mathrm{base}}(D)\|\mathcal{M}_{\mathrm{base}}(D')\big) \le \frac{\alpha e_\star^2}{2\eta^2\sigma^2}\min\left\{T, \frac{4e^{2B}}{1-e^{-\gamma}}\right\}. \tag{52}$$

Proof. Set $q = e^{-\gamma} \in (0,1)$, $\Gamma = e^B \ge 1$, and $s = \eta\sigma$. We first construct a feasible schedule for a dominating recursion with an injection $e_\star$ at every update. Define the backward weights

$$Q_0 = 1, \qquad Q_k = \prod_{v=T-k}^{T-1} c_v, \quad 1 \le k \le T. \tag{53}$$

For $0 \le j \le k \le T$, the interval condition applied to the last $k-j$ updates gives

$$Q_k \le \Gamma q^{k-j} Q_j. \tag{54}$$

The case $j = k$ follows from $\Gamma \ge 1$.

Fix $h \in \{1, \dots, T\}$ and define

$$G_h = \sum_{j=0}^{h-1} Q_j, \qquad G_T = \sum_{k=0}^{T-1} Q_k. \tag{55}$$

Because $Q_0 = 1$, $G_h > 0$. Allocate

$$u_t^{\mathrm{F}} = \begin{cases} 0, & t < T-h, \\ e_\star G_T/G_h, & t \ge T-h. \end{cases} \tag{56}$$

Let $h_0^{\mathrm{F}} = 0$ and $h_{t+1}^{\mathrm{F}} = c_t h_t^{\mathrm{F}} + e_\star - u_t^{\mathrm{F}}$. Before the final window, the weighted shift at time $T-h$ is

$$\left(\prod_{v=T-h}^{T-1} c_v\right) h_{T-h}^{\mathrm{F}} = e_\star \sum_{k=h}^{T-1} Q_k = e_\star(G_T - G_h). \tag{57}$$

For $T-h \le m \le T$, put $w_t = \prod_{v=t+1}^{T-1} c_v$. Unrolling the recursion through the window gives

$$\begin{aligned}\left(\prod_{v=m}^{T-1} c_v\right) h_m^{\mathrm{F}} &= e_\star(G_T - G_h) + \left(e_\star - \frac{e_\star G_T}{G_h}\right)\sum_{t=T-h}^{m-1} w_t \\ &= e_\star(G_T - G_h)\left(1 - \frac{\sum_{t=T-h}^{m-1} w_t}{G_h}\right) \ge 0.\end{aligned} \tag{58}$$

The inequality holds because the partial sum is between 0 and $\sum_{t=T-h}^{T-1} w_t = G_h$. At $m = T$, the sum equals G_h, so $h_T^{\mathrm{F}} = 0$. Thus the schedule is feasible. Its cost is

$$\frac{\alpha}{2s^2}\sum_{t<T}(u_t^{\mathrm{F}})^2 = \frac{\alpha e_\star^2}{2s^2} h\left(\frac{G_T}{G_h}\right)^2. \tag{59}$$

For $k \ge h$, multiply equation 54 by q^j and sum over $j = 0, \dots, h-1$:

$$Q_k \sum_{j=0}^{h-1} q^j \le \Gamma q^k G_h. \tag{60}$$

Summing over $k = h, \dots, T-1$ and dividing by G_h gives

$$\frac{G_T}{G_h} \le 1 + \Gamma q^h \frac{1-q^{T-h}}{1-q^h} \le \frac{\Gamma}{1-q^h}. \tag{61}$$

The last inequality uses $\Gamma \ge 1$.

