AUDITING OF UNLEARNING ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating whether an unlearning algorithm has truly removed the influence of its forget set remains an open challenge. We introduce a hypothesis-testing auditor that uses membership-inference attacks to compute high-confidence empirical lower bounds on the zero-concentrated differential privacy (zCDP) parameter ρ of certified unlearning, with a direct audit of the Gaussian differential privacy parameter μ as well; rejecting a candidate ρ establishes that a claimed certificate cannot hold. Our bounds are computed under a black-box threat model in which the adversary additionally observes the initialization and the per-epoch shuffling order. On CIFAR-100, certified methods such as model clipping (Koloskova et al., 2025) and rewind-to-delete (Mu & Klabjan, 2025) yield small lower bounds consistent with their stated guarantees, whereas uncertified methods—including Hessian-based unlearning (Zhang et al., 2025), gradient difference, gradient ascent, retain fine-tuning, and distillation-based forgetting methods such as BadTeacher (Chundawat et al., 2023), SCRUB Kurmanji et al. (2023), DELETE (Zhou et al., 2025)—often yield large lower bounds, revealing substantial residual influence of the forget set. Hessian-based unlearning is certified only for convex losses, and that certification does not transfer to the nonconvex networks we study. We further audit uncertified methods on Shakespeare and scale our evaluation to Llama-3.2-1B (Grattafiori et al., 2024) on TOFU, a dataset of fictitious authors (Maini et al., 2024). Overall, our auditor provides a practical tool for refuting unlearning claims across model architectures, datasets, and scales.

1 INTRODUCTION

Machine unlearning is the task of removing the influence of specific training samples from an already-trained model, ideally yielding a model that behaves as if those samples had never been seen during training Cao & Yang (2015). Simply deleting a data point from the training set does not achieve this: its information remains embedded in the learned parameters and can often be recovered by an adversary with query access, e.g. via membership inference attacks Shokri et al. (2016); Carlini et al. (2022). This motivates *exact* unlearning, which produces a model with the same distribution as one retrained from scratch on the retained data Bourtoule et al. (2021); Ginart et al. (2019), and *approximate* unlearning, which relaxes this requirement in exchange for efficiency Guo et al. (2020); Sekhari et al. (2021); Neel et al. (2021); Golatkar et al. (2020); Kurmanji et al. (2023).

While exact unlearning offers the strongest guarantee, it is often impractical at scale. Approximate methods instead aim to make the distribution of an unlearned model close to that of a reference model trained without the forget set. Some methods are heuristic, including Fisher/Hessian-based scrubbing Golatkar et al. (2020), gradient ascent, retain fine-tuning, interleaved ascent–descent schemes and distillation-based unlearning methods such BadTeacher Chundawat et al. (2023), DELETE Zhou et al. (2025) and SCRUB Kurmanji et al. (2023). Certified methods give a distributional guarantee, but most require convexity or strong convexity Guo et al. (2020); Sekhari et al. (2021); Neel et al. (2021); Qiao et al. (2025); Zhang et al. (2025); only a few recent methods address genuinely nonconvex losses Koloskova et al. (2025); Mu & Klabjan (2025); Chien et al. (2024); Chourasia & Shah (2023).

We study certification through ρ zero-concentrated differential privacy (zCDP) (Bun & Steinke, 2016) and also discuss other notions of privacy such as $(\varepsilon, 0)$ indistinguishability Dwork & Roth (2014) and Gaussian-DP Dong et al. (2022) in Appendix C. Our goal is to use an empirical audit to reject a claimed level of privacy, in the spirit of differential-privacy auditing Jagielski et al. (2020); Nasr et al. (2023); Steinke et al. (2023).

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.

While our audit draws on differential-privacy auditing, the two settings differ in an important way: DP guarantees indistinguishability under a change in a single data point, whereas certified unlearning gives the analogous guarantee with respect to the deletion of an *arbitrary subset* of forget points. Unlearning maybe viewed closer to group privacy Dwork & Roth (2014) with respect to the forget set and while a group private algorithm satisfies standard instinguishability with respect to one point for the same privacy parameter, the reverse does not hold true unless scaled by appropriate constants. Off-the-shelf DP auditors therefore only certify the weaker neighbouring-by-one guarantee and yield loose privacy lower bounds in the unlearning setting. Our meta-algorithm Algorithm 1 is structurally inspired by the one-run auditor of Steinke et al. (2023), but exploiting the stronger subset-level guarantee requires new bounds tailored to unlearning.

- **Threat model.** We adopt a strong adversary with black-box knowledge of the learning algorithm, the unlearning algorithm, the data-loading shuffler, and the initial model weights at the start of training.
- **Auditor.** Leveraging the definition of ρ-zCDP unlearning with respect to arbitrary forget subsets, we design a hypothesis test that rejects the null hypothesis $\rho < \rho_{\mathrm{LB}}$ with controlled Type I error (see Lemma 4.2). Other notions of privacy are discussed in Appendix C.
- **Empirical evaluation.** We develop two membership-inference instantiations in Section 5 and evaluate them on certified and heuristic unlearning methods using CIFAR-100 and Shakespeare, as well as on the larger Llama-3.2-1B language model Grattafiori et al. (2024) using the synthetic TOFU author dataset Maini et al. (2024), and find that forget-set accuracy does not track the audited bound in either direction (Section 7.3).

2 Related work

Empirically lower-bounding the privacy parameter ε of an (ε, δ)-DP mechanism Jagielski et al. (2020); Nasr et al. (2021; 2023); Annamalai & De Cristofaro (2024) proceeds by planting canaries in a pair of neighbouring datasets and converting distinguishing accuracy into a bound via the hypothesis-testing view of DP Kairouz et al. (2015); Wasserman & Zhou (2010); Steinke et al. (2023); Mahloujifar et al. (2025) reduce the cost to a single training run for (ε, δ)-DP and f-DP by auditing many canaries jointly. We borrow that joint-canary idea, but the transfer to unlearning is not direct: a subset-level guarantee yields weaker per-canary signal than the single-point neighbouring relation, so we aggregate across runs and convert the statistic into a lower bound ρ_{LB} on the zCDP certified-unlearning parameter—which requires a reduction through the common retraining reference and a Rényi change-of-measure argument. To tighten the bound we replace loss-difference scoring with a LiRA-style Carlini et al. (2022) per-canary likelihood-ratio test on logit-scaled confidences.

A separate line audits unlearning directly: backdoor-based verification Sommer et al. (2022); Zhang et al. (2024a), output-difference audits Wang et al. (2025b), and sample-level completeness scores Wang et al. (2025a); Triantafillou et al. (2024); Gu et al. (2025). Membership inference itself runs from Shokri et al. (2016); Yeom et al. (2018) to likelihood-ratio attacks Carlini et al. (2022); Zarifzadeh et al. (2024), and its unlearning variants Chen et al. (2021); Bertran et al. (2024) exploit the *difference* between the original and unlearned models. Our audit differs from both groups in what it assumes and what it returns: it queries only the unlearned model, not the pre-unlearning model as well, and it reports an algorithm-level bound comparable against a claimed certificate rather than a per-example completeness score. See Appendix A for a more detailed comparison.

3 Problem setup

3.1 Unlearning and the reference distribution

Let $\mathcal{A}$ be a randomized training algorithm that maps a dataset $\mathcal{D}$ to a model $\mathcal{A}(\mathcal{D})$. Given a forget set $\mathcal{D}_f \subseteq \mathcal{D}$, write $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{D}_f$ for the retain set. An unlearning algorithm $\mathcal{U}$ receives the trained model, the full dataset, and the forget set, and returns $f_u := \mathcal{U}\big(\mathcal{A}(\mathcal{D}), \mathcal{D}, \mathcal{D}_f\big)$ Its output should be close in distribution to that of a reference algorithm $\overline{\mathcal{A}}$ run without the forget set, while being cheaper than training the reference model from scratch.

3.2 zCDP-CERTIFIED UNLEARNING

For probability measures P, Q on a common measurable space and an order $\alpha > 1$, the Rényi divergence is

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

with $D_\alpha(P\|Q) = +\infty$ when P is not absolutely continuous with respect to Q. Zero-concentrated differential privacy (zCDP) requires the linear all-orders bound $D_\alpha \leq \rho\alpha$ (Bun & Steinke, 2016); we use it in the bidirectional certified-unlearning comparison below. Rényi DP and its relation to approximate DP are developed by Mironov (2017).

3.3 THREAT MODEL

We consider a black-box adversary that additionally observes (i) the initial model x_0 at the start of training and (ii) the shuffling order π of the training points within each epoch. It does not observe the remaining randomness used during training or unlearning. Because (x_0, π) are public to the adversary, the certified-unlearning comparison must hold conditional on them, and the reference algorithm may depend on them. Let $\mathcal{A}_{x_0,\pi}(\mathcal{D})$ denote training initialized at x_0 with shuffling order π. We define unlearning as indistinguishability from a retain-set reference algorithm $\overline{\mathcal{A}}_{x_0,\pi}$ (Koloskova et al., 2025). Ordinary retraining is recovered by taking $\overline{\mathcal{A}}_{x_0,\pi} = \mathcal{A}_{x_0,\pi}$.

Definition 3.1 (ρ-zCDP-certified unlearning under our threat model)**.** We say that $\mathcal{U}$ is ρ-zCDP-certified for $\mathcal{A}$ under our threat model if, for every x_0 and π, there exists a reference algorithm $\overline{\mathcal{A}}_{x_0,\pi}$ such that, for every forget set $\mathcal{D}_f \subseteq \mathcal{D}$ and every $\alpha > 1$, the probability distributions P_u and P_r of $\mathcal{U}\big(\mathcal{A}_{x_0,\pi}(\mathcal{D}), \mathcal{D}, \mathcal{D}_f\big)$ and $\overline{\mathcal{A}}_{x_0,\pi}(\mathcal{D}_r)$ respectively satisfy

$$D_\alpha(P_u\|P_r) \leq \rho\alpha \qquad \text{and} \qquad D_\alpha(P_r\|P_u) \leq \rho\alpha.$$

The two directions are explicit because unlearned and reference outputs do not arise from a symmetric neighbouring-dataset relation. The scalar ρ controls the entire curve $\varepsilon_\alpha = \rho\alpha$, and exact unlearning is $\rho = 0$. Gaussian-DP and $(\varepsilon, 0)$-DP formulations are deferred to Appendix C.

Our auditor returns a data-dependent lower bound ρ_{LB} on ρ. At confidence level $1 - \zeta$, it tests the composite null $H_0 : \rho \leq \rho_0$. Thus, rejecting a candidate ρ_0 establishes that the claimed zCDP guarantee is too small. Appendix C.3 gives a monotone rescaling $\varepsilon^{(c)}(\rho)$ onto the more familiar (ε, δ) scale; it is interpretive only, not a confidence bound.

4 THE AUDITING ALGORITHM

We now describe the meta-algorithm underlying our audit; pseudocode appears in Algorithm 1. The auditor has black-box query access to $\mathcal{A}$ and $\mathcal{U}$. It partitions a candidate forget pool into m batches and, in each run $\ell \in [L]$, samples a balanced sign vector

$$\mathcal{S}_m := \{s \in \{\pm 1\}^m : 0 \leq \#\{j : s_j = +1\} - \#\{j : s_j = -1\} \leq 1\}.$$

The coordinates with $S_j^{(\ell)} = +1$ determine which candidate batches are included in the training set and subsequently forgotten and the co-ordinates with $S_j^{(\ell)} = -1$ were never seen either during training or unlearning. From the resulting unlearned model [1], the auditor returns a prediction $\widehat{S}^{(\ell)} \in \{-1, 0, +1\}^m$ with $r/2$ positive and $r/2$ negative entries.Its overlap score is

$$V^{(\ell)} := \sum_{j=1}^{m} \max\{0, \widehat{S}_j^{(\ell)} S_j^{(\ell)}\} \in \{0, \ldots, r\}.$$

A score above the chance baseline indicates that the unlearned model retains information about which candidate batches were present. The construction extends the many-canary auditing paradigm for differential privacy Steinke et al. (2023); Nasr et al. (2023); Pillutla et al. (2024) to subset-level

[1] As defined in Definition C.5, the auditor observes the initial model x_0 and the data shuffling order π.

unlearning guarantees. For the analysis, fix the dataset, initial model x_0, shuffling order π, and any calibration information generated independently of the evaluation sign vector. The evaluation procedure induces a randomized mechanism

$$\mathcal{M} : \mathcal{S}_m \longrightarrow \{-1, 0, +1\}^m, \qquad \mathcal{M}(S) := \widehat{S}.$$

For $S \in \mathcal{S}_m$, let $\mathcal{D}_f(S)$ be the forget subset encoded by S and define

$$f_u(S) := \mathcal{U}\big(\mathcal{A}(\mathcal{D}_r \cup \mathcal{D}_f(S)), \mathcal{D}_r \cup \mathcal{D}_f(S), \mathcal{D}_f(S)\big), \qquad f_r := \overline{\mathcal{A}}(\mathcal{D}_r).$$

The mechanism $\mathcal{M}(S)$ accesses the hidden vector S only through $f_u(S)$.

4.1 From zCDP-certified unlearning to local Rényi DP

Certified unlearning compares each $f_u(S)$ with the common reference f_r, whereas the overlap audit requires an all-pairs comparison between $\mathcal{M}(S_1)$ and $\mathcal{M}(S_2)$. Rényi divergence does not satisfy an ordinary triangle inequality, so the following reduction uses its weak triangle inequality.[2]

Lemma 4.1 (zCDP unlearning implies local Rényi DP)**.** *Suppose $(\mathcal{A}, \mathcal{U})$ is ρ-zCDP-certified under our threat model (Definition 3.1). Then, for every $\gamma > 1$, the induced audit mechanism $\mathcal{M}$ satisfies*

$$D_\gamma(\mathrm{Law}(\mathcal{M}(S_1)) \| \mathrm{Law}(\mathcal{M}(S_2))) \le \varepsilon_\gamma^{\mathrm{loc}}(\rho) := 4\rho\gamma$$

for every $S_1, S_2 \in \mathcal{S}_m$.

Following the all-input-pairs formulation of local DP Bebensee (2019), we call $\gamma \mapsto \varepsilon_\gamma^{\mathrm{loc}}(\rho)$ the *local Rényi DP curve* of $\mathcal{M}$ on $\mathcal{S}_m$. This proof follows identically to the transitivity argument used for group privacy in Bun & Steinke (2016, Proposition 9) and reproduced in Appendix B.1.

4.2 The Rényi-DP overlap audit

In this section, we use a hypothesis-testing framework to translate the observed overlap scores $\{V^{(\ell)}\}_{\ell=1}^L$ into lower bounds on the local Rényi DP curve of $\mathcal{M}$. Let

$$\mathcal{T} := \{t \in \{-1, 0, +1\}^m : \#\{j : t_j = +1\} = \#\{j : t_j = -1\} = r/2\}, \qquad M' := |\mathcal{S}_m| = \binom{m}{\lfloor m/2 \rfloor}.$$

For $u \in \{0, \dots, r\}$, define the chance-overlap mass

$$\pi(u) := \frac{1}{M'} \sum_{\substack{\alpha_1, \alpha_2 \in \{0, \dots, r/2\} \\ \alpha_1 + \alpha_2 = u}} \binom{m-r}{\lceil (m-r)/2 \rceil - (\alpha_1 - \alpha_2)} \binom{r/2}{\alpha_1} \binom{r/2}{\alpha_2},$$

and set $\overline{\Pi}(v) := \sum_{u=\lceil v \rceil}^r \pi(u)$. This is the overlap distribution when a balanced S is independent of a prediction in $\mathcal{T}$.

Lemma 4.2 (Rényi-DP overlap-score audit)**.** *Let $\mathcal{M} : \mathcal{S}_m \to \{-1, 0, +1\}^m$ have a local Rényi-DP curve $\{\varepsilon_\gamma^{\mathrm{loc}}\}_{\gamma > 1}$, and suppose $\mathcal{M}(S) \in \mathcal{T}$ almost surely. Draw $S^{(1)}, \dots, S^{(L)}$ independently and uniformly from $\mathcal{S}_m$, draw the corresponding mechanism outputs independently, and define the overlap scores $V^{(1)}, \dots, V^{(L)}$ as above. Then, for every $v \in \mathbb{R}$,*

$$\Pr\left[\frac{1}{L}\sum_{\ell=1}^L V^{(\ell)} \ge v\right] \le \inf_{\gamma > 1} \exp\left(\frac{\gamma - 1}{\gamma} L \varepsilon_\gamma^{\mathrm{loc}}\right) \left[\inf_{\theta \ge 0} \exp\left(L \log \sum_{u=0}^r e^{\theta u} \pi(u) - \theta L v\right)\right]^{\frac{\gamma-1}{\gamma}}, \tag{1}$$

$$\Pr\Big[\mathrm{Median}\{V^{(\ell)}\}_{\ell=1}^L \ge v\Big] \le \inf_{\gamma > 1} \exp\left(\frac{\gamma - 1}{\gamma} L \varepsilon_\gamma^{\mathrm{loc}}\right) \left[\binom{L}{\lceil L/2 \rceil} \overline{\Pi}(v)^{\lceil L/2 \rceil}\right]^{\frac{\gamma-1}{\gamma}}. \tag{2}$$

[2] For comparison, composing two (ε, δ) comparisons through the common reference gives only $(2\varepsilon, (1 + e^\varepsilon)\delta)$-indistinguishability. This amplified slack is particularly damaging for our auditing Remark C.4.