Choose $h = T$ in equation 59 to obtain the finite-horizon bound $\alpha e_\star^2 T/(2s^2)$. For the uniform bound, put $x = 1 - q = 1 - e^{-\gamma}$ and $h_0 = \lceil 1/x \rceil$. If $h_0 \le T$, then $h_0 x \in [1, 2]$ and $q^{h_0} \le e^{-h_0 x}$. Hence

$$h_0 \left(\frac{G_T}{G_{h_0}} \right)^2 \le \frac{\Gamma^2 h_0}{(1 - e^{-h_0 x})^2} = \frac{\Gamma^2}{x} \frac{h_0 x}{(1 - e^{-h_0 x})^2} \le \frac{4\Gamma^2}{x}. \tag{62}$$

For the last inequality, the derivative of $f(y) = y/(1 - e^{-y})^2$ is negative at $y = 1$, positive at $y = 2$, and has only one zero on $[1, 2]$; hence its maximum is attained at an endpoint. Directly, $f(1) < 4$ and $f(2) < 4$. If $h_0 > T$, then $T < h_0 \le 2/x$, and the finite-horizon bound is already at most $4\Gamma^2/x$ because $\Gamma \ge 1$.

It remains to pass from full injection to the actual access pattern. Let $(u_t^{\mathrm{F}}, h_t^{\mathrm{F}})$ be any of the feasible schedules above. Starting from $h_0^I = 0$, define

$$u_t^I = \min\{u_t^{\mathrm{F}}, c_t h_t^I + a_t I_t\}, \qquad h_{t+1}^I = c_t h_t^I + a_t I_t - u_t^I. \tag{63}$$

Equivalently, $h_{t+1}^I = [c_t h_t^I + a_t I_t - u_t^{\mathrm{F}}]_+$. We prove by induction that $0 \le h_t^I \le h_t^{\mathrm{F}}$. The claim is true at $t = 0$. If it holds at t, then

$$c_t h_t^I + a_t I_t \le c_t h_t^{\mathrm{F}} + e_\star,$$

and therefore

$$h_{t+1}^I = [c_t h_t^I + a_t I_t - u_t^{\mathrm{F}}]_+ \le [c_t h_t^{\mathrm{F}} + e_\star - u_t^{\mathrm{F}}]_+ = h_{t+1}^{\mathrm{F}}.$$

Thus $h_T^I = 0$, and $u_t^I \le u_t^{\mathrm{F}}$ makes the sparse schedule no more expensive than the full schedule. The immediate allocation $u_t = a_t I_t$ gives the additional bound

$$\mathcal{C}_\alpha(I) \le \frac{\alpha e_\star^2}{2s^2} \sum_{t<T} I_t \le \frac{\alpha e_\star^2 T}{2s^2}.$$

Taking the smaller of the two bounds and substituting $s = \eta\sigma$ proves equation 52. The final-model divergence is bounded by the causal certificate from Theorem 1. □

B Proof of the Adaptive Privacy Guarantee

Theorem 6 (Privacy of certification-adaptive training)**.** *Under the public conditions of the monitor rule in Section 5, for every $\alpha > 1$ the joint output $(Z, \theta_{\widehat{T}(Z)})$ satisfies*

$$D_\alpha\big(\mathcal{M}_{\mathrm{adapt}}^{\mathrm{joint}}(D) \| \mathcal{M}_{\mathrm{adapt}}^{\mathrm{joint}}(D')\big) \le \frac{\alpha|\mathcal{Q}|}{2\sigma_{\mathrm{m}}^2} \max_{t \in \mathcal{Q}} \left(\frac{C + \kappa_t R_t}{b} \right)^2 + \frac{\alpha \bar{e}_\star^2}{2\sigma^2} \min\left\{ T, \frac{4e^{2B_\star}}{1 - e^{-\gamma}} \right\}. \tag{64}$$

When $\mathcal{Q} = \varnothing$, the first term is defined to be zero.