The proof, based on a change of measure between a matched experiment and an independent null experiment, is deferred to Appendix B. Combining Lemmas 4.1 and 4.2 and substituting $\varepsilon_\gamma^{\mathrm{loc}}(\rho) = 4\rho\gamma$, let $B_\rho(v)$ be the right-hand side of Equation (1) or Equation (2), for the mean or median statistic, and let v_{obs} be its observed value. We define $\rho_{\mathrm{LB}} := \sup\big(\{\rho \geq 0 : B_\rho(v_{\mathrm{obs}}) \leq \zeta\} \cup \{0\}\big)$. We report ρ_{LB} throughout; the interpretive rescaling $\varepsilon^{(c)}(\rho_{\mathrm{LB}})$, which carries no lower-confidence-bound interpretation, is deferred to Appendix C.3. The following corollary ensures that we reject every null $H_0 : \rho = \rho_0$ with $\rho_0 < \rho_{\mathrm{LB}}$ at significance level ζ.

Corollary 4.3 (Monotonicity in the zCDP parameter)**.** *For a fixed observed threshold v, both Rényi bounds $B_\rho(v)$ after the substitution $\varepsilon_\gamma^{\mathrm{loc}} = \varepsilon_\gamma^{\mathrm{loc}}(\rho)$ are non-decreasing in ρ.*

We prove this corollary in Appendix B. A direct audit of a claimed Gaussian-DP parameter, using the GDP trade-off rather than a conversion through Rényi DP, is given in Appendix C.

Remark 4.4 (Auditor design choices). The number of candidate batches m and the reported support size r jointly control power. Larger m makes identifying the balanced sign vector harder and raises the maximum attainable bound, but it also reduces the per-batch signal. The support size r lets the auditor abstain on uncertain batches; empirically the resulting bound often follows an inverted-U as r grows. We study both choices in Appendix F.

5 Two instantiations of the auditor meta-algorithm

We now instantiate the abstract mechanism $\mathcal{M}$ from Algorithm 1 in two ways, and contrast both with a pairwise baseline adapted from DP auditing. Both instantiations use repeated runs of training and unlearning to calibrate a likelihood-based predictor, but differ in what they predict from the final unlearned model. Instantiation I scores each candidate forget batch independently and is suitable when m is too large to enumerate $\mathcal{S}_m$; Instantiation II treats each balanced sign vector as a hypothesis and exploits cross-batch correlations, but is practical only when $|\mathcal{S}_m|$ is enumerable.

For a model f, write $f(x)_y$ for the probability assigned to label y on input x and $\phi(p) := \log\Big(\frac{p}{1-p}\Big)$ for the logit transform. We write $\mathcal{N}(s; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(s-\mu)^2}{2\sigma^2}\Big)$ for the Gaussian likelihood. Modelling attack scores as Gaussian is only a predictor-design choice; it is unrelated to the auditing guarantees in Lemma 4.2, which make no distributional assumption on these scores. Appendix E.1 compares the resulting audit lower bounds when the auditor's per-example score is this logit transform versus the cross-entropy loss.

Instantiation I: Batchwise inclusion/exclusion prediction. When m is large, enumerating $\mathcal{S}_m$ is infeasible, so we score each batch independently. For every forget example, we fit two Gaussians to the logit scores collected across Γ independent calibration runs $\mathcal{U}(\mathcal{A}(\mathcal{D}(S)), \mathcal{D}, \mathcal{D}_f(S))$ with $S \sim \mathrm{Unif}(\mathcal{S}_m)$: an *in*-distribution conditioned on $S_j = +1$ and an *out*-distribution conditioned on $S_j = -1$, in the spirit of LiRA Carlini et al. (2022).[3] At evaluation, for each batch we sum per-point log-likelihoods under the in- and out-distributions separately and take their difference to obtain a batch score Λ_j. We then predict $+1$ for the top $r/2$ batches by Λ_j, -1 for the bottom $r/2$, and 0 for the rest.[4] Calibration runs and evaluation runs are drawn independently, which is required for the audit's lower bound to be valid. See Algorithm 2 for pseudocode.

Instantiation II: Joint sign-vector prediction. When $|\mathcal{S}_m|$ is enumerable, we treat each candidate sign vector as a hypothesis, capturing cross-batch correlations that batchwise scoring ignores. During calibration, for every candidate $\widetilde{S} \in \mathcal{S}_m$ and every forget example (x, y), we fit a Gaussian to the logit scores collected from Γ independent runs of $\mathcal{U}(\mathcal{A}(\mathcal{D}(\widetilde{S})), \mathcal{D}, \mathcal{D}_f(\widetilde{S}))$. At evaluation, we score each candidate by the cumulative log-likelihood of the observed scores under its fitted Gaussians and return the highest-scoring candidate. Since the auditor commits to a full sign vector, $r = m$ here. See Algorithm 3 for pseudocode.

[3] See Appendix E.1 for a comparison of logit score versus cross-entropy loss; the two yield similar values on most datasets.

[4] We consider only even r for our audits.

Pairwise auditor (baseline). As a baseline we adapt the pairwise distinguishability auditor of Nasr et al. (2021), which distinguishes two unlearned models trained on different forget sets. The original auditor converts true/false positive rates into a lower bound under (ε, δ)-indistinguishability using Clopper–Pearson confidence bounds. We derive its analogue for ρ-zCDP-certified unlearning using a Rényi change-of-measure inequality in Appendix D.4.

6 Datasets and auditor setup

CIFAR-100 (Krizhevsky, 2009). We train a `TinyNetCIFAR100` model to $\sim 55\%$ validation accuracy. The forget set $\mathcal{D}_f$ contains 10% of training points $(4{,}500)$[5]; we use a *uniform* split (sampled at random) and an *adversarial* split (forget and retain drawn from largely disjoint label classes). Audit batch size is $B = 1$ for uncertified and $B = 750$ for certified algorithms; the rationale and full construction are in Appendices F.4 and H.1.

Shakespeare (Shakespeare, 1997). Following McMahan et al. (2017) we treat each role as a client, subsample 300 roles, and partition at the role level into 33 forget roles ($\sim 10\%$ of training characters) and 267 retain roles. The model is a 2-layer character-level LSTM reaching ~ 0.53 accuracy; we evaluate only uncertified methods, using the batchwise inclusion/exclusion auditor with $m = 400$ candidate batches. Full construction, score definition, and model details are in Appendix H.2.

TOFU dataset of fictitious authors (Maini et al., 2024) Following Maini et al. (2024), we consider TOFU, a dataset comprising 200 fictitious authors with 20 question–answer pairs per author. We designate 20 authors as the candidate forget pool and use the remaining 180 authors as a fixed retain set. The resulting 400 candidate question–answer pairs are partitioned into singleton batches, yielding $m = 400$. In the training run, we fintune Llama-3.2-1B-Instruct Grattafiori et al. (2024) on the training set (retain + sampled forget set) and then apply unlearning on the selected forget set. Full construction, score definition, and model details are in Appendix H.3.

7 Unlearning Algorithms Audited

Our primary analysis applies Lemma 4.2 and reports the zCDP lower bound ρ_{LB} throughout; its interpretive conversion to an (ε, δ) scale is given in Appendix C.3. Mean and median aggregations are compared in Appendix E.2 with confidence parameter is $\zeta = 0.05$ throughout. All bounds are one-sided: a large ρ_{LB} falsifies a claimed guarantee, but a small one does not establish that a method unlearns.

We organise the audited algorithms into two groups: certified algorithms with formal (ε, δ) guarantees (Section 7.1), and heuristic algorithms without such guarantees (Section 7.2). Section 7.3 discusses them in detail, and full hyperparameters and per-method experimental details are deferred to Appendix I. Additional membership inference attacks are reported in Appendix K.

7.1 Certified unlearning algorithms

Gradient clipping, model clipping (Koloskova et al., 2025), R2D (Mu & Klabjan, 2025), and Langevin unlearning (Chien et al., 2024) all admit ρ-zCDP guarantees — stated directly, or derivable. For Langevin unlearning, the learner's limiting distribution to satisfy a Log-Sobolev Inequality (LSI). Model clipping and R2D additionally admit natural μ-GDP certificates. Appendix G derives these consequences from the statements proved in the original papers. We audit the latter two.

Model clipping. The algorithm of Koloskova et al. (2025) interleaves projected noisy gradient steps on an ℓ_2-ball of radius C_2 with a noiseless retain-set fine-tuning phase, the number of noisy steps being set by their Theorem 4.2 to meet the target (ε, δ) guarantee. Sweeping ε, C_2, and σ over five orders of magnitude (up to $\varepsilon = 10^5$), the zCDP audit yields $\rho_{\mathrm{LB}} = 0$ across the entire grid at the prescribed stopping time. Truncating the noisy phase and raising C_2/σ beyond what

[5] Certified-unlearning guarantees typically depend on the full dataset size n_{pt} and the forget-set size m_{pt} (e.g. **?**). In every setting we therefore fix both the retain-set and forget-set sizes. Because the sign vector is balanced (Section 4), each run forgets the same number of points, so all L runs audit a single, fixed guarantee.

Theorem 4.2 admits removes the certificate, and the audit then tracks unlearning as it proceeds: ρ_{LB} falls from roughly 0.7 to 0 over the first 100 steps under both splits (Figure 1a), so the auditor responds to how much noisy updating has been performed rather than to whether a certificate is claimed. Configurations, sweeps, and per-step diagnostics are in Appendices I.1 and I.1.3.

Rewind-to-delete. Rewind-to-delete (R2D) provides certified unlearning in the full-batch setting by adding final-step Gaussian noise calibrated, per Theorem 3.1 of Mu & Klabjan (2025), to make the trained and unlearned models (ε, δ)-indistinguishable. The noise scale, sensitivity, and other choices are detailed in Appendix I.2. The audit is not vacuous here, but it never approaches the certificate: $\rho_{\mathrm{LB}} \approx 0.2$ even at claimed ρ as large as 500, under both splits, so no configuration is falsified (Figure 1b).

7.2 Uncertified unlearning algorithms

We audit four heuristic unlearning algorithms on both CIFAR-100 and Shakespeare: Hessian-based unlearning, gradient difference retraining, gradient ascent, and retain fine-tuning. We additionally audit four class-based distillation methods—BadTeacher, SCRUB, SCRUB+R, and DELETE—on CIFAR-100. All use the batchwise predictor with $L = 10$ evaluation runs; base-training settings, learning-rate schedules, and the per-dataset batching and support sizes are deferred to Appendices J, J.2, J.3 and L.1. The headline TOFU results use $r/m = 0.25$.

Hessian-based unlearning. We audit the Hessian-based unlearning algorithm of Zhang et al. (2025), which performs a Newton-style update on the forget-set gradient followed by addition of Gaussian noise to every parameter. Zhang et al. (2025) prove certification under loss convexity, an assumption violated by the TinyNetCIFAR100 and Shakespeare models used here. In this nonconvex regime the audit returns a large zCDP lower bound under uniform splitting, sharply exposing that the published certification does not transfer beyond its convexity assumptions; the split-dependent behaviour follows the broader pattern of Section 7.3. The implementation details and the final noise sweep are described in Appendix J.1 and Appendix L.3 respectively.

Gradient-based methods. Gradient difference (GD) inserts a λ-weighted forget-gradient step after every q retain steps for the first E_f epochs; gradient ascent (GA) ascends the forget loss for two epochs; retain fine-tuning (Retain FT) never touches $\mathcal{D}_f$. GA and Retain FT then fine-tune on the retain set. The GD bound falls with more frequent and more heavily weighted forget steps (smaller q, larger λ and E_f) (see Table 18). Retain FT is among the most detectable methods on both datasets: never touching the forget set does not, by itself, remove its influence. The GD and GA update rules, the base-training and fine-tuning schedules, and the full GD sweep are in Appendix J.2, Appendix J.3, Appendix J, and Appendix L.2, respectively.

Distillation-based unlearning. We audit four distillation-based unlearners on CIFAR-100: BadTeacher (Chundawat et al., 2023), SCRUB and SCRUB+R (Kurmanji et al., 2023), and DELETE (Zhou et al., 2025). All distil the trained model into a student, differing in the target used on $\mathcal{D}_f$: BadTeacher distils towards a randomly initialised teacher, SCRUB and SCRUB+R maximise divergence from the trained model, and DELETE masks and renormalises the teacher's logits on the forgotten classes. Being class-level methods—uniform splitting would place one class in both $\mathcal{D}_f$ and $\mathcal{D}_r$—they are audited under adversarial splitting only. Full objectives and the rewind step distinguishing SCRUB+R are in Appendix J.5.

Our attack is least successful against BadTeacher ($\rho_{\mathrm{LB}} = 0.5$), followed by SCRUB (10.1) and SCRUB+R (11.4), and most successful against DELETE (89.4), even though DELETE reaches 1% forget accuracy at the best utility ($69/49$). Low forget-set accuracy therefore does not imply a small ρ. We believe this order reflects how each forget objective uses the trained model's predictions. BadTeacher's forget target ignores the trained model's prediction at the same input, and SCRUB and SCRUB+R push away from it; either could remove the per-example confidence patterns the attack exploits. DELETE's target instead keeps the trained model's relative odds among surviving classes, and so could carry those patterns into the student. The bound decays monotonically in the unlearning rate for all four (Appendix L.1).

7.2.1 UNLEARNING ALGORITHMS ON TOFU WITH LLAMA-3.2-1B-INSTRUCT

Beyond GA, GD, and RetainFT, we evaluate two language-model unlearning algorithms: NPO-RT (Zhang et al., 2024b), a saturating forget objective relative to the frozen pre-unlearning model plus a retain term, and SimNPO (Fan et al., 2024), its reference-free, length-normalized variant. Details are in Appendix J.6, and comparisons with TOFU's diagnostics in Appendix M.

GA yields the smallest audited lower bound, but not evidence of successful unlearning: after two epochs, its loss approaches random-token prediction and degradation extends beyond the forget set, indicating collapse. Among methods that avoid collapse, GD yields the smallest bound (still significant $\rho_{\mathrm{LB}} \sim 5.594$), followed by NPO-RT. RetainFT and SimNPO reach the largest bound reportable at the chosen support size, with every reported membership guess correct for both.

7.3 COMPARISON OF LOWER BOUNDS ACROSS METHODS

Results are in Table 1, with sweeps in Appendix L.

Pairwise baseline. With the same $2L = 10$ executions, the pairwise auditor (Section 5) returns $\rho_{\mathrm{LB}} = 0$ for every CIFAR-100 entry: at 5 runs per hypothesis, Clopper–Pearson intervals are too wide even under perfect separation (Appendix D.4). It reaches 10.17 on the convex mechanisms of Section 8, where a large number of runs is permissible.

Certified versus uncertified methods. The audit separates the two groups cleanly. Methods whose certificate holds in the regime they are deployed in—model clipping and R2D—yield bounds orders of magnitude below the claimed parameter ($\rho_{\mathrm{LB}} = 0$ throughout for model clipping, ≈ 0.2 for R2D against ρ up to 500), so neither is falsified. Methods without a valid certificate in this non-convex regime often yield large bounds.

Comparing the uncertified methods. The ordering depends on both the split and the dataset. Under adversarial splitting on CIFAR-100, GD, retain fine-tuning, and gradient ascent reach the same 2.7–2.9% forget accuracy, yet their bounds range from 2.0 to 38.6; GD and the Hessian-based method (2.8) leave substantially less auditable signal than retain fine-tuning and gradient ascent: both of the former compute a correction aimed at the retain-only optimum, whereas gradient ascent degrades forget-set performance without targeting that solution. The Hessian-based method is also by far the most split-sensitive, moving from the smallest bound under adversarial splitting to the largest under uniform (66.9). On the text datasets the ordering differs: gradient ascent gives the smallest bound on both Shakespeare (0) and TOFU (0.009) and retain fine-tuning among the largest, reversing the CIFAR-100 pattern; on TOFU this reflects collapse, on Shakespeare it does not.

Forget accuracy versus the audit. Low forget-set accuracy can coexist with a strong audit signal, as for DELETE and, under adversarial splitting, retain fine-tuning and gradient ascent on CIFAR-100 (above). Conversely, among distillation methods and on TOFU, a low audit bound can accompany model collapse or utility loss (Section 7.2.1 and Appendix L.1).

Uniform vs. adversarial splits. Across both certified and uncertified methods, ρ_{LB} under uniform splitting typically exceeds adversarial (with retain fine-tuning and gradient ascent on CIFAR-100 as mild exceptions). Uniform splits share feature support between retain and forget, leaving residual memorisation Cheng et al. (2026) and exploitable semantic adjacency Ebrahimpour-Boroojeny et al. (2025); adversarial splits do not.

8 UNLEARNING ON CONVEX LOSSES

The certified algorithms audited above yield very loose bounds ($\rho_{\mathrm{LB}} = 0$ for model clipping even at target $\varepsilon = 10^5$, and similarly weak bounds for R2D), raising the question of how much of this gap reflects auditor slack rather than genuine unlearning. We therefore turn to *convex* losses, adopting the stronger certified-unlearning notion of Sekhari et al. (2021) (Definition N.1), which compares the unlearned output directly to the same mechanism invoked with an empty forget set—exactly the pair the auditor observes, so no transitivity step is needed. We audit two mechanisms: the *Perturbed*

Setting	Method	ρ_{LB}
Shakespeare	GD	3.444
	Retain FT	5.398
	Grad. ascent	0
	Hessian	8.763
TOFU	GA	0.009
	GD	5.594
	Retain FT	15.092
	NPO-RT	8.162
	SimNPO	15.092

(a) Shakespeare and TOFU. On Shakespeare, retain/forget accuracy is $\sim$0.56 and test accuracy $\sim$0.53 for nearly every method and setting.

		ρ_{LB}		R/F/T accuracy (%)	
Class	Method	U	A	U	A
Distillation-based	BadTeacher	–	0.5	–	45/8/38
	DELETE	–	89.4	–	69/1/49
	SCRUB	–	10.1	–	49/2/37
	SCRUB+R	–	11.4	–	46/3/34
Gradient	GD	42.5	2.0	92/66/54	92/2.9/51
	Retain FT	37.5	38.6	93/65/54	94/2.8/51
	Grad. ascent	35.7	38.6	93/65/54	94/2.7/51
	Hessian	66.9	2.8	69/67/51	55/9/37

(b) CIFAR-100. U = uniform; A = adversarial. Class-level methods are shown at their best-utility unlearning rate; per-method configurations are in Appendix J and the hyper-parameter sweep in Tables 17 and 18.

Table 1: Audit lower bounds for unlearning on Shakespeare and TOFU (left) and CIFAR-100 (right).

ε	ρ	$\rho_{\mathrm{LB}}^{(p)}$	ρ_{LB}
0.1	0.0017	0.0000	0.0000
2	0.2394	0.0050	0.0015
10	3.032	0.0666	0.0228
50	27.79	0.6147	0.1870
100	65.51	1.416	0.3640
400	322.5	5.778	0.6711
800	685.6	10.16	0.6867

(a) Convex cubic loss (Perturbed Newton).

ε	ρ	$\rho_{\mathrm{LB}}^{(p)}$	ρ_{LB}
0.1	0.0017	0.0000	0.0000
1	0.0754	0.0272	0.0092
2	0.2394	0.1000	0.0334
5	1.051	0.4659	0.1442
10	3.032	1.310	0.3421
20	8.214	3.254	0.5822
100	65.51	10.17	0.6867

(b) Output perturbation.

Table 2: Lower bounds for Perturbed Newton unlearning on the cubic loss (left) and for output perturbation on a linear-regression task (right). ε and ρ are the theoretical upper bounds ($\delta = 10^{-3}$; Definitions N.1 and N.2); $\rho_{\mathrm{LB}}^{(p)}$ and ρ_{LB} are the pairwise and joint sign-vector audits, both direct convex versions carrying no transitivity penalty. Full grids, including the pure-DP audits, are in Tables 22 and 23.

Newton unlearning of Sekhari et al. (2021, Algorithm 1) on a bounded cubic loss in dimension $d = 1$, and a simple *output perturbation* mechanism that clips the trained model and adds Gaussian noise. Both the pairwise auditor and the $m = 2$, $r = 2$ joint sign-vector instantiation of Algorithm 1 use $L = 100{,}000$ evaluation runs per hypothesis, affordable here because a single convex fit is cheap. Pipeline, remaining settings, and a logistic-loss instantiation are in Appendices N and O.

Table 2 shows these direct audits to be substantially tighter than the certified-mechanism audits above. Both saturate once empirical prediction is perfect, attaining the largest value the evaluation budget and confidence level permit: in Table 23 the bound is unchanged from $\varepsilon = 100$ through $\varepsilon = 800$. The logistic loss instead yields a zero lower bound (Appendix N).

9 Conclusion

We introduce a hypothesis-testing auditor that converts membership-inference attacks into high-confidence lower bounds on the zCDP parameter of certified unlearning, for arbitrary forget subsets. Neither certified method we audit is refuted: model clipping yields a zero bound at its prescribed stopping time, and R2D yields bounds far below its claimed parameter. Uncertified methods often yield much larger bounds, and low forget-set accuracy does not imply a small one. Large bounds can also arise when a method is used outside the assumptions of its certificate: the Hessian-based method of Zhang et al. (2025), certified only for convex losses, yields a large bound on nonconvex networks under uniform splitting.

In future work, since our auditor observes only model outputs, extending it to internal representations, which may retain information that output-level metrics miss (Ha et al., 2025; Siddiqui et al., 2025), might yield larger bounds. Tighter analyses of existing certified algorithms, or new algorithms with smaller certified parameters, might also narrow the gap between our lower bounds and the certified parameters from above.

AI USE STATEMENT

We used generative AI tools to assist with experimental design, code development, and the polishing and writing of parts of this paper. Every AI-assisted statement and line of code was reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and software artifacts.

ETHICS STATEMENT

This work was conducted in accordance with the ICLR Code of Ethics, and we are not aware of any ethical violations arising from the research. The work studies privacy and security properties of machine-unlearning algorithms with the goal of improving their transparency and accountability. We report the scope and limitations of our auditing methods so that empirical lower bounds are not misinterpreted as privacy guarantees.

REPRODUCIBILITY STATEMENT

The experimental setups, datasets, unlearning methods, auditor configurations, training and evaluation procedures, hyperparameters, and computational resources are described in detail in the main paper and appendices. Complete algorithmic descriptions and proofs are also provided, and the accompanying code supports reproduction of the reported experiments.

REFERENCES

Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *J. Mach. Learn. Res.*, 18(1):4148–4187, January 2017. ISSN 1532-4435.

Meenatchi Sundaram Muthu Selva Annamalai and Emiliano De Cristofaro. Nearly tight black-box auditing of differentially private machine learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis testing interpretations and rényi differential privacy. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2496–2506, 2020.

Björn Bebensee. Local differential privacy: a tutorial. *arXiv preprint arXiv:1907.11908*, 2019.

Martín Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Reconstruction attacks on machine unlearning: Simple models are vulnerable. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings, Part I, of the 14th International Conference on Theory of Cryptography - Volume 9985*, pp. 635–658, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 9783662536407. doi: 10.1007/978-3-662-53641-4_24. URL https://doi.org/10.1007/978-3-662-53641-4_24.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 463–480, 2015.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles, 2022. URL https://arxiv.org/abs/2112.03570.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 896–911, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3484756. URL https://doi.org/10.1145/3460120.3484756.

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine unlearning under retain-forget entanglement, 2026. URL https://arxiv.org/abs/2603.26569.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: a new perspective of noisy gradient descent for machine unlearning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Rishav Chourasia and Neil Shah. Forget unlearning: towards true data-deletion in machine learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7210–7217, 2023. doi: 10.1609/aaai.v37i6.25879.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):3–37, 2022. doi: 10.1111/rssb.12454.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3e4a38f228427ab819ba7899003a44b1-Abstract-Datasets_and_Benchmarks_Track.html.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Ali Ebrahimpour-Boroojeny, Yian Wang, and Hari Sundaram. On the necessity of output distribution reweighting for effective class unlearning, 2025. URL https://arxiv.org/abs/2506.20893.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning, 2024. URL https://arxiv.org/abs/2410.07163.

Antonio A. Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. *Making AI forget you: data deletion in machine learning*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yuechun Gu, Jiajie He, and Keke Chen. Auditing approximate machine unlearning for differentially private models. In *IEEE International Conference on Data Mining (ICDM)*, 2025.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

SeungBum Ha, Saerom Park, and Sung Whan Yoon. Unlearning's blind spots: Over-unlearning and prototypical relearning attack, 2025. URL https://arxiv.org/abs/2506.01318.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1376–1385. PMLR, 2015.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks, 2025. URL https://arxiv.org/abs/2506.06985.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

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

Siqiao Mu and Diego Klabjan. Rewind-to-delete: Certified machine unlearning for nonconvex functions, 2025. URL https://arxiv.org/abs/2409.09778.

Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary instantiation: Lower bounds for differentially private machine learning, 2021. URL https://arxiv.org/abs/2101.04535.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *USENIX Security Symposium*, pp. 1631–1648, 2023.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory (ALT)*, pp. 931–962, 2021.

Krishna Pillutla, Galen Andrew, Peter Kairouz, H. Brendan McMahan, Alina Oprea, and Sewoong Oh. Unleashing the power of randomization in auditing differentially private ML. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xinbao Qiao, Meng Zhang, Ming Tang, and Ermin Wei. Hessian-free online certified unlearning. In *International Conference on Learning Representations (ICLR)*, 2025.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: algorithms for machine unlearning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

William Shakespeare. *The Complete Works of William Shakespeare*. Project Gutenberg, 1997. URL https://www.gutenberg.org/ebooks/100. Project Gutenberg edition.

Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *CoRR*, abs/1610.05820, 2016. URL http://arxiv.org/abs/1610.05820.

Shoaib Ahmed Siddiqui, Adrian Weller, David Krueger, Gintare Karolina Dziugaite, Michael Curtis Mozer, and Eleni Triantafillou. From dormant to deleted: Tamper-resistant unlearning through weight-space regularization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

David Marco Sommer, Liwei Song, Sameer Wagh, and Prateek Mittal. Athena: Probabilistic verification of machine unlearning. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2022(3):268–290, 2022.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run, 2023. URL https://arxiv.org/abs/2305.08846.

Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, Lisheng Sun Hosoya, Sergio Escalera, Gintare Karolina Dziugaite, Peter Triantafillou, and Isabelle Guyon. Are we making progress in unlearning? findings from the first NeurIPS unlearning competition. *arXiv preprint arXiv:2406.09073*, 2024.

Cheng-Long Wang, Qi Li, Zihang Xiang, Yinzhi Cao, and Di Wang. Towards lifecycle unlearning commitment management: Measuring sample-level unlearning completeness. In *USENIX Security Symposium*, 2025a.

Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. TAPE: Tailored posterior difference for auditing of machine unlearning. In *Proceedings of the ACM on Web Conference (WWW)*, 2025b.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. doi: 10.1198/jasa.2009.tm08651. URL https://doi.org/10.1198/jasa.2009.tm08651.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *International Conference on Machine Learning (ICML)*, pp. 58244–58282, 2024.

Binchi Zhang, Zihan Chen, Cong Shen, and Jundong Li. Verification of machine unlearning is fragile. In *International Conference on Machine Learning (ICML)*, 2024a.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep neural networks, 2025. URL https://arxiv.org/abs/2408.00920.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024b. URL https://arxiv.org/abs/2404.05868.

Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20350–20359, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_Decoupled_Distillation_to_Erase_A_General_Unlearning_Method_for_Any_CVPR_2025_paper.html.

APPENDIX ROADMAP

- **Background and theory.** Appendix A extends the related work; Appendix B proves Lemmas 4.1 and 4.2 and Corollary 4.3; Appendix C gives direct audits under pure DP and Gaussian DP and the interpretive conversion $\varepsilon^{(c)}$.
- **The auditor.** Appendix D gives pseudocode for both instantiations and the pairwise baseline; Appendix E tests robustness to the per-example score (logit versus cross-entropy) and to the aggregation rule (mean versus median); Appendix F studies how the audit depends on the number of candidate batches m, the batch size B, and the reported support size r, and compares the two instantiations.
- **Algorithms and setup.** Appendix G derives the certificates of the certified algorithms; Appendix H describes datasets, models and batching; Appendix I and Appendix J give configurations for the certified and uncertified methods, respectively.
- **Additional results.** Appendix K compares, on CIFAR-100, with a basic membership-inference attack that fits a logistic classifier to the per-example losses of candidates sampled into the forget set versus those not sampled; Appendix L reports hyperparameter sweeps; Appendix M relates TOFU utility metrics to the audit.
- **Convex setting.** Appendix N and Appendix O cover the Perturbed Newton and output-perturbation audits, respectively. Appendix P lists compute resources.

A EXTENDED RELATED WORK

This appendix expands the three lines of work summarised in Section 2: privacy auditing for differential privacy, audits designed specifically for unlearning, and membership inference.

A.1 AUDITING DIFFERENTIAL PRIVACY

Empirically lower-bounding the privacy parameter ε of an (ε, δ)-DP mechanism has been studied in Jagielski et al. (2020); Nasr et al. (2021; 2023); Annamalai & De Cristofaro (2024). The common template plants a canary in one of a pair of neighbouring datasets, runs the mechanism, and asks an adversary which dataset produced the observed release. The hypothesis-testing characterisation of DP Kairouz et al. (2015); Wasserman & Zhou (2010) converts the adversary's true- and false-positive rates into a lower bound on ε, with a confidence interval—typically Clopper–Pearson—accounting for the finite number of trials. Because that interval shrinks only as the number of independent training runs grows, the approach is expensive: driving the audit's Type I and Type II errors low enough to certify a large ε can require thousands of runs, which is prohibitive for deep models.

Two developments reduce this cost. Steinke et al. (2023) observe that many canaries can be planted in a *single* run and audited jointly, with each canary's inclusion drawn independently, so that the number of Bernoulli trials scales with the number of canaries rather than the number of runs. Mahloujifar et al. (2025) give the analogous construction for f-DP. Our auditor inherits the joint-canary idea, with two departures forced by the unlearning setting. First, an unlearning guarantee constrains a *subset*-level comparison—the unlearned model against a reference trained without the forget set—rather than the single-point neighbouring relation that standard DP audits exploit, so the per-canary signal is weaker and a single run does not suffice; we therefore aggregate a joint statistic across runs. Second, the comparison the auditor can actually perform is between two unlearned outputs, whereas the definition compares each to a common retraining reference, so the reduction passes through that reference and incurs the Rényi change-of-measure step of Lemma 4.1.

A further difference concerns the attack rather than the accounting. Steinke et al. (2023) and Mahloujifar et al. (2025) score canaries by a loss difference, which is convenient but discards the per-example calibration that makes modern membership inference effective. We instead fit, for each canary, Gaussians to its logit-scaled confidence under the in-forget-set and never-seen hypotheses and threshold the resulting likelihood ratio, in the manner of LiRA Carlini et al. (2022). Since the validity of the bound does not depend on the predictor (Section 5), a stronger attack translates directly into a larger reportable bound.

A.2 AUDITS DESIGNED FOR UNLEARNING

A separate literature audits unlearning directly rather than auditing a DP guarantee. *Backdoor-based verification* asks the data owner to poison the points it will later request be deleted, and tests whether the backdoor survives; Athena Sommer et al. (2022) is representative. Zhang et al. (2024a) show that such schemes are fragile against a dishonest provider, who can detect and neutralise the backdoor without genuinely unlearning, so a passed backdoor test is evidence only against a provider who is not adversarial in this specific way. *Output-difference* audits such as TAPE Wang et al. (2025b) train a reconstructor on the gap between the pre- and post-unlearning models, recovering information about the forget set from the update itself. *Sample-level* audits Wang et al. (2025a); Triantafillou et al. (2024); Gu et al. (2025) return a per-example score measuring how completely an individual point appears to have been removed.

These differ from our audit in what they certify. A per-example score answers "does this point still appear to be in the model?", which is the right question for a data subject but does not compose into a statement about the algorithm. Our auditor instead returns a lower confidence bound on the algorithm's zCDP certified-unlearning parameter, which is directly comparable against a claimed certificate and can therefore falsify it. The two are complementary: a per-example score localises residual memorisation, while an algorithm-level bound adjudicates a published guarantee. Audits under pure DP and Gaussian DP, reported separately, appear in Appendix C.

A.3 MEMBERSHIP INFERENCE

Membership inference began with shadow-model classifiers Shokri et al. (2016) and simple loss thresholds Yeom et al. (2018), and was sharpened substantially by per-example likelihood-ratio tests—LiRA Carlini et al. (2022) and RMIA Zarifzadeh et al. (2024)—which calibrate the decision threshold to each example rather than applying a global one. This calibration is what makes the attack informative at the low false-positive rates that matter for auditing, and it is the component we adopt.

In the unlearning setting, Chen et al. (2021) and Bertran et al. (2024) exploit the *difference* between the original and the unlearned model to infer membership in, or reconstruct, the deleted set. This is a strictly stronger access model than ours: it presumes the adversary observes the model both before and after the deletion request. Our auditor queries only the unlearned model, which is the access a deployed system actually exposes, and correspondingly reports what can be established from that alone.