Proof. For a checkpoint $t \in \mathcal{Q}$, let $d_t = (C + \kappa_t R_t) I_t / b$. The reverse triangle inequality and the fixed-slot adjacency give

$$|q_t(D) - q_t(D')| = \Big| \|\widehat{g}_t^D(\phi_0)\| - \|\widehat{g}_t^{D'}(\phi_0)\| \Big| \le \|\widehat{g}_t^D(\phi_0) - \widehat{g}_t^{D'}(\phi_0)\| \le d_t. \tag{65}$$

Let $Y_t = q_t(D) + \xi_t$ with $\xi_t \sim \mathcal{N}(0, \sigma_{\mathrm{m}}^2)$, and define the corresponding Y_t' for D'. The exact Gaussian RDP formula gives

$$D_\alpha(\mathcal{N}(q_t(D), \sigma_{\mathrm{m}}^2) \| \mathcal{N}(q_t(D'), \sigma_{\mathrm{m}}^2)) = \frac{\alpha (q_t(D) - q_t(D'))^2}{2\sigma_{\mathrm{m}}^2} \le \frac{\alpha d_t^2}{2\sigma_{\mathrm{m}}^2}. \tag{66}$$

The threshold bit $Z_t = \mathbf{1}\{Y_t \ge \omega\}$ is deterministic post-processing of Y_t, so it has no larger RDP cost. The monitor noises are independent across checkpoints, hence the transcript cost is the sum of the per-checkpoint costs:

$$\rho_{\mathrm{m}} \le \frac{\alpha}{2\sigma_{\mathrm{m}}^2} \sum_{t \in \mathcal{Q}} d_t^2 \le \frac{\alpha |\mathcal{Q}|}{2\sigma_{\mathrm{m}}^2} \max_{t \in \mathcal{Q}} \left(\frac{C + \kappa_t R_t}{b} \right)^2. \tag{67}$$

The last expression is interpreted as zero when $\mathcal{Q}$ is empty.

It remains to bound the training law conditional on a transcript. Fix a bit vector z and let $\widehat{T}(z)$ be the stopping time produced by the controller. The candidate anchor schedule and the stopping rule are functions of public quantities and z only, so both neighboring runs use the same branch and the same deterministic stopping index $\widehat{T}(z)$. The training noises are independent of the monitor noises and have the same law in both runs.

We first verify the interval condition on this branch. The recurrence in Section 5.1 satisfies

$$H_t = \max\left\{0, \max_{0\le j<t}\left[\sum_{v=j}^{t-1}\log c_v + \gamma(t-j)\right]\right\}. \tag{68}$$

This identity follows by induction: it is true for $t=0$; replacing t by $t+1$ adds the new term $\log c_t+\gamma$ to the maximization and retains the previous terms through the additive recurrence followed by the maximum with zero. On every completed update, the gate requires $\eta L_t \le \exp(B_\star - H_t - \gamma)$. Since $\lambda_t \ge \lambda_t^{\mathrm{safe}}$ and $0 \le \lambda_t \le 1/\eta$, we have

$$c_t = 1 - \eta\lambda_t + \eta L_t \le \exp(B_\star - H_t - \gamma). \tag{69}$$

Consequently $H_{t+1} \le B_\star$, and the identity above implies

$$\prod_{v=j}^{k-1} c_v \le \exp(B_\star - \gamma(k-j)) \quad \text{for every } 0 \le j < k \le \widehat{T}(z).$$

Thus every fixed transcript branch satisfies the interval condition with $B = B_\star$ and γ.

If $\widehat{T}(z) = 0$, both runs release the public initialization and the conditional training divergence is zero. Assume therefore $\widehat{T}(z) \ge 1$. For this branch, the update-space map discrepancy satisfies $a_t I_t = \eta\bar{a}_t I_t$ with $\bar{a}_t = (C + \kappa_t R_t)/b$, so its envelope is at most $e_\star = \eta\bar{e}_\star$. Applying the epoch-independent theorem to the prefix of length $\widehat{T}(z)$ gives

$$D_\alpha(P_z^D \| P_z^{D'}) \le \rho_{\mathrm{train}} := \frac{\alpha\bar{e}_\star^2}{2\sigma^2}\min\left\{T, \frac{4e^{2B_\star}}{1-e^{-\gamma}}\right\}. \tag{70}$$

We used the declared horizon T to obtain a bound uniform over all stopping times $\widehat{T}(z) \le T$.

Let $p_D(z)$ and $p_{D'}(z)$ denote the probabilities of transcript z, and let $P_z^D, P_z^{D'}$ be the conditional laws of the final model on that branch. The joint laws factor as

$$\mu_D(z, \mathrm{d}\theta) = p_D(z)P_z^D(\mathrm{d}\theta), \qquad \mu_{D'}(z, \mathrm{d}\theta) = p_{D'}(z)P_z^{D'}(\mathrm{d}\theta).$$

For discrete z, direct substitution into the definition of Rényi divergence gives

$$\begin{aligned}\exp((\alpha-1)D_\alpha(\mu_D\|\mu_{D'})) &= \sum_z p_D(z)^\alpha p_{D'}(z)^{1-\alpha}\exp\Big((\alpha-1)D_\alpha(P_z^D\|P_z^{D'})\Big)\\ &\le e^{(\alpha-1)\rho_{\mathrm{train}}}\sum_z p_D(z)^\alpha p_{D'}(z)^{1-\alpha}.\end{aligned} \tag{71}$$

The remaining sum is the RDP moment of the monitor transcript and is at most $e^{(\alpha-1)\rho_{\mathrm{m}}}$. Taking logarithms and dividing by $\alpha-1$ yields $D_\alpha(\mu_D\|\mu_{D'}) \le \rho_{\mathrm{m}} + \rho_{\mathrm{train}}$, which is exactly equation 64. Releasing only $\theta_{\widehat{T}(Z)}$ discards Z and is therefore post-processing. □

C Detailed Experimental Results

This appendix reports the complete experiments. The privacy analysis uses the exact Gaussian-smoothed expectation. The utility implementation uses one antithetic probe pair, so its utility values describe the finite-probe implementation rather than a utility guarantee for the exact-expectation mechanism.

Epochs	Causal certificate	Epoch envelope	Step RDP	Chien–Li 2025
1	0.4124	0.4124	0.4124	0.8321
20	1.8249	1.9009	1.9009	3.9469
50	2.5048	3.0722	3.0722	6.5073
100	2.7117	4.0000	4.4511	9.6280
200	2.7309	4.0000	6.5073	14.4663
500	2.7310	4.0000	10.9560	25.5411

Table 1: Complete final-model privacy-loss sweep at $\delta = 10^{-5}$ for the fixed noise scale.

Epochs	Public certification check (ms)	Causal accountant (ms)
20	0.0075	0.0417
100	0.0247	0.1856
200	0.0699	0.4075
500	0.1090	0.9445

Table 2: Complete runtime sweep on one CPU thread.

C.1 Experimental setting

All utility runs start from the same public full-model fine-tuning of the development split and use fixed-slot add/remove adjacency with $n = b = 8192$, $\delta = 10^{-5}$, no subsampling amplification, and zero gradient for an empty slot. The public task-fine-tuned initialization has test accuracy 0.7523 on SST-2, 0.8632 on AG News, 0.8777 on CIFAR-10, and 0.5436 on CIFAR-100. The smoothed anchor uses Gaussian smoothing at the clipping scale and a proximal strength accepted by the public certification check. DP-SGD and DP-Adam use step-wise RDP accounting. We also report runs that start from the original pretrained checkpoint; those runs are separate from the public task-fine-tuned initialization.

C.2 Privacy certificate and computation

Table 1 reports every privacy-loss point in the fixed-noise sweep. The causal certificate rises from 0.4124 at one epoch to 2.7310 at 500 epochs, while step-wise RDP grows from 0.4124 to 10.9560. The epoch envelope reaches 4.0 at 100 epochs and remains there. The Chien–Li bound grows to 25.5411 at 500 epochs. Thus the causal certificate is nearly constant over the long horizon in this certified schedule. Table 2 gives the complete runtime sweep: at 500 epochs, the public certification check takes 0.109 ms and the causal accountant takes 0.9445 ms on one CPU thread.

C.3 Utility after public task fine tuning

Table 3 gives all four privacy budgets for the four task-fine-tuned datasets. The smoothed anchor is highest at every listed budget on SST-2. DP-SGD is highest on AG News, while DP-Adam is highest on CIFAR-10 and CIFAR-100. The smoothed anchor remains close to the public initialization on the two CIFAR datasets. Figure 2 shows all four utility curves.