B PROOFS FOR THE ZCDP/RÉNYI AUDIT

B.1 FROM ZCDP-CERTIFIED UNLEARNING TO LOCAL RÉNYI DP

Proof of Lemma 4.1. Let P_i be the law of $f_u(S_i)$ and let R be the law of f_r. For conjugate $p, q > 1$, the weak triangle inequality (Mironov, 2017, Proposition 11) gives

$$D_\gamma(P_1\|P_2) \le \frac{\gamma - 1/p}{\gamma - 1} D_{p\gamma}(P_1\|R) + D_{q(\gamma-1/p)}(R\|P_2).$$

Both orders on the right are greater than one: clearly $p\gamma > 1$, and, using $q = p/(p-1)$,

$$q(\gamma - 1/p) = \frac{p\gamma - 1}{p - 1} > 1.$$

The bidirectional zCDP assumption can therefore be applied at both orders, giving

$$\begin{aligned} D_\gamma(P_1\|P_2) &\le \rho \frac{\gamma - 1/p}{\gamma - 1} p\gamma + \rho q(\gamma - 1/p) \\ &= \rho(p\gamma - 1)\left(\frac{\gamma}{\gamma - 1} + \frac{1}{p - 1}\right). \end{aligned}$$

Set $x = p - 1 > 0$. Since $p\gamma - 1 = (\gamma - 1) + \gamma x$, the last expression equals

$$\rho\left(2\gamma + \frac{\gamma^2 x}{\gamma - 1} + \frac{\gamma - 1}{x}\right).$$

By the arithmetic–geometric mean inequality,

$$\frac{\gamma^2 x}{\gamma-1}+\frac{\gamma-1}{x}\geq 2\gamma,$$

with equality at $x=(\gamma-1)/\gamma$. Thus the minimizing conjugate exponents are

$$p^\star=\frac{2\gamma-1}{\gamma},\qquad q^\star=\frac{2\gamma-1}{\gamma-1}.$$

For this choice, the two divergence orders are

$$p^\star\gamma=2\gamma-1>1,\qquad q^\star(\gamma-1/p^\star)=2\gamma>1,$$

and hence

$$D_\gamma(P_1\|P_2)\leq 4\rho\gamma.$$

The audit output is a post-processing of $f_u(S_i)$, so data processing completes the proof. □

B.2 THE OVERLAP-SCORE AUDIT

We now restate and prove the overlap-score lemma from Section 4.2.

Lemma (Restatement of Lemma 4.2). Let $\mathcal{M}:\mathcal{S}_m\to\{-1,0,+1\}^m$ have a local Rényi-DP curve $\{\varepsilon_\gamma^{\mathrm{loc}}\}_{\gamma>1}$, and suppose $\mathcal{M}(S)\in\mathcal{T}$ almost surely. Draw $S^{(1)},\dots,S^{(L)}$ independently and uniformly from $\mathcal{S}_m$, draw the corresponding mechanism outputs independently, and define the overlap scores $V^{(1)},\dots,V^{(L)}$ as above. Then, for every $v\in\mathbb{R}$,

$$\Pr\left[\frac{1}{L}\sum_{\ell=1}^{L}V^{(\ell)}\geq v\right]\leq\inf_{\gamma>1}\exp\left(\frac{\gamma-1}{\gamma}L\varepsilon_\gamma^{\mathrm{loc}}\right)\left[\inf_{\theta\geq 0}\exp\left(L\log\sum_{u=0}^{r}e^{\theta u}\pi(u)-\theta Lv\right)\right]^{\frac{\gamma-1}{\gamma}},\tag{3}$$

$$\Pr\Big[\mathrm{Median}\{V^{(\ell)}\}_{\ell=1}^{L}\geq v\Big]\leq\inf_{\gamma>1}\exp\left(\frac{\gamma-1}{\gamma}L\varepsilon_\gamma^{\mathrm{loc}}\right)\left[\binom{L}{\lceil L/2\rceil}\overline{\Pi}(v)^{\lceil L/2\rceil}\right]^{\frac{\gamma-1}{\gamma}}.\tag{4}$$

Proof. For $s\in\mathcal{S}_m$, let P_s denote the distribution of $\mathcal{M}(s)$ and define the mixture

$$\overline{P}:=\frac{1}{M'}\sum_{t\in\mathcal{S}_m}P_t.$$

Consider two experiments on $\mathcal{S}_m\times\mathcal{T}$. In the matched experiment,

$$Q_1(S=s,\widehat{S}\in B)=\frac{1}{M'}P_s(B),$$

whereas in the independent null experiment,

$$Q_0(S=s,\widehat{S}\in B)=\frac{1}{M'}\overline{P}(B).$$

Thus S and $\widehat{S}$ are independent under Q_0, while their marginal distributions agree with those under Q_1. Let $P_1:=Q_1^{\otimes L}$ and $P_0:=Q_0^{\otimes L}$ denote the laws of the L independent pairs.

Divergence between the experiments. Fix $\gamma>1$. Local Rényi DP gives $D_\gamma(P_s\|P_t)\leq\varepsilon_\gamma^{\mathrm{loc}}$ for every $s,t\in\mathcal{S}_m$. For fixed P, the quantity

$$\exp\big((\gamma-1)D_\gamma(P\|Q)\big)=\int\left(\frac{dP}{dQ}\right)^\gamma dQ$$

is convex in Q. Consequently,

$$D_\gamma(P_s\|\overline{P})\leq\varepsilon_\gamma^{\mathrm{loc}}\qquad\text{for every } s\in\mathcal{S}_m.$$

Using the definitions of Q_1 and Q_0,

$$\begin{aligned}\exp\big((\gamma-1)D_\gamma(Q_1\|Q_0)\big)&=\frac{1}{M'}\sum_{s\in\mathcal{S}_m}\exp\big((\gamma-1)D_\gamma(P_s\|\overline{P})\big)\\&\leq\exp\big((\gamma-1)\varepsilon_\gamma^{\mathrm{loc}}\big).\end{aligned}$$

Hence $D_\gamma(Q_1\|Q_0)\leq\varepsilon_\gamma^{\mathrm{loc}}$. Additivity of Rényi divergence under product distributions yields

$$D_\gamma(P_1\|P_0)\leq L\varepsilon_\gamma^{\mathrm{loc}}.$$

Change of measure. For any event A depending on the L pairs, Hölder's inequality (or the probability-preservation inequality for Rényi divergence (Mironov, 2017, Proposition 10)) gives

$$P_1(A) \le \exp\left(\frac{\gamma-1}{\gamma} L\varepsilon_\gamma^{\mathrm{loc}}\right) P_0(A)^{\frac{\gamma-1}{\gamma}}. \tag{5}$$

Null overlap distribution. Under P_0, each $S^{(\ell)}$ is uniform on $\mathcal{S}_m$ and independent of $\widehat{S}^{(\ell)}$. For a fixed $t \in \mathcal{T}$, let α_1 count agreements on the $+1$ coordinates of t and let α_2 count agreements on its -1 coordinates. The overlap is $u = \alpha_1 + \alpha_2$. Once these agreement positions are chosen, balance of S forces the number of positive signs on the zero block of t to equal $\lceil (m-r)/2 \rceil - (\alpha_1 - \alpha_2)$. Therefore the number of $s \in \mathcal{S}_m$ with overlap u is

$$N_u = \sum_{\substack{\alpha_1,\alpha_2 \in \{0,\dots,r/2\} \\ \alpha_1+\alpha_2=u}} \binom{m-r}{\lceil (m-r)/2 \rceil - (\alpha_1-\alpha_2)} \binom{r/2}{\alpha_1} \binom{r/2}{\alpha_2}.$$

This count does not depend on t, so

$$P_0[V^{(\ell)} = u] = \frac{N_u}{M'} = \pi(u), \qquad P_0[V^{(\ell)} \ge v] = \overline{\Pi}(v).$$

Mean statistic. For $\theta \ge 0$, independence under the null and Markov's inequality give

$$P_0\left[\frac{1}{L}\sum_{\ell=1}^{L} V^{(\ell)} \ge v\right] \le \exp\left(L \log \sum_{u=0}^{r} e^{\theta u}\pi(u) - \theta L v\right).$$

Optimizing over θ, substituting this bound into Equation (5), and then optimizing over $\gamma > 1$ proves Equation (3).

Median statistic. If the median is at least v, at least $\lceil L/2 \rceil$ of the independent scores are at least v. A union bound over those subsets gives

$$P_0\Big[\mathrm{Median}\{V^{(\ell)}\}_{\ell=1}^{L} \ge v\Big] \le \binom{L}{\lceil L/2 \rceil} \overline{\Pi}(v)^{\lceil L/2 \rceil}.$$

Applying Equation (5) and optimizing over $\gamma > 1$ proves Equation (4). □

B.3 Proof of Corollary 4.3

We now prove the following corollary Corollary 4.3 showing monotonicity of the computed probability bounds in Lemma 4.2.

Corollary (Monotonicity in the zCDP parameter). For a fixed observed threshold v, both Rényi overlap bounds after the substitution $\varepsilon_\gamma^{\mathrm{loc}} = \varepsilon_\gamma^{\mathrm{loc}}(\rho)$ are non-decreasing in ρ.

Proof. For every fixed $\gamma > 1$, the only ρ-dependent factor is $\exp\big((\gamma-1)L\varepsilon_\gamma^{\mathrm{loc}}(\rho)/\gamma\big)$, which is non-decreasing because $\varepsilon_\gamma^{\mathrm{loc}}(\rho)$ is linear and non-decreasing in ρ. The infimum over γ of functions that are all non-decreasing in the same scalar is also non-decreasing. □

C Audits under $(\varepsilon, 0)$ DP and Gaussian DP

We first instantiate the auditor for pure $(\varepsilon, 0)$-DP and then give a direct audit of the Gaussian-DP parameter. Unlike pure DP, approximate DP does not compose transitively through the common reference without significantly enlarging the δ-slack, which weakens the induced pairwise audit; see Remark C.4.

C.1 AUDIT UNDER $(\varepsilon, 0)$-DP

For distributions P, Q, write $P \approx_{\varepsilon,\delta} Q$ when, for every measurable event A,

$$P(A) \le e^{\varepsilon} Q(A) + \delta \qquad \text{and} \qquad Q(A) \le e^{\varepsilon} P(A) + \delta.$$

Definition C.1 ((ε, δ)-DP-certified unlearning (Koloskova et al., 2025))**.** We say that $\mathcal{U}$ is $(\varepsilon, 0)$-DP-certified for $\mathcal{A}$ under our threat model if, for every public initialization x_o and shuffling order π, there exists a reference algorithm $\overline{\mathcal{A}}_{x_o,\pi}$ such that the following holds for every forget set $\mathcal{D}_f \subseteq \mathcal{D}$. Writing $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{D}_f$, let P_u and P_r be the laws of

$$\mathcal{U}\big(\mathcal{A}_{x_o,\pi}(\mathcal{D}), \mathcal{D}, \mathcal{D}_f\big) \qquad \text{and} \qquad \overline{\mathcal{A}}_{x_o,\pi}(\mathcal{D}_r),$$

respectively. We require $P_u \approx_{\varepsilon,\delta} P_r$.

The audit compares two unlearned-model distributions, whereas certified unlearning compares each of them with the common retrained reference. The following reduction accounts for this difference.

Composing through the common reference amplifies the δ slack, as Remark C.4 below explains; we therefore restrict this audit to $\delta = 0$, and the lemmas that follow are stated only for that case.

Lemma C.2 (Pure-DP reduction through the reference)**.** *If $\mathcal{U}$ is $(\varepsilon, 0)$-DP-certified according to Definition C.1, then the induced audit mechanism $\mathcal{M}$ is 2ε-locally differentially private: for every $s_1, s_2 \in \mathcal{S}_m$ and measurable event A,*

$$\Pr[\mathcal{M}(s_1) \in A] \le e^{2\varepsilon} \Pr[\mathcal{M}(s_2) \in A].$$

Proof. Let P_{s_i} be the law of the unlearned model for encoding s_i and let R be the common reference law. Then

$$P_{s_1}(A) \le e^{\varepsilon} R(A) \le e^{2\varepsilon} P_{s_2}(A).$$

The reverse inequality follows by exchanging s_1 and s_2, and post-processing transfers both inequalities to $\mathcal{M}$. □

We now state the resulting overlap audit. Recall from Section 4.2 that $M' = |\mathcal{S}_m|$, that $\pi(u)$ is the chance overlap mass, and that $N_u := M'\pi(u)$ is the number of balanced sign vectors having overlap u with any fixed prediction in $\mathcal{T}$. For a local privacy parameter $\eta \ge 0$, define the pointwise upper bound and its tail by

$$\pi_{\eta}^{\mathrm{DP}}(u) := N_u \frac{e^{\eta}}{e^{\eta} + M' - 1}, \qquad \overline{\Pi}_{\eta}^{\mathrm{DP}}(v) := \sum_{u=\lceil v \rceil}^{r} \pi_{\eta}^{\mathrm{DP}}(u). \tag{6}$$

Draw $S^{(1)}, \dots, S^{(L)}$ independently and uniformly from $\mathcal{S}_m$, draw the corresponding outputs $\widehat{S}^{(\ell)} = \mathcal{M}(S^{(\ell)})$ independently, and define

$$V^{(\ell)} := \sum_{j=1}^{m} \max\{0, \widehat{S}_j^{(\ell)} S_j^{(\ell)}\}, \qquad Z_{\mathrm{mean}} := \frac{1}{L} \sum_{\ell=1}^{L} V^{(\ell)}, \qquad Z_{\mathrm{med}} := \mathrm{Median}\{V^{(\ell)}\}_{\ell=1}^{L}.$$

For a certified-unlearning parameter ε, set

$$B_{\varepsilon,\mathrm{mean}}^{\mathrm{pure}}(v) := \min\left\{1, \inf_{\lambda \ge 0} \exp\left(L \log \sum_{u=0}^{r} e^{\lambda u} \pi_{2\varepsilon}^{\mathrm{DP}}(u) - \lambda L v\right)\right\}, \tag{7}$$

$$B_{\varepsilon,\mathrm{med}}^{\mathrm{pure}}(v) := \min\left\{1, \binom{L}{\lceil L/2 \rceil} \left(\overline{\Pi}_{2\varepsilon}^{\mathrm{DP}}(v)\right)^{\lceil L/2 \rceil}\right\}. \tag{8}$$

The argument 2ε incorporates the reduction in Lemma C.2.

Lemma C.3 (Pure-DP overlap audit)**.** *If the unlearning algorithm is $(\varepsilon, 0)$-DP-certified, then for every $v \in \mathbb{R}$,*

$$\Pr[Z_{\mathrm{mean}} \ge v] \le B_{\varepsilon,\mathrm{mean}}^{\mathrm{pure}}(v), \qquad \Pr[Z_{\mathrm{med}} \ge v] \le B_{\varepsilon,\mathrm{med}}^{\mathrm{pure}}(v).$$

Proof. Write $\eta = 2\varepsilon$. By Lemma C.2, $\mathcal{M}$ is η-locally differentially private. Fix a prediction $t \in \mathcal{T}$ in the support of $\widehat{S}$ and let $p_s(t) := \Pr\Big[\widehat{S} = t \mid S = s\Big]$. For every $s, s' \in \mathcal{S}_m$, local DP applied to the singleton event $\{t\}$ gives

$$p_{s'}(t) \geq e^{-\eta} p_s(t).$$

Because S is uniform on $\mathcal{S}_m$, Bayes' rule therefore yields

$$\Pr\Big[S = s \mid \widehat{S} = t\Big] = \frac{p_s(t)}{\sum_{s' \in \mathcal{S}_m} p_{s'}(t)} \leq \frac{p_s(t)}{p_s(t) + (M'-1)e^{-\eta} p_s(t)} = \frac{e^{\eta}}{e^{\eta} + M' - 1}.$$

Exactly N_u choices of s have overlap u with a fixed t. Consequently,

$$\Pr\Big[V^{(\ell)} = u \mid \widehat{S}^{(\ell)} = t\Big] = \sum_{s \in \mathcal{S}_m :\, V(s,t) = u} \Pr\Big[S^{(\ell)} = s \mid \widehat{S}^{(\ell)} = t\Big] \leq N_u \frac{e^{\eta}}{e^{\eta} + M' - 1} = \pi_{\eta}^{\mathrm{DP}}(u),$$

where $V(s,t) := \sum_{j=1}^{m} \max\{0, t_j s_j\}$. Averaging over t gives $\Pr\big[V^{(\ell)} = u\big] \leq \pi_{\eta}^{\mathrm{DP}}(u)$ and hence, for every $\lambda \geq 0$,

$$\mathbb{E}\Big[e^{\lambda V^{(\ell)}}\Big] \leq \sum_{u=0}^{r} e^{\lambda u} \pi_{\eta}^{\mathrm{DP}}(u).$$

Independence across runs and Markov's inequality now give

$$\Pr[Z_{\mathrm{mean}} \geq v] \leq \exp\left(L \log \sum_{u=0}^{r} e^{\lambda u} \pi_{\eta}^{\mathrm{DP}}(u) - \lambda L v \right).$$

Taking the infimum over $\lambda \geq 0$ and truncating at one proves Equation (7).

For the median, put $k := \lceil L/2 \rceil$. The pointwise bound above implies

$$p_v := \Pr\Big[V^{(\ell)} \geq v\Big] \leq \sum_{u=\lceil v \rceil}^{r} \pi_{\eta}^{\mathrm{DP}}(u) = \overline{\Pi}_{\eta}^{\mathrm{DP}}(v).$$

If $Z_{\mathrm{med}} \geq v$, some subset of k runs must all satisfy $V^{(\ell)} \geq v$. A union bound over the $\binom{L}{k}$ such subsets and independence therefore give

$$\Pr[Z_{\mathrm{med}} \geq v] \leq \binom{L}{k} p_v^k \leq \binom{L}{k} \left(\overline{\Pi}_{\eta}^{\mathrm{DP}}(v)\right)^k.$$

Truncating at one and substituting $\eta = 2\varepsilon$ proves Equation (8). □

For $a \in \{\mathrm{mean}, \mathrm{med}\}$ and observed value $v_{\mathrm{obs}} = Z_a$, define

$$\varepsilon_{\mathrm{LB},a} := \sup\left(\left\{\varepsilon \geq 0 : B_{\varepsilon,a}^{\mathrm{pure}}(v_{\mathrm{obs}}) \leq \zeta\right\} \cup \{0\}\right). \tag{9}$$

The bounds are non-decreasing in ε, since $\pi_{2\varepsilon}^{\mathrm{DP}}(u)$ is non-decreasing in ε. Therefore, every null $H_0 : \varepsilon^{\star} \leq \varepsilon_0$ with $\varepsilon_0 < \varepsilon_{\mathrm{LB},a}$ is rejected at level at most ζ. Computing the median bound and inverting it requires only a scalar search over ε; the mean bound additionally requires the one-dimensional optimization over λ.

Remark C.4 (Why we restrict this audit to $\delta = 0$). If each unlearned distribution is only (ε, δ)-indistinguishable from the reference, then

$$P_{s_1}(A) \leq e^{2\varepsilon} P_{s_2}(A) + (1 + e^{\varepsilon})\delta.$$

Thus the induced pairwise guarantee is $(2\varepsilon, (1 + e^{\varepsilon})\delta)$, rather than $(2\varepsilon, 2\delta)$. A direct local approximate-DP overlap bound can be derived, but its additive slack is amplified by the combinatorial audit output space and is typically vacuous in our regimes. We therefore state the clean pure-DP audit and do not include a separate lemma for $\delta > 0$.

C.2 AUDIT UNDER GAUSSIAN DP

This section derives an audit of the Gaussian-DP parameter directly from its hypothesis-testing semantics.

Let P and Q be distributions on a common measurable space $(\mathcal{X}, \mathcal{F})$. A randomized test is a measurable function $\phi : \mathcal{X} \to [0,1]$, where $\phi(x)$ is the probability of rejecting the null P after observing x. We write $\mathbb{E}_P[\phi] := \int_{\mathcal{X}} \phi \, dP$; hence $\mathbb{E}_P[\phi]$ and $1 - \mathbb{E}_Q[\phi]$ are the test's type-I and type-II error probabilities, respectively. Following Dong et al. (2022), define the hypothesis-testing trade-off function

$$T(P,Q)(a) := \inf_{\substack{\phi:\mathcal{X}\to[0,1] \text{ measurable} \\ \mathbb{E}_P[\phi]\le a}} \left(1 - \mathbb{E}_Q[\phi]\right), \qquad a \in [0,1].$$

Let

$$G_\mu(a) := \Phi\big(\Phi^{-1}(1-a) - \mu\big)$$

be the trade-off between $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$. A pair is μ-GDP when its trade-off dominates G_μ.

Definition C.5 (μ-GDP-certified unlearning)**.** We say that $\mathcal{U}$ is μ-GDP-certified for $\mathcal{A}$ under our threat model if, for every public initialization x_o and shuffling order π, there exists a reference algorithm $\overline{\mathcal{A}}_{x_o,\pi}$ such that the following holds for every forget set $\mathcal{D}_f \subseteq \mathcal{D}$. Writing $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{D}_f$, let P_u and P_r denote the laws of

$$\mathcal{U}\big(\mathcal{A}_{x_o,\pi}(\mathcal{D}), \mathcal{D}, \mathcal{D}_f\big) \qquad \text{and} \qquad \overline{\mathcal{A}}_{x_o,\pi}(\mathcal{D}_r),$$

respectively. We require

$$T(P_u, P_r) \ge G_\mu \qquad \text{and} \qquad T(P_r, P_u) \ge G_\mu.$$

In the audit, all encoded forget sets leave the same retain set $\mathcal{D}_r$. Accordingly, we write $P_s := \mathrm{Law}(f_u(s))$ for $s \in \mathcal{S}_m$ and $R := \mathrm{Law}(f_r) = \mathrm{Law}(\overline{\mathcal{A}}_{x_o,\pi}(\mathcal{D}_r))$ for their common reference law.

For $\varepsilon \ge 0$, the hockey-stick divergence from P to Q is (Dong et al., 2022, Section 2.4 and Proposition 6)

$$E_\varepsilon(P\|Q) := \sup_{A\in\mathcal{F}} \big\{P(A) - e^\varepsilon Q(A)\big\} = \int_{\mathcal{X}} (p - e^\varepsilon q)_+ \, d\nu, \tag{10}$$

where ν is any measure dominating P and Q, $p = dP/d\nu$, $q = dQ/d\nu$, and $(x)_+ := \max\{x, 0\}$. Thus P and Q are (ε,δ)-indistinguishable exactly when $E_\varepsilon(P\|Q) \le \delta$ and $E_\varepsilon(Q\|P) \le \delta$.

The exact approximate-DP profile associated with G_μ is (Dong et al., 2022, Proposition 6 and Corollary 1)

$$\theta_\varepsilon(\mu) := \Phi\left(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}\right) - e^\varepsilon \Phi\left(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\right), \tag{11}$$

with the value at $\mu = 0$ defined by continuity. In particular, every μ-GDP comparison is $(\varepsilon, \theta_\varepsilon(\mu))$-indistinguishable. This is a property of the Gaussian trade-off function and does not assume that P and Q are themselves Gaussian.

C.2.1 REDUCTION THROUGH THE COMMON REFERENCE

The group-privacy theorem for f-DP composes trade-off guarantees along a path (Dong et al., 2022, Section 2.5, Theorem 3). Specializing its two-edge formula to Gaussian trade-offs gives

$$1 - (1 - G_\mu)^{\circ 2} = G_{2\mu}.$$

Apply this result to the path P_{s_1}, R, P_{s_2}. The two comparisons in Definition C.5 supply the required guarantee in either direction along each edge. Hence, for every $s_1, s_2 \in \mathcal{S}_m$,

$$T(P_{s_1}, P_{s_2}) \ge G_{2\mu} \qquad \text{and} \qquad T(P_{s_2}, P_{s_1}) \ge G_{2\mu}. \tag{12}$$

Thus passing through the common reference costs a factor of two in the GDP parameter; no Gaussian-envelope approximation is needed.

C.2.2 MATCHED AND INDEPENDENT EXPERIMENTS

For $s \in \mathcal{S}_m$, let K_s be the law of the audit prediction $\widehat{S}$ given $S = s$, and let $\overline{K} := |\mathcal{S}_m|^{-1} \sum_{t \in \mathcal{S}_m} K_t$. Define one-run matched and independent experiments by

$$Q_1(S = s, \widehat{S} \in B) = \frac{1}{|\mathcal{S}_m|} K_s(B), \qquad Q_0(S = s, \widehat{S} \in B) = \frac{1}{|\mathcal{S}_m|} \overline{K}(B).$$

The local GDP guarantee is preserved by the post-processing that produces $\widehat{S}$. Convexity of hockey-stick divergence in each distribution gives

$$E_\varepsilon(K_s \| \overline{K}) \le \frac{1}{|\mathcal{S}_m|} \sum_{t \in \mathcal{S}_m} E_\varepsilon(K_s \| K_t) \le \theta_\varepsilon(2\mu),$$

and the analogous reverse inequality. Averaging over s proves that Q_1, Q_0 are 2μ-GDP. The GDP tensor-product identity states that the product of experiments with parameters $\mu_1, \dots, \mu_L$ is $\sqrt{\sum_{\ell=1}^{L} \mu_\ell^2}$-GDP (Dong et al., 2022, Section 3.1, Corollary 2). Taking every $\mu_\ell = 2\mu$ therefore gives

$$Q_1^{\otimes L}, Q_0^{\otimes L} \quad \text{are} \quad 2\sqrt{L}\,\mu\text{-GDP}. \tag{13}$$

C.2.3 DIRECT OVERLAP-SCORE TEST

Use the overlap scores $V^{(\ell)}$ and aggregations $Z_{\mathrm{mean}}, Z_{\mathrm{med}}$ defined in Appendix C.1, now viewing the L matched pairs as drawn independently from Q_1. Using the chance-overlap mass π and tail $\overline{\Pi}$ from Section 4.2, define the corresponding null tail bounds

$$q_{\mathrm{mean}}(v) := \min\left\{1, \inf_{\lambda \ge 0} \exp\left(L \log \sum_{u=0}^{r} e^{\lambda u} \pi(u) - \lambda L v\right)\right\}, \tag{14}$$

$$q_{\mathrm{med}}(v) := \min\left\{1, \binom{L}{\lceil L/2 \rceil} \overline{\Pi}(v)^{\lceil L/2 \rceil}\right\}. \tag{15}$$

These are the same combinatorial null calculations used in the main Rényi audit and satisfy

$$\Pr_{Q_0^{\otimes L}}[Z_{\mathrm{mean}} \ge v] \le q_{\mathrm{mean}}(v), \qquad \Pr_{Q_0^{\otimes L}}[Z_{\mathrm{med}} \ge v] \le q_{\mathrm{med}}(v).$$

If P, Q are ν-GDP and $Q(A) \le q$, the GDP trade-off gives

$$P(A) \le 1 - G_\nu(q) = \Phi\big(\Phi^{-1}(q) + \nu\big).$$

Combining this identity with Equation (13) yields the direct GDP analogue of the main overlap theorem.

Lemma C.6 (Direct GDP overlap audit)**.** *Suppose the unlearning algorithm is μ-GDP-certified according to Definition C.5. Then, for every threshold $v \in \mathbb{R}$,*

$$\Pr_{Q_1^{\otimes L}}[Z_{\mathrm{mean}} \ge v] \le B^{\mathrm{GDP}}_{\mu,\mathrm{mean}}(v) := \Phi\Big(\Phi^{-1}(q_{\mathrm{mean}}(v)) + 2\sqrt{L}\,\mu\Big),$$

$$\Pr_{Q_1^{\otimes L}}[Z_{\mathrm{med}} \ge v] \le B^{\mathrm{GDP}}_{\mu,\mathrm{med}}(v) := \Phi\Big(\Phi^{-1}(q_{\mathrm{med}}(v)) + 2\sqrt{L}\,\mu\Big). \tag{16}$$

The inverse Gaussian CDF is interpreted in the extended sense at zero and one.

For $a \in \{\mathrm{mean}, \mathrm{med}\}$, let Z_a, q_a, and $B^{\mathrm{GDP}}_{\mu,a}$ denote the corresponding mean or median quantities defined above. Choose one such aggregation rule and let v_{obs} be the realized value of Z_a. Because $B^{\mathrm{GDP}}_{\mu,a}(v_{\mathrm{obs}})$ is non-decreasing in the claimed GDP parameter, define

$$\mu_{\mathrm{LB},a} := \sup\left(\left\{\mu \ge 0 : B^{\mathrm{GDP}}_{\mu,a}(v_{\mathrm{obs}}) \le \zeta\right\} \cup \{0\}\right). \tag{17}$$

This lower bound can be computed by first evaluating $q_{\mathrm{obs}} := q_a(v_{\mathrm{obs}})$. If $q_{\mathrm{obs}} \ge \zeta$, the audit returns $\mu_{\mathrm{LB},a} = 0$. Otherwise, monotonicity of Φ shows that Equation (17) has the closed form

$$\mu_{\mathrm{LB},a} = \max\left\{0, \frac{\Phi^{-1}(\zeta) - \Phi^{-1}(q_{\mathrm{obs}})}{2\sqrt{L}}\right\}. \tag{18}$$

For the median, q_{obs} is available directly from Equation (15); for the mean, it requires the one-dimensional Chernoff optimization in Equation (14). No further numerical inversion is required. Observing the corresponding tail event rejects every null $H_0 : \mu^\star \le \mu_0$ with $\mu_0 < \mu_{\text{LB},a}$ at level at most ζ. Thus $\mu_{\text{LB},a}$ is a lower confidence bound on the optimal GDP parameter itself; no Rényi-DP or zCDP conversion is used.

C.3 Interpretive conversions to an (ε, δ) scale

Our audits report ρ and μ, whereas much of the unlearning literature is stated in (ε, δ) terms. For comparability we fix the conversion slack to $\delta = 10^{-3}$ throughout and record the two conversions here. Neither is a confidence bound: each is an order-preserving rescaling of an audited lower bound, reported only so that magnitudes can be read on a familiar scale.

From zCDP. Applying the refined Rényi-DP conversion of Balle et al. (2020, Theorem 20) to the zCDP curve $\varepsilon_\gamma^{\text{RDP}} = \rho\gamma$ gives

$$\varepsilon^{(c)}(\rho) := \max\left\{0, \inf_{\gamma>1}\left[\rho\gamma + \log\left(1 - \frac{1}{\gamma}\right) - \frac{\log(\delta) + \log\gamma}{\gamma - 1}\right]\right\}, \tag{19}$$

where "c" denotes conversion, the infimum selects the tightest order, and the outer maximum enforces $\varepsilon \ge 0$. For a *valid* zCDP parameter ρ this yields $(\varepsilon^{(c)}(\rho), 10^{-3})$-indistinguishability; evaluated at ρ_{LB}, whose input is only a lower bound on ρ, it is neither a lower nor an upper confidence bound on the smallest valid $(\varepsilon, 10^{-3})$ parameter. The simpler but looser standard conversion would give $\rho + 2\sqrt{\rho\log(10^3)}$.

From Gaussian DP. Using the same reporting choice $\delta = 10^{-3}$, define

$$\varepsilon_{\text{GDP}}^{(c)}(\mu) := \inf\left\{\varepsilon \ge 0 : \theta_\varepsilon(\mu) \le 10^{-3}\right\}. \tag{20}$$

For $\mu > 0$, $\theta_\varepsilon(\mu)$ decreases continuously from $\theta_0(\mu) = 2\Phi(\mu/2) - 1$ to zero. Hence Equation (20) is zero when $10^{-3} \ge \theta_0(\mu)$ and otherwise is the unique positive solution of $\theta_\varepsilon(\mu) = 10^{-3}$, which we compute by bisection. This is the exact GDP-to-approximate-DP conversion for a valid μ-GDP parameter; evaluated at $\mu_{\text{LB},a}$ it carries the same caveat as Equation (19) above.

C.4 Audits under the alternative privacy notions

We apply the pure-DP and direct GDP audits above to the uncertified methods reported in the main paper. The tables also repeat the zCDP audit so that the three independently audited lower bounds can be compared in one place. In particular, ε_{LB} is the certified-unlearning lower bound at $\delta = 0$. The quantities ρ_{LB} and μ_{LB} are direct lower confidence bounds under zCDP and GDP, respectively, and are not halved. For a common numerical scale, $\varepsilon^{(c)}(\rho_{\text{LB}})$ and $\varepsilon_{\text{GDP}}^{(c)}(\mu_{\text{LB}})$ are the corresponding forward conversions at $\delta = 10^{-3}$. As explained in Equations (19) and (20), these converted values are interpretive and are not themselves confidence lower bounds on the optimal approximate-DP parameter.

The three notions disagree on the numerical value of the bound but not on what it says about the methods: in every table below, ε_{LB}, ρ_{LB} and μ_{LB} induce the same ranking. This is by construction rather than coincidence—at a fixed audit configuration all three are inverted from the same observed guesses, so each is a monotone function of the same statistic. The choice of privacy notion therefore fixes the scale on which distinguishability is reported, not which methods are judged more distinguishable.

C.4.1 SHAKESPEARE AUDITS

Method	Parameters	Pure DP	zCDP		GDP	
		$\varepsilon_{\mathrm{LB}}$	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	μ_{LB}	$\varepsilon^{(c)}_{\mathrm{GDP}}(\mu_{\mathrm{LB}})$
GD	$q = 1$, $\lambda = 1.5$, $E_f = 4$	8.92	3.444	12.02	2.72	11.44
Retain fine-tune	—	13.34	5.398	16.29	3.38	15.48
Gradient ascent	$E_f = 2$	0.0	0.0	0.0	0.00	0.00
Hessian-based	—	20.77	8.763	22.84	4.29	21.71

Table 3: Audited lower bounds on Shakespeare for the LLR overlap attack with $r = 100$, $m = 400$ forget files, $L = 10$ test runs, and confidence $1 - \zeta = 0.95$, at the configurations reported in the main paper.