Dataset	Method	$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$
SST-2	Smoothed anchor	**0.7534**	**0.7557**	**0.7546**	**0.7546**
	DP-SGD	0.7179	0.7489	0.7328	0.7339
	DP-Adam	0.7282	0.7339	0.7397	0.7500
AG News	Smoothed anchor	0.8641	0.8638	0.8628	0.8624
	DP-SGD	**0.8667**	**0.8736**	**0.8754**	**0.8754**
	DP-Adam	0.8638	0.8659	0.8696	0.8721
CIFAR-10	Smoothed anchor	0.8776	0.8777	0.8777	0.8777
	DP-SGD	0.8803	0.8842	0.8873	0.8978
	DP-Adam	**0.8830**	**0.8869**	**0.8947**	**0.9006**
CIFAR-100	Smoothed anchor	0.5437	0.5437	0.5437	0.5437
	DP-SGD	0.5359	0.5500	0.5645	0.5855
	DP-Adam	**0.5488**	**0.5620**	**0.5779**	**0.5907**

Table 3: Test accuracy after a shared public task-fine-tuned initialization. The smoothed-anchor column uses a finite-probe Gaussian-smoothed update and its causal certificate; DP-SGD and DP-Adam use step-wise Rényi composition. $\delta = 10^{-5}$. Highest accuracy is underlined.

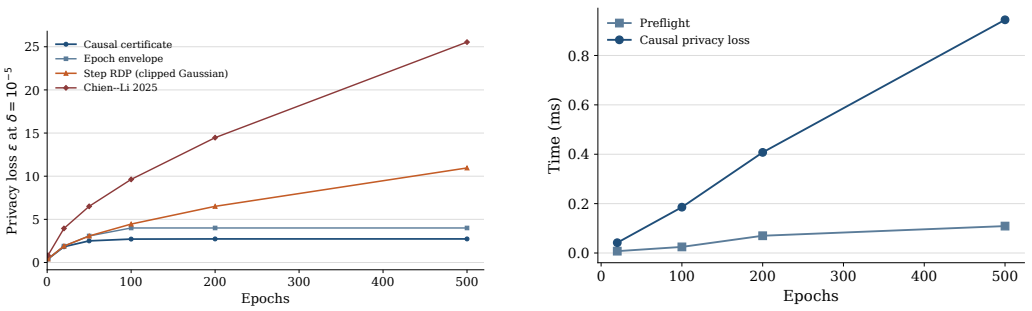


Figure 1: Complete privacy and runtime curves. The left panel shows final-model privacy loss versus epochs; the right panel shows the public certification check and causal accountant runtime.