C.4.2 DISTILLATION-BASED METHODS

Method	Pure DP	zCDP		GDP	
	$\varepsilon_{\mathrm{LB}}$	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	μ_{LB}	$\varepsilon^{(c)}_{\mathrm{GDP}}(\mu_{\mathrm{LB}})$
DELETE	189.23	89.44	136.73	13.49	131.80
BadTeacher	1.66	0.46	3.36	1.03	3.25
SCRUB	23.59	10.06	25.20	4.59	23.96
SCRUB+R	26.47	11.39	27.55	4.88	26.20

Table 4: Audits of DELETE after 20 unlearning epochs, of BadTeacher after one unlearning epoch, and of SCRUB and SCRUB+R at unlearning rate 1×10^{-3} (all mean-aggregation, $L = 10$ runs, $m = 4500$ candidate forget points, $r = 3000$ guesses; DELETE and BadTeacher are shown at unlearning rate 3×10^{-3}) on CIFAR-100. All use the adversarial split, as these are class-based methods. The pure-DP, zCDP, and GDP lower bounds are audited directly from the same guesses; $\varepsilon_{\mathrm{LB}}$ here is half of the induced local-DP bound $\varepsilon^{\mathrm{LDP}}_{\mathrm{LB}}$.

The independently computed pure-DP and GDP quantities for Hessian-based unlearning are reported at $r = 3000$, $\sigma = 10^{-3}$ in Table 5; they cannot be obtained by converting a zCDP lower bound.

C.4.3 GRADIENT-BASED METHODS

Setting	Params	$\varepsilon_{\mathrm{LB}}$		ρ_{LB} [$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$]		μ_{LB} [$\varepsilon^{(c)}_{\mathrm{GDP}}(\mu_{\mathrm{LB}})$]	
		U	A	U	A	U	A
GD	$q = 1$, $E_f = 10$	83.50	0.00	38.36 [68.9]	0.0 [0.0]	8.87 [65.9]	0.0 [0.0]
	$q = 2$, $E_f = 10$	92.04	5.59	42.45 [74.6]	2.02 [8.5]	9.33 [71.5]	2.1 [8.1]
	$q = 4$, $E_f = 10$	99.77	26.99	46.17 [79.8]	11.63 [28.0]	9.72 [76.5]	4.93 [26.6]
Method comparison	Retain fine-tune	81.76	84.02	37.5 [67.7]	38.6 [69.2]	8.8 [64.8]	8.9 [66.3]
	Gradient ascent ($E_f = 2$)	77.90	83.92	35.68 [65.1]	38.56 [69.1]	8.56 [62.2]	8.9 [66.2]
	Hessian-based ($\sigma = 10^{-3}$)	142.84	7.50	66.9 [107.7]	2.8 [10.6]	11.7 [103.5]	2.5 [10.1]

Table 5: Pure-DP, zCDP, and GDP audits of gradient-based unlearning on CIFAR-100 — gradient difference (GD) swept over q, retain fine-tuning, gradient ascent, and Hessian-based unlearning at final noise $\sigma = 10^{-3}$. $L = 10$ runs, $m = 4500$ candidate forget batches, $r = 3000$ guesses; U and A denote uniform and adversarial splits. Bracketed values are interpretive forward conversions, not confidence lower bounds.

D AUDITOR ALGORITHMS: PSEUDOCODE

This appendix gives pseudocode for the meta-algorithm of Section 4 and the two auditor instantiations (*batchwise inclusion/exclusion* and *joint sign vector prediction*) of Section 5; see Algorithms 1 to 3.

D.1 META-ALGORITHM

Algorithm 1 Meta-algorithm for auditor

Require: Batched dataset $\mathcal{D} \in \mathcal{B}^n$ partitioned as $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ with $|\mathcal{D}_f| = m$
Require: Algorithms $(\mathcal{A}, \mathcal{U})$, reported support size r, and number of runs L
1: **for** $\ell = 1, 2, \ldots, L$ **do**
2: Sample $S^{(\ell)} \sim \mathrm{Unif}(\mathcal{S}_m)$ and unlearnt model $f_u^{(\ell)} \leftarrow \mathcal{U}\big(\mathcal{A}(\mathcal{D}(S^{(\ell)})),\, \mathcal{D}(S^{(\ell)}),\, \mathcal{D}_f(S^{(\ell)})\big)$.
3: *Induced mechanism* $\mathcal{M}(S^{(\ell)})$: From $f_u^{(\ell)}$, output $\widehat{S}^{(\ell)} \in \{-1, 0, +1\}^m$ with exactly $r/2$ entries $+1$ and $r/2$ entries -1.
4: Compute overlap score $V^{(\ell)} := \sum_{j=1}^m \max\{0,\, \widehat{S}_j^{(\ell)} S_j^{(\ell)}\}$.
5: **end for**
6: **Return** summary statistics of $\{V^{(\ell)}\}_{\ell=1}^L$ (e.g., mean or median).

D.2 BATCHWISE INCLUSION/EXCLUSION PREDICTION (INSTANTIATION I)

We now present the audit algorithm Algorithm 2 below for batchwise inclusion/exclusion based predictor which is the first instantiation of Algorithm 1

Algorithm 2 Auditor Instantiation I: Batchwise inclusion/exclusion prediction

Require: Batched dataset $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ with $|\mathcal{D}_f| = m$; algorithms $(\mathcal{A}, \mathcal{U})$; calibration rounds Γ; support size r
Ensure: Mechanism $\mathcal{M} : \mathcal{S}_m \to \{-1, 0, +1\}^m$
1: // CALIBRATION PHASE (RUN ONCE, OFFLINE)
2: **for** $\gamma = 1, \ldots, \Gamma$ **do**
3: $S^{(\gamma)} \sim \mathrm{Unif}(\mathcal{S}_m), \quad \mathcal{D}^{(\gamma)} \leftarrow \mathcal{D}_r \cup \mathcal{D}_f(S^{(\gamma)})$
4: $f^{(\gamma)} \leftarrow \mathcal{U}\big(\mathcal{A}(\mathcal{D}^{(\gamma)}),\, \mathcal{D}^{(\gamma)},\, \mathcal{D}_f(S^{(\gamma)})\big)$
5: $s_{x,y}^{(\gamma)} \leftarrow \phi\big(f^{(\gamma)}(x)_y\big)$ for all $(x, y) \in \mathcal{D}_f$
6: **end for**
7: For each $j \in [m]$ and $(x, y) \in \mathcal{D}_{f,j}$, fit Gaussians:

$$(\mu_{x,y}^{\pm},\, \sigma_{x,y}^{\pm}) \;\leftarrow\; \mathrm{fit}\Big(\big\{s_{x,y}^{(\gamma)} : S_j^{(\gamma)} = \pm 1\big\}_{\gamma=1}^{\Gamma}\Big)$$

8: // MECHANISM $\mathcal{M}(S)$ (APPLIED AT EVALUATION)
9: $f \leftarrow \mathcal{U}\big(\mathcal{A}(\mathcal{D}(S)),\, \mathcal{D}(S),\, \mathcal{D}_f(S)\big), \quad s_{x,y} \leftarrow \phi(f(x)_y)$
10: For each batch $j \in [m]$, compute

$$\Lambda_j \;\leftarrow\; \sum_{(x,y)\in\mathcal{D}_{f,j}} \log \frac{\mathcal{N}\big(s_{x,y};\, \mu_{x,y}^{+},\, \sigma_{x,y}^{+}\big)}{\mathcal{N}\big(s_{x,y};\, \mu_{x,y}^{-},\, \sigma_{x,y}^{-}\big)}$$

11: $\widehat{S}_j \leftarrow +1$ for top-$r/2$ values of Λ_j; -1 for bottom-$r/2$; 0 otherwise
12: **return** $\widehat{S} \in \{-1, 0, +1\}^m$

D.3 JOINT SIGN-VECTOR PREDICTION (INSTANTIATION II)

We now present the audit algorithm Algorithm 3 below for joint sign -vector prediction which is the second instantiation of Algorithm 1.

Algorithm 3 Auditor Instantiation II: Joint sign-vector prediction

Require: Batched dataset $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ with $|\mathcal{D}_f| = m$; algorithms $(\mathcal{A}, \mathcal{U})$; calibration rounds Γ
Ensure: Mechanism $\mathcal{M} : \mathcal{S}_m \to \mathcal{S}_m$

// CALIBRATION PHASE (RUN ONCE, OFFLINE)
for all $\widetilde{S} \in \mathcal{S}_m$ **do**
  for $\gamma = 1, \dots, \Gamma$ **do**
    $f^{(\gamma)} \leftarrow \mathcal{U}\big(\mathcal{A}(\mathcal{D}(\widetilde{S})),\, \mathcal{D}(\widetilde{S}),\, \mathcal{D}_f(\widetilde{S})\big)$
    $s^{(\gamma)}_{x,y}(\widetilde{S}) \leftarrow \phi\big(f^{(\gamma)}(x)_y\big)$ for all $(x, y) \in \mathcal{D}_f$
  end for
  For each $(x, y) \in \mathcal{D}_f$, fit Gaussian:

$$(\mu_{x,y}(\widetilde{S}),\, \sigma_{x,y}(\widetilde{S})) \;\leftarrow\; \mathrm{fit}\Big(\big\{s^{(\gamma)}_{x,y}(\widetilde{S})\big\}_{\gamma=1}^{\Gamma}\Big)$$

end for
// MECHANISM $\mathcal{M}(S)$ (APPLIED AT EVALUATION)
  $f \leftarrow \mathcal{U}\big(\mathcal{A}(\mathcal{D}(S)),\, \mathcal{D}(S),\, \mathcal{D}_f(S)\big), \quad s_{x,y} \leftarrow \phi(f(x)_y)$
  For each $\widetilde{S} \in \mathcal{S}_m$, compute

$$\Lambda(\widetilde{S}) \;\leftarrow\; \sum_{(x,y)\in\mathcal{D}_f} \log \mathcal{N}\Big(s_{x,y};\, \mu_{x,y}(\widetilde{S}),\, \sigma_{x,y}(\widetilde{S})\Big)$$

  return $\widehat{S} \;\leftarrow\; \arg\max_{\widetilde{S}\in\mathcal{S}_m} \Lambda(\widetilde{S})$

D.4 PAIRWISE AUDITOR (BASELINE)

Nasr et al. (2021) audit DP by having the adversary distinguish two models trained on neighbouring datasets. We adapt this baseline to unlearning by distinguishing the two unlearned models $f_u^{(i)} := \mathcal{U}\big(\mathcal{A}(\mathcal{D}_r \cup \mathcal{D}_f^{(i)}),\, \mathcal{D}_r \cup \mathcal{D}_f^{(i)},\, \mathcal{D}_f^{(i)}\big)$ for $i \in \{1, 2\}$, corresponding to two distinct forget sets $\mathcal{D}_f^{(1)}, \mathcal{D}_f^{(2)}$. Calibration and prediction are identical to Instantiation II with $m = 2$ and $\mathcal{D}_{f,j} = \mathcal{D}_f^{(j)}$; only the conversion from the confusion matrix to a privacy lower bound differs.

Let P_i denote the distribution of one independently generated $f_u^{(i)}$, and let A be the event that the calibrated attack predicts class 1. Thus

$$P_1(A) = \mathrm{TPR}, \qquad P_0(A) = \mathrm{FPR}, \qquad P_1(A^c) = \mathrm{FNR}, \qquad P_0(A^c) = \mathrm{TNR}.$$

Suppose the independent evaluation set contains n_1 draws from P_1 and n_0 draws from P_0, of which k_1 and k_0 are classified as class 1, respectively. Writing $\mathrm{Beta}^{-1}(q; a, b)$ for the q-quantile of a $\mathrm{Beta}(a, b)$ distribution, define the one-sided Clopper–Pearson endpoints

$$\mathrm{TPR}^{\mathrm{low}} := \mathrm{Beta}^{-1}\left(\frac{\zeta}{2}; k_1, n_1 - k_1 + 1\right),$$
$$\mathrm{FPR}^{\mathrm{high}} := \mathrm{Beta}^{-1}\left(1 - \frac{\zeta}{2}; k_0 + 1, n_0 - k_0\right),$$

with the usual boundary conventions $\mathrm{TPR}^{\mathrm{low}} = 0$ when $k_1 = 0$ and $\mathrm{FPR}^{\mathrm{high}} = 1$ when $k_0 = n_0$. Each endpoint has error probability $\zeta/2$, so a union bound gives simultaneous coverage at least $1 - \zeta$. Set

$$\mathrm{TNR}^{\mathrm{low}} := 1 - \mathrm{FPR}^{\mathrm{high}}, \qquad \mathrm{FNR}^{\mathrm{high}} := 1 - \mathrm{TPR}^{\mathrm{low}}.$$

Lower bound under (ε, δ)-indistinguishability. The pairwise auditor of Nasr et al. (2021) reports the lower bound

$$\varepsilon_{\mathrm{LB}}^{(p)} = \max\left\{\log \frac{1 - \delta - \mathrm{FPR}^{\mathrm{high}}}{\mathrm{FNR}^{\mathrm{high}}},\; \log \frac{1 - \delta - \mathrm{FNR}^{\mathrm{high}}}{\mathrm{FPR}^{\mathrm{high}}}\right\}, \tag{21}$$

where a term is omitted when its logarithm does not have a positive argument. To see why this is a lower bound, suppose P_0 and P_1 are (ε, δ)-indistinguishable. Applying the defining inequalities to A in the direction P_1 versus P_0, and to A^c in the reverse direction, gives

$$1 - \mathrm{FNR} \le e^{\varepsilon}\mathrm{FPR} + \delta,$$
$$1 - \mathrm{FPR} \le e^{\varepsilon}\mathrm{FNR} + \delta.$$

Rearranging gives the population version of Equation (21). On the simultaneous Clopper–Pearson coverage event, replacing FPR and FNR by their upper confidence bounds can only decrease both lower-bound expressions. Therefore $\varepsilon_{\mathrm{LB}}^{(p)}$ is a one-sided $(1-\zeta)$ lower confidence bound on the smallest ε for which the two distributions are (ε, δ)-indistinguishable.

Lower bound under ρ-zCDP-certified unlearning. We now derive an analogous lower bound for the zCDP parameter. Under a putative ρ-zCDP-certified-unlearning guarantee, Lemma 4.1 gives, in both directions,

$$D_\gamma(P_i \| P_j) \le \varepsilon_\gamma^{\mathrm{loc}}(\rho) = 4\rho\gamma, \qquad \gamma > 1.$$

For any event E, the Rényi change-of-measure inequality used in Equation (5), now applied to one test draw, gives

$$P_i(E) \le \exp\left(\frac{\gamma-1}{\gamma}\varepsilon_\gamma^{\mathrm{loc}}(\rho)\right) P_j(E)^{(\gamma-1)/\gamma}.$$

Taking $(i, j, E) = (1, 0, A)$ and rearranging yields

$$\rho \ge \frac{\frac{\gamma}{\gamma-1}\log \mathrm{TPR} - \log \mathrm{FPR}}{4\gamma}. \tag{22}$$

Applying the same argument with $(i, j, E) = (0, 1, A^c)$ yields the symmetric TNR/FNR expression. Accordingly, for every $\gamma > 1$, define

$$b_\gamma^{(+)} := \frac{\gamma}{\gamma-1}\log \mathrm{TPR}^{\mathrm{low}} - \log \mathrm{FPR}^{\mathrm{high}},$$
$$b_\gamma^{(-)} := \frac{\gamma}{\gamma-1}\log \mathrm{TNR}^{\mathrm{low}} - \log \mathrm{FNR}^{\mathrm{high}}.$$

The second expression applies the same inequality to A^c after reversing P_0 and P_1. With the convention that a term having a zero lower confidence bound contributes no positive lower bound, the pairwise auditor reports

$$\boxed{\rho_{\mathrm{LB}}^{(p)} := \sup_{\gamma>1} \frac{\max\{0, b_\gamma^{(+)}, b_\gamma^{(-)}\}}{4\gamma}.} \tag{23}$$

On the simultaneous Clopper–Pearson coverage event, replacing TPR by its lower bound and FPR by its upper bound can only decrease the right-hand side of Equation (22); the analogous monotonicity holds for the TNR/FNR expression. The same coverage event works for every γ, so optimizing over γ requires no additional multiple-testing correction. Therefore $\rho_{\mathrm{LB}}^{(p)}$ is a one-sided $(1-\zeta)$ lower confidence bound on the smallest valid zCDP parameter.

The sample counts n_0, n_1 (equal to L in our balanced experiments) enter only through the Clopper–Pearson endpoints. There is no factor of L multiplying $\varepsilon_\gamma^{\mathrm{loc}}$ above because each entry of the confusion matrix classifies one mechanism release. Such a factor would be appropriate only if a single attack decision jointly observed an L-fold product release; in that case the resulting lower bound on the per-release ρ would be divided by L.

Why the baseline is vacuous at deep-model run counts. This dependence on L alone is what makes the pairwise auditor uninformative in Section 7.3. At the deep-model budget of $2L = 10$ executions, i.e. $L = 5$ per hypothesis, and $\zeta = 0.05$, the endpoints are $\mathrm{TPR}^{\mathrm{low}} = \mathrm{TNR}^{\mathrm{low}} = 0.4782$ and $\mathrm{FPR}^{\mathrm{high}} = \mathrm{FNR}^{\mathrm{high}} = 0.5218$ even when the empirical confusion matrix is perfect, so both the pure-DP and zCDP bounds vanish before any property of the mechanism is consulted. The same auditor is competitive once runs are cheap enough to make L large, as in Section 8.

Method \ r	200		1000		2000		3000		4000	
	Logit	Loss	Logit	Loss	Logit	Loss	Logit	Loss	Logit	Loss
uni. (gradient ascent)	13.44	7.39	38.86	28.97	40.30	33.95	35.68	31.05	30.18	26.58
adv. (gradient ascent)	11.42	11.22	33.69	33.32	41.55	41.19	38.56	38.38	31.08	31.21
uni. (GD, $q = 1$)	13.31	6.25	37.58	31.31	44.05	38.16	38.36	36.03	31.00	29.84
adv. (GD, $q = 1$)	0.01	0.01	0.01	0.21	0.00	0.17	0.00	0.11	0.00	0.11
uni. (GD, $q = 2$)	13.00	7.92	40.07	33.15	46.11	40.19	42.45	37.66	35.33	32.51
adv. (GD, $q = 2$)	1.48	1.35	2.86	2.85	2.27	2.37	2.02	1.99	1.49	1.53
uni. (GD, $q = 4$)	13.18	8.73	41.90	35.98	50.88	42.19	46.17	40.08	37.61	34.11
adv. (GD, $q = 4$)	5.12	4.83	12.00	11.83	13.10	13.17	11.63	11.90	9.20	9.15

Table 6: ρ-zCDP lower bounds from logit score versus cross-entropy loss, across gradient ascent and gradient difference (GD) at varying q, for both uniform (uni.) and adversarial (adv.) splits.

Method \ r	200		1000		2000		3000		4000	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
uni. (GA)	13.44	6.36	38.86	18.60	40.30	19.64	35.68	17.08	30.18	14.75
adv. (GA)	11.42	5.95	33.69	17.00	41.55	20.59	38.56	19.15	31.08	15.12
uni. (GD, $q = 1$)	13.31	6.36	37.58	18.15	44.05	21.88	38.36	18.66	31.00	14.50
adv. (GD, $q = 1$)	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
uni. (GD, $q = 2$)	13.00	6.58	40.07	18.60	46.11	23.33	42.45	20.82	35.33	17.51
adv. (GD, $q = 2$)	1.48	0.61	2.86	1.43	2.27	1.12	2.02	0.89	1.49	0.77
uni. (GD, $q = 4$)	13.18	6.36	41.90	20.44	50.88	25.31	46.17	22.48	37.61	17.85
adv. (GD, $q = 4$)	5.12	2.26	12.00	6.01	13.10	6.60	11.63	5.76	9.20	4.53

Table 7: ρ-zCDP lower bounds from mean versus median aggregation of overlap scores, across gradient ascent (GA) and gradient difference (GD) at varying q, for both uniform (uni.) and adversarial (adv.) splits.

E Robustness of the auditor: alternative score and aggregation

In this appendix we sanity-check two design choices of our auditor. First, we replace the logit score $\phi(f(x)_y)$ used throughout the paper with the per-example cross-entropy loss and recompute the lower bounds (Appendix E.1). Second, we replace the mean aggregation of overlap scores with the median (Appendix E.2). In both cases the directional trends with respect to the support size r are preserved, indicating that our reported bounds are not artefacts of a particular score function or aggregation rule.

E.1 Logit score versus cross-entropy score

Table 6 compares the lower bounds obtained when the auditor's per-example score is the cross-entropy loss versus the logit score, for gradient ascent and gradient difference (GD) at $q \in \{1, 2, 4\}$ on both the uniform and adversarial splits. The two scores yield similar directional trends in r, with the logit score giving slightly tighter bounds in most settings.

E.2 Mean versus median aggregation

Table 7 compares the lower bounds obtained from the mean and median aggregations of the overlap scores. The two aggregations agree to within a few percent in nearly all settings, and both reproduce the inverted-U dependence on r noted in Appendix F — the lower bound first increases with r and then decreases.

F VARYING BATCH SIZES B (AND EQUIVALENTLY m) AND REPORTED SUPPORT SIZE r

This appendix gives the detailed analysis of how the lower bound ρ_{LB} depends on the number of forget batches m (equivalently the audit batch size B, since $m = |\mathcal{D}_f|/B$) and the reported support size r. The compact takeaways are summarised in Remark 4.4.

F.1 MAXIMUM ATTAINABLE BOUND UNDER PERFECT PREDICTION (IN m)

The maximum bound implied by a perfect prediction — overlap score m with support size $r = m$ — grows with m, and for ρ_{LB} the ceiling has a closed form. A perfect prediction makes the inner Chernoff term of Lemma 4.2 exact, and at $r = m$ the chance-overlap mass is $\pi(m) = 1/|\mathcal{S}_m|$; substituting $\varepsilon_\gamma^{\mathrm{loc}} = 4\rho\gamma$ and optimising over γ gives

$$\rho_{\mathrm{LB}}^{\max} = \frac{1}{4}\left(\sqrt{\log|\mathcal{S}_m|} - \sqrt{\tfrac{1}{L}\log\tfrac{1}{\zeta}}\right)^2.$$

At $L = 10$ and $\zeta = 0.05$ this is 0.35 at $m = 6$ and 14.53 at $m = 100$, and since $\log|\mathcal{S}_m| = \log\binom{m}{\lfloor m/2\rfloor}$ grows like $m\log 2$, the ceiling grows linearly in m. Intuitively, identifying the correct sign vector is harder when m is large, since the number of candidates $|\mathcal{S}_m|$ grows exponentially in m. A perfect prediction at large m therefore reflects much stronger distinguishing power, and certifies a correspondingly larger bound.

The pure-DP bound of Lemma C.3 behaves the same way: with $L = 10$ independent runs it yields $\varepsilon_{\mathrm{LB}} \geq 4.00$ at $m = 6$ and $\varepsilon_{\mathrm{LB}} \geq 67.87$ at $m = 100$.

F.2 CHOICE OF AUDIT BATCH SIZE B

Since $m = |\mathcal{D}_f|/B$, smaller batches enlarge the hypothesis space $\mathcal{S}_m$ and raise the attainable lower bound — but only if the adversary can pick the right hypothesis. Uncertified algorithms admit a strong per-batch signal, so small B (large m) is preferable. Certified algorithms have tight guarantees and weak per-batch signal; here, large m leaves the adversary unable to beat random guessing among exponentially many candidates, and one must use a larger B (smaller m) to obtain any nonzero bound. In addition, as discussed in Section 5, we use the *joint sign vector* prediction for certified unlearning algorithms, which is precisely why having a small batch size (large m) is computationally infeasible. Thus, in Section 7 we stick to $B = 1$ for uncertified algorithms and $B = 750$ for certified algorithms on CIFAR-100.

F.3 DEPENDENCE ON THE SUPPORT SIZE r

The support size r lets the auditor abstain on uncertain batches, improving the overlap on those it does report. Given m, the maximum bound *conditional on perfect overlap* grows with r: at $m = 100$, the zCDP ceiling is 8.60 at $r = 60$, rising to 14.53 at $r = 100$, obtained from the expression above with $\log|\mathcal{S}_m|$ replaced by $\log\binom{m}{\lfloor m/2\rfloor} - \log\binom{m-r}{\lceil (m-r)/2\rceil}$. The pure-DP ceiling follows the same curve, $\varepsilon_{\mathrm{LB}} \geq 40.83$ at $r = 60$ rising to $\varepsilon_{\mathrm{LB}} \geq 67.87$ at $r = 100$. In practice, the empirical lower bound traces an inverted-U in r. It first increases for two compounding reasons: the maximum attainable bound at perfect overlap grows with r, and the auditor's reported predictions remain accurate while there are still confident batches to report. As r grows further, however, perfect overlap becomes harder to attain — the auditor is forced to commit to low-confidence batches, which dilute the overlap score and pull the bound back down.

In Table 8, we display the lower bound against $r \cdot B$ (equivalently, $4500\,r/m$, since $m = |\mathcal{D}_f|/B = 4500/B$) for varying batch sizes $B \in \{1, 10, 100, 500\}$ on the ascent-on-forget unlearning algorithm. For both $B = 1$ and $B = 10$, the lower bound first increases with r and then decreases. This inverted-U effect is more pronounced at smaller batch sizes (larger m).

F.4 BATCHWISE INCLUSION/EXCLUSION VERSUS JOINT SIGN-VECTOR PREDICTION

The choice of auditor instantiation depends on whether the unlearning algorithm is certified. For certified methods such as model clipping (Koloskova et al., 2025) and R2D (Mu & Klabjan, 2025),

$r \cdot B$ \ B	$B = 1$ ($m = 4500$)	$B = 10$ ($m = 450$)	$B = 100$ ($m = 45$)	$B = 500$ ($m = 9$)
100	7.27	0.97	N/A	N/A
200	13.44	2.28	0.10	N/A
400	22.51	5.05	0.30	N/A
1000	38.86	13.00	1.06	0.09
2000	40.30	23.10	2.46	0.27
3000	35.68	29.33	3.91	0.48
3500	32.77	30.53	N/A	N/A
4000	30.18	29.72	5.35	0.68

Table 8: ρ-zCDP lower bounds across batch sizes for the ascent on forget set setting (uniform splitting).

Gaussian noise injected during unlearning makes the in- and out-distributions for a single batch nearly indistinguishable, so per-batch likelihood ratios rarely yield overlap scores above the random baseline of $r/2$. We therefore use the *joint sign-vector* predictor, which calibrates a separate Gaussian for every candidate $\widetilde{S} \in \mathcal{S}_m$ and aggregates evidence across all forget points before committing to a single sign vector. Since this requires independent calibration runs for each of the $|\mathcal{S}_m|$ candidates, we restrict to small m — concretely, batch size $B = 750$ giving $m = 6$ forget batches. For uncertified methods, the per-batch signal is strong enough that this issue does not arise; we obtain the strongest bounds at large m, where $|\mathcal{S}_m|$ is astronomically large and enumeration is infeasible, and so default to the *batchwise inclusion/exclusion* predictor.

Experiment with batchwise inclusion/exclusion. We also evaluated the *batchwise inclusion/exclusion* auditor on the uncertified variant of model clipping, with $C_2 = 5$ and $\sigma = 10^{-4}$ as in Section 7.1. We used batch sizes $B = 750$ and $B = 100$, with $r = 6$ and $r = 40$, respectively. The resulting ρ-zCDP lower bounds were only 0.002 and $1e-5$ even just after the *first* step of unlearning. We believe this is because, after noise is added to the model at each step, the predictions of the resulting models under inclusion and exclusion of a forget batch start to overlap substantially. Once this overlap becomes large, distinguishing the two cases through batchwise membership inference becomes difficult, leading to only a negligible lower bound.

G Privacy certificates of certified unlearning algorithms

This appendix records how the published certificates for gradient clipping, model clipping, and rewind-to-delete (R2D) induce zCDP and Gaussian-DP certificates. We distinguish guarantees stated in the original papers from consequences derived here. All comparisons are conditional on the public information in our threat model and are with respect to the common reference algorithm in the corresponding certified-unlearning definition. Langevin unlearning is mentioned only in the main text because its published two-sided Rényi bounds already yield a zCDP certificate directly.

G.1 Certificate-transfer principles

Recall from Definition C.5 that a pair P, Q is μ-GDP when its hypothesis-testing trade-off dominates G_μ. The following implication holds for every GDP pair, not only when P and Q are themselves Gaussian.

Lemma G.1 (GDP implies zCDP)**.** *If P, Q are μ-GDP, then, for every $\alpha > 1$,*

$$D_\alpha(P\|Q), D_\alpha(Q\|P) \le \frac{\alpha\mu^2}{2}.$$

Consequently, μ-GDP-certified unlearning is also ρ-zCDP-certified unlearning for

$$\rho = \frac{\mu^2}{2}.$$

Proof. Let $P_0 = \mathcal{N}(0, 1)$ and $P_1 = \mathcal{N}(\mu, 1)$. The GDP assumption is $T(P, Q) \ge T(P_0, P_1)$. By Blackwell's theorem (Dong et al., 2022, Theorem 2), there is a common randomized post-processing

kernel mapping (P_0, P_1) to (P, Q). Rényi data processing and the Gaussian-shift identity (Mironov, 2017, Proposition 7) give

$$D_\alpha(P\|Q) \le D_\alpha(P_0\|P_1) = \frac{\alpha\mu^2}{2}.$$

Applying the same argument in the reverse direction gives $D_\alpha(Q\|P) \le D_\alpha(P_1\|P_0) = \alpha\mu^2/2$. □

Two additional extraction rules will be useful. First, a two-sided Rényi curve ε_α yields the zCDP coefficient

$$\rho_{\mathrm{RDP}} := \sup_{\alpha>1} \frac{\varepsilon_\alpha}{\alpha}. \tag{24}$$

In particular, $D_\alpha(P\|Q), D_\alpha(Q\|P) \le \alpha R$ for every $\alpha > 1$ gives R-zCDP. Second, if a two-sided hockey-stick profile obeys

$$E_\varepsilon(P\|Q), E_\varepsilon(Q\|P) \le h(\varepsilon) \quad \text{for every } \varepsilon \ge 0,$$

then its Gaussian envelope

$$\mu_h := \inf\{\mu \ge 0 : h(\varepsilon) \le \theta_\varepsilon(\mu) \text{ for every } \varepsilon \ge 0\} \tag{25}$$

is a valid GDP parameter by the primal–dual characterization in Dong et al. (2022, Proposition 6). Here E_ε is the hockey-stick divergence from Equation (10), and θ_ε is the exact Gaussian privacy profile from Equation (11).

G.2 Gradient clipping

Theorem A.9 of Koloskova et al. (2025) concerns the *gradient-clipping* algorithm. Let

$$c_t := 1 - \gamma_t\lambda, \qquad s_t := 2\gamma_t C_1, \qquad w_t := \prod_{k=t+1}^{T-1} c_k, \qquad w_{-1} := \prod_{k=0}^{T-1} c_k,$$

where an empty product equals one. The total propagated displacement is

$$B_{\mathrm{GC}} := 2C_0 w_{-1} + \sum_{t=0}^{T-1} w_t s_t.$$

The theorem states that, for any $a_t \ge 0$ satisfying $\sum_{t=0}^{T-1} w_t a_t = B_{\mathrm{GC}}$,

$$D_q(P_u\|P_r) \le q \sum_{t=0}^{T-1} \frac{a_t^2}{2\sigma_t^2}.$$

The same argument applies after interchanging the two trajectories. Optimizing over (a_t) gives the published Rényi certificate in the form

$$D_q(P_u\|P_r), D_q(P_r\|P_u) \le qR_{\mathrm{GC}}, \qquad R_{\mathrm{GC}} := \frac{B_{\mathrm{GC}}^2}{2\sum_{t=0}^{T-1} w_t^2\sigma_t^2}. \tag{26}$$

Therefore gradient clipping is ρ_{GC}-zCDP-certified with $\rho_{\mathrm{GC}} = R_{\mathrm{GC}}$. This is a direct use of the paper's Rényi bound, not a conversion from its reported (ε, δ) guarantee.

For completeness, the algorithm also admits a generally looser GDP certificate that we derive here. At its first noisy step, the two Gaussian centres are separated by at most

$$\Delta_{\mathrm{GC}}^{(1)} := 2c_0C_0 + s_0.$$

Thus that step is $\mu_{\mathrm{GC}}^{(1)}$-GDP for

$$\mu_{\mathrm{GC}}^{(1)} := \frac{\Delta_{\mathrm{GC}}^{(1)}}{\sigma_0}. \tag{27}$$

All later steps use only the common retain set and are randomized post-processing, so this GDP certificate is preserved. By Lemma G.1, it also yields the valid, but typically weaker, zCDP value $(\mu_{\mathrm{GC}}^{(1)})^2/2$.

Method	Published certificate	GDP parameter	zCDP parameter
Gradient clipping	Rényi, Theorem A.9	$\mu_{\mathrm{GC}}^{(1)}$ (derived, coarse)	R_{GC}
Model clipping	Hockey-stick, Theorem 4.2	μ_{MC} (derived envelope)	$\mu_{\mathrm{MC}}^2/2$
R2D	Gaussian mechanism, Theorem 3.1	$\Delta_{\mathrm{R2D}}/\sigma$	$\Delta_{\mathrm{R2D}}^2/(2\sigma^2)$

Table 9: Certificates for the algorithms discussed in detail in this appendix. "Derived" marks a consequence established here from the published distributional bound rather than a parameter explicitly reported by the original paper.