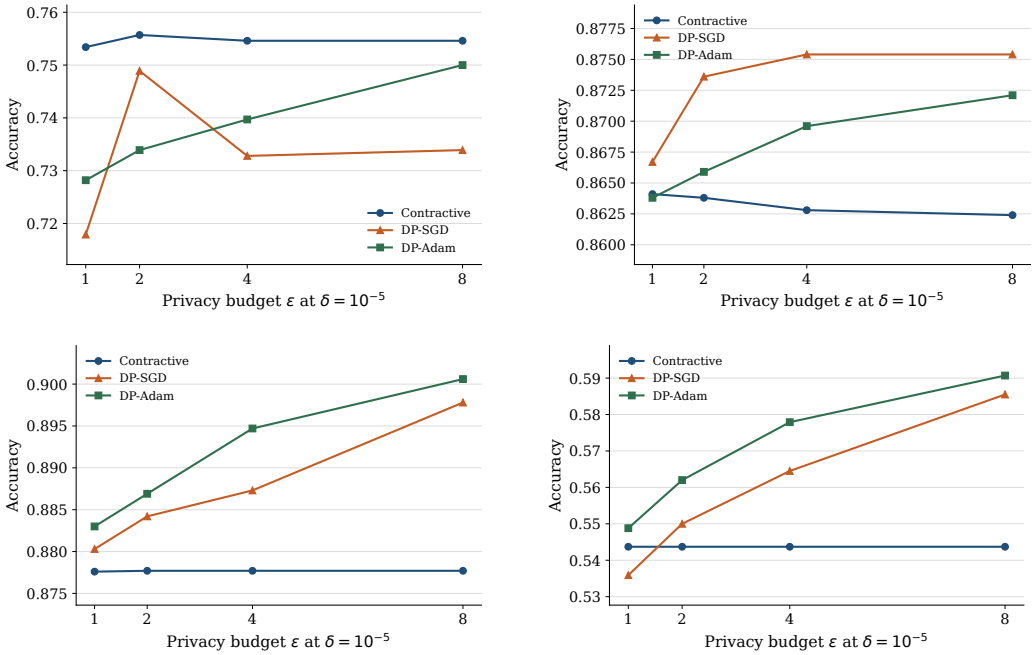


Figure 2: Utility after the shared public task-fine-tuned initialization for SST-2, AG News, CIFAR-10, and CIFAR-100. Each curve uses the privacy budget reported by its own accountant.

C.4 Utility from the original pretrained checkpoint

The package also contains runs that start from the original pretrained checkpoint rather than the public task-fine-tuned initialization. These results are reported separately because the certificate does not cover the public fine-tuning stage. On SST-2, the contractive anchor stays near chance, between 0.5092 and 0.5172, whereas DP-SGD and DP-Adam reach about 0.60 over most budgets. On CIFAR-10, DP-Adam is highest at every listed budget and reaches 0.7548 at $\varepsilon = 10$; the proximal anchor reaches 0.6041 and DP-SGD reaches 0.5823 at the same budget. Tables 4 and 5 report every row, and Figure 3 shows the corresponding curves.

ε	DP-SGD	Weak anchor	Contractive anchor	DP-Adam
1	0.5757	0.5745	0.5172	0.5757
2	0.5940	0.5837	0.5149	0.5940
3	0.5975	0.5917	0.5138	0.5975
4	0.5952	0.5929	0.5138	0.5952
5	0.5952	0.5940	0.5149	0.5952
6	0.5963	0.5940	0.5149	0.5963
7	0.5963	0.5940	0.5149	0.5963
8	0.5952	0.5940	0.5149	0.5952
9	0.5952	0.5940	0.5092	0.5092
10	0.5092	0.5092	0.5092	0.5092

Table 4: SST-2 utility from the original pretrained checkpoint.

ε	DP-SGD	DP-Adam	Proximal anchor
1	0.1193	0.2610	0.1198
2	0.2088	0.4233	0.2063
4	0.4624	0.6001	0.4606
6	0.5915	0.6769	0.5911
8	0.6405	0.7221	0.6336
10	0.5823	0.7548	0.6041

Table 5: CIFAR-10 utility from the original pretrained checkpoint.

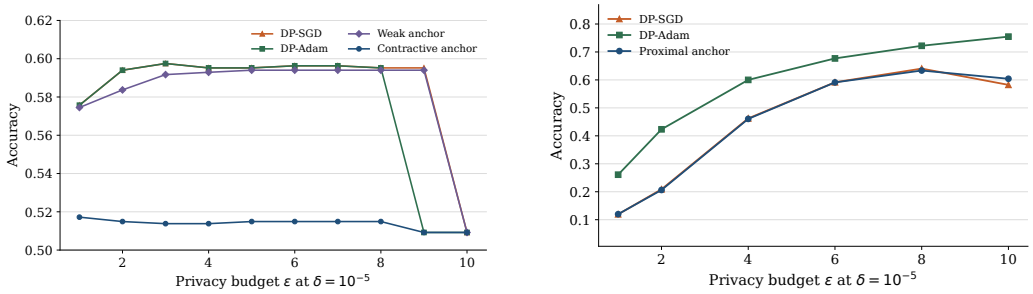


Figure 3: Utility from the original pretrained checkpoint on SST-2 and CIFAR-10.