G.3 Model clipping

Model clipping is analyzed separately in Theorem 4.2 of Koloskova et al. (2025). Define

$$r_0 := \frac{2C_0}{\sigma_0}, \qquad r_2 := \frac{2C_2}{\sigma}.$$

The proof of that theorem, using its Lemmas A.2–A.3, gives the complete two-sided hockey-stick profile

$$E_\varepsilon(P_u\|P_r), E_\varepsilon(P_r\|P_u) \le h_{\mathrm{MC}}(\varepsilon) := \theta_\varepsilon(r_0)\theta_\varepsilon(r_2)^T. \tag{28}$$

The paper uses this profile to report an (ε, δ) guarantee. To obtain a single GDP parameter instead, define

$$\mu_{\mathrm{MC}} := \inf\left\{\mu \ge 0 : h_{\mathrm{MC}}(\varepsilon) \le \theta_\varepsilon(\mu) \text{ for every } \varepsilon \ge 0\right\}. \tag{29}$$

Then model clipping is μ_{MC}-GDP-certified and, by Lemma G.1, is ρ_{MC}-zCDP-certified with

$$\rho_{\mathrm{MC}} = \frac{\mu_{\mathrm{MC}}^2}{2}. \tag{30}$$

This GDP-envelope step is our consequence of the paper's full profile. A simple closed-form certificate follows from $h_{\mathrm{MC}}(\varepsilon) \le \theta_\varepsilon(r_0)$:

$$\mu_{\mathrm{MC}} \le \frac{2C_0}{\sigma_0}, \qquad \rho_{\mathrm{MC}} \le \frac{2C_0^2}{\sigma_0^2}.$$

Numerically evaluating Equation (29) can retain the amplification from all T model-clipping steps and can be substantially tighter than this closed-form value.

G.4 Rewind-to-delete

R2D adds isotropic Gaussian noise of variance σ^2 to both the trained and unlearned releases. Under the conditions of Theorem 3.1 of Mu & Klabjan (2025), the proof bounds the distance between their unperturbed centres by

$$\Delta_{\mathrm{R2D}} := \frac{2mG}{Ln}\left[\left(1 + \frac{\eta L n}{n-m}\right)^{T-K} - 1\right](1+\eta L)^K. \tag{31}$$

Consequently, the released distributions are equal-covariance Gaussian shifts with

$$\mu_{\mathrm{R2D}} = \frac{\Delta_{\mathrm{R2D}}}{\sigma}, \qquad \rho_{\mathrm{R2D}} = \frac{\Delta_{\mathrm{R2D}}^2}{2\sigma^2}. \tag{32}$$

The first identity is the GDP parameter for the worst-case centre separation, and the second follows either from Lemma G.1 or directly from the Rényi divergence between equal-covariance Gaussians. The original theorem calibrates σ to state an (ε, δ) guarantee; Equation (32) records the GDP and zCDP guarantees implicit in the same Gaussian release.

G.5 Summary of certified algorithms

Each zCDP parameter in Table 9 can be used with the main audit through Lemma 4.1; each GDP parameter can instead be audited directly using Lemma C.6. These are alternative parameterizations of an upper-bound certificate. A lower bound returned by one audit should not be relabeled as a lower bound on the other privacy parameter merely by applying an upper-bound conversion.

H DATASETS, MODELS, AND AUDITOR BATCHING

This appendix gives the full data-construction and model details for the three datasets described in Section 6.

H.1 CIFAR-100

Forget/retain construction. For CIFAR-100 (Krizhevsky, 2009), we first split the original training set into a training portion (90%) and a validation portion (10%). From the training portion, we designate 10% of the points as the forget set $\mathcal{D}_f$ ($4{,}500$ points), and use the remaining $40{,}500$ points as the retain set $\mathcal{D}_r$. We consider two types of splits. In the *uniform* split, the forget and retain sets are formed by sampling points uniformly at random. In the *adversarial* split, the forget and retain sets are chosen from largely disjoint CIFAR-100 label classes, with overlap allowed in at most one label class. We then independently shuffle the retain and forget sets, and partition both $\mathcal{D}_r$ and $\mathcal{D}_f$ into batches of size B.

Auditor batching. For uncertified unlearning algorithms we use $B = 1$, while for certified unlearning algorithms we use $B = 750$. Accordingly, we use the *batchwise inclusion/exclusion* auditor (Algorithm 2) for uncertified methods and the *joint sign-vector* auditor (Algorithm 3) for certified methods. A discussion of why this split between auditor instantiations is appropriate is given in Appendix F.4, with a more detailed study in Appendix F.

Model. We use a `TinyNetCIFAR100` model, a lightweight CNN that attains around 55% accuracy on the CIFAR-100 test set after 400 epochs of training. The architecture consists of two convolutional blocks, each containing two 3×3 convolutional layers (with 64 and 128 filters respectively) with Group Normalization (8 groups) and ReLU activations, followed by 2×2 max pooling. After the two blocks, global average pooling collapses the spatial dimensions, and a single linear layer maps to the 100-class output. All weights are Kaiming-initialized. The model contains no dropout or batch normalization; Group Normalization is used throughout for training stability. We use a cosine or one-cycle learning-rate scheduler with stochastic gradient descent optimizer.

H.2 SHAKESPEARE

Forget/retain construction. Following McMahan et al. (2017), we treat each speaking role in each play as a separate client, parsing speaker labels from the text to extract per-role dialogue. Roles with fewer than 2 dialogue blocks are discarded, and for our experiments we randomly subsample 300 roles. Each role's dialogue blocks are split chronologically into train (80%), validation (10%), and test (10%); validation and test sets are formed by aggregating across all 300 roles. The training data is then partitioned into a retain set $\mathcal{D}_r$ and a forget set $\mathcal{D}_f$ *at the role level* — each role is assigned entirely to one set, ensuring zero character overlap between $\mathcal{D}_r$ and $\mathcal{D}_f$. Roles are randomly shuffled and greedily assigned to $\mathcal{D}_f$ until approximately 10% of training characters are accumulated, yielding 33 forget roles ($\mathcal{D}_f$, 135,300 characters, $\sim$10.0% of train) and 267 retain roles ($\mathcal{D}_r$, 1,221,363 characters, $\sim$90.0% of train).

For unlearning experiments, the forget set $\mathcal{D}_f$ is further partitioned into $B = 400$ equal batches: the dialogue blocks from all forget roles are shuffled at the block level (preserving within-block coherence) and split into 400 contiguous chunks $\mathcal{D}_{f,1}, \ldots, \mathcal{D}_{f,400}$. For each run, 200 forget chunks are sampled uniformly at random from the 400. Each chunk and the retain set are independently converted to sliding-window samples of length 80 characters, where x is the input sequence and y is the target sequence shifted by one character. The retain samples and sampled forget samples are then pooled and shuffled to form the training dataset. The retain and forget datasets used during evaluation are constructed by the same sliding-window procedure applied to the retain and sampled forget text respectively.

Auditor batching. On Shakespeare we evaluate only uncertified unlearning methods (gradient ascent, gradient difference, retain fine-tuning, and Hessian-based unlearning of Zhang et al. (2025)); we therefore use the *batchwise inclusion/exclusion* auditor (Algorithm 2) throughout, as on CIFAR-100.

Model. The model is a 2-layer stacked character-level LSTM following McMahan et al. (2017). Each input character is embedded into an 8-dimensional space, processed through two LSTM layers each with 256 hidden units, and projected to a softmax output layer over the vocabulary of $|\mathcal{V}| = 100$ characters. The full model has 825,220 parameters and is trained with an unroll length of $L = 80$ characters. For the Shakespeare dataset, the score $\phi(f(x)_y)$ for a sample (x, y) is the mean per-character cross-entropy loss $-\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t \mid x_{\le t})$ assigned by the model f, where $T = 80$ is the sequence length.

H.3 TOFU DATASET

Dataset. TOFU (Maini et al., 2024) is a synthetic question-answering corpus of 200 *fictitious* author profiles, each with 20 question–answer pairs about that author's life and work, for 4,000 rows in total (the `full` configuration of `locuslab/TOFU`). Because the authors are invented, this knowledge does not occur in natural text, and TOFU was released after the December 2023 pretraining cutoff that Meta reports for Llama-3.2-1B-Instruct. We therefore expect the model's competence on these facts to come from our fine-tuning, although we cannot verify what every post-training stage of the released model saw. This is what makes TOFU suitable for an unlearning audit in a way that natural-text corpora are not: the auditor's negative condition requires candidates the model is unlikely to have seen, which is far harder to argue for real-world data. Rows are stored in author-major order, so row i belongs to author $\lfloor i/20 \rfloor$. Questions and answers average 11.3 and 25.3 words respectively, so at our sequence length of 512 tokens no example is truncated.

Splitting. Unlike the role-level partition used on Shakespeare, the TOFU partition is *pinned* rather than randomly drawn: the candidate pool is authors 180–199, exactly TOFU's published `forget10` split, and the retain set $\mathcal{D}_r$ is authors 0–179 (3,600 QA pairs), exactly `retain90`. The reason is that TOFU's evaluation files are built around that boundary — the perturbed configurations required to compute truth ratios cover only authors 0–19 (`retain_perturbed`) and 180–199 (`forget10_perturbed`), leaving 160 authors permanently unscorable — so a random draw of 20 candidates cuts across it and leaves most candidates without perturbed data. Pinning makes the audit's partition coincide with the one the evaluation data presupposes, and does not affect the validity of the bound, which requires the candidate pool to be fixed before any run and the *sign vectors* to be i.i.d. uniform over $\mathcal{S}_m$ (Section 4), not the pool itself to be randomly drawn.

Auditor batching. The main experiments use singleton QA-pair batches ($B = 1$, $m = 400$), as summarised in Section 6. We additionally sweep $B \in \{20, 10, 4, 1\}$, corresponding to $m \in \{20, 40, 100, 400\}$. Batches are formed within authors and never combine QA pairs from different authors; thus $B = 20$ gives one author-level canary, whereas $B = 1$ gives one canary per QA pair. Because the sign vectors are exactly balanced, every configuration includes 200 candidate pairs and trains on the same 3,800 examples. Consequently, the sweep changes only the granularity of the audited hypothesis, not the training-set size. We use $\Gamma = 20$ calibration runs and $L = 10$ evaluation runs throughout.

Model. The model is `meta-llama/Llama-3.2-1B-Instruct` (Grattafiori et al., 2024), a 16-layer decoder-only transformer with a vocabulary of $|\mathcal{V}| = 128{,}256$ tokens and 1,235,814,400 parameters, which we fine-tune in full (no adapters) in `bfloat16` with gradient checkpointing and SDPA attention. Prompts are rendered with the model's own chat template. Critically, the prompt and the answer are tokenized *separately* and concatenated, so that the prompt/answer boundary is exact by construction rather than located post hoc in a templated string; all prompt and padding positions are masked out of the loss. For TOFU the score $\phi(f(x)_y)$ for a QA pair $z = (q, a)$ is the teacher-forced mean per-token cross-entropy over the answer tokens only,

$$\ell_z(f) \;=\; -\frac{1}{|a|}\sum_{t=1}^{|a|} \log p_f(a_t \mid q,\, a_{<t})\,, \tag{33}$$

computed in FP32 with the model in evaluation mode. Answers average 36.7 tokens. Per-example means are accumulated from per-example token sums rather than from a batch-level mean, so that examples are not implicitly weighted by answer length. Masking the prompt matters: were question tokens included, every candidate's score would be dominated by prompt perplexity, which barely depends on membership, and most of the attack signal would be lost.

Fine-tuning (training algorithm in our audtitng setup). Each run fine-tunes the base checkpoint on $\mathcal{D}(S)$ for 5 epochs with AdamW at a learning rate of 1×10^{-5} decayed linearly to zero with no warmup, weight decay 0.01, gradient-norm clipping at 1.0, a maximum sequence length of 512 tokens, and an effective batch size of 16 (micro-batch 2 with 8 gradient accumulation steps), giving $1{,}190$ optimizer steps over the $3{,}800$ examples. Examples are emitted in a *single fixed global permutation* of all $4{,}000$ rows, restricted to the examples present in that run, and this order is not reshuffled between epochs. Training loss falls from 1.695 after the first epoch to 1.372 at convergence, and a run takes $\sim$17.5 minutes on one A100-80GB at a peak of 11.6 GiB.

We describe the unlearning algorithms in detail in Appendix J.6.

I Certified Unlearning algorithms: hyperparameters and audits

This appendix gives the full setup for each unlearning algorithm audited in Section 7. Appendices I.1 and I.2 cover the certified algorithms; Appendices J and J.1 to J.4 cover the four general heuristic algorithms on CIFAR-100 and Shakespeare; and Appendix J.6 gives the corresponding objectives and hyperparameters for TOFU. Hyperparameter-sweep results are collected in Appendix L.

I.1 Model clipping

I.1.1 Algorithm

Let $\hat{x}$ denote the trained model produced by $\mathcal{A}$ prior to unlearning. The unlearning procedure of Koloskova et al. (2025) consists of two phases.

Phase 1: noisy projection. For $t = 0, 1, \ldots, T_{\text{noisy}} - 1$, the iterate is updated by

$$x_0 = \hat{x} + \xi_0, \qquad x_{t+1} = \Pi_{C_2}\big(x_t - \gamma\big(g_t + \lambda x_t\big)\big) + \xi_{t+1}, \tag{34}$$

where Π_{C_2} is projection onto the ℓ_2-ball of radius C_2, g_t is the gradient of the loss on the retain set evaluated at x_t, γ is the step size, λ is the ℓ_2 regularisation strength, and the noise satisfies $\xi_0 \sim \mathcal{N}(0, \sigma_0^2 \mathbb{I}_d)$ and $\xi_{t+1} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ for $t \geq 0$. The number of noisy steps T_{noisy} is set by Theorem 4.2 of Koloskova et al. (2025) so that the iterate $x_{T_{\text{noisy}}}$ satisfies the target (ε, δ) guarantee.

Phase 2: fine-tuning. Starting from $x_{T_{\text{noisy}}}$, the algorithm runs plain gradient descent on the retain-set loss — with no noise and no projection — until the total step budget is exhausted.

I.1.2 Hyperparameters

Following Koloskova et al. (2025), we train each model for 400 epochs using a batch size of 128, corresponding to 317 optimisation steps per epoch, and run each unlearning method for 50 epochs. For training, we use a one-cycle learning-rate schedule with a maximum learning rate of 0.06 and a regularisation coefficient of 5×10^{-4}. On CIFAR-100, we use the same training configuration for gradient ascent (GA), gradient difference (GD), and retain fine-tuning. For Hessian-based unlearning, we instead use a different learning rate and scheduling scheme, as detailed in Appendix J.

For unlearning Phase 1, we use a constant learning rate of 10^{-3} for the noisy steps; for Phase 2 we use a one-cycle schedule with maximum 0.1 and regulariser 5×10^{-4}. The initial clipping radius just before unlearning starts is $C_0 = 30$. The same regulariser is applied during the retain-set fine-tuning phase. The sweep over the certified $(\varepsilon, C_2, \sigma)$ grid is reported in Table 10.

I.1.3 Uncertified variant of model clipping

For the uncertified experiment of Figure 1a we set $C_2 = 5$ and $\sigma = 10^{-4}$, terminate the clipping/noise-addition phase after 10 epochs (3,170 update steps), and then fine-tune as above. The ratio C_2/σ is far larger than what Theorem 4.2 of Koloskova et al. (2025) admits, so no certified (ε, δ) guarantee applies; we report ρ_{LB} at every step t. To obtain tighter bounds in this regime we use $L = 500$ runs. For context, 1 epoch corresponds to 317 steps in this setup, so Figure 1a effectively covers up to roughly 2 epochs.

C	σ	ε	Noisy steps	ρ_{LB} (U/A)		$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$ (U/A)	
				Step 1	Epoch 50	Step 1	Epoch 50
0.2	0.1	1	40	0/0	0/0	0/0	0/0
0.1	0.02	10	N/A	0/0	0/0	0/0	0/0
0.225	0.001	10^5	4041	0/0	0/0	0/0	0/0

Table 10: Model-clipping audits under uniform (U) and adversarial (A) CIFAR-100 splits, evaluated at the first noisy step and after 50 epochs. The converted quantity $\varepsilon^{(c)}(\rho_{\mathrm{LB}})$ is the interpretive scale of Equation (19), not an ε confidence bound.

I.2 Rewind-to-delete

We take $T = 40$ training epochs and $K = 5$ rewind epochs. Following Mu & Klabjan (2025, Theorem 3.1), the unlearned model θ''_K is obtained by loading the checkpoint θ'_{T-K} and then training for K epochs on the retain set; final-step Gaussian noise of scale σ is added to ensure (ε, δ)-indistinguishability between the trained model θ'_T and θ''_K. Since we consider $\varepsilon > 1$, the noise parameter σ [6] must satisfy

$$\Phi\big(-\tfrac{\varepsilon\sigma}{\Delta} + \tfrac{\Delta}{2\sigma}\big) - e^{\varepsilon}\,\Phi\big(-\tfrac{\varepsilon\sigma}{\Delta} - \tfrac{\Delta}{2\sigma}\big) \le \delta,$$

where Φ is the standard Gaussian CDF and $\Delta = \|\theta'_T - \theta''_K\|$ is the sensitivity. Both T and K are kept small because the resulting σ grows exponentially in T and K. Theoretical guarantees for R2D hold only under full-batch gradient descent, so we use full-batch updates throughout; on CIFAR-100 this yields a base model with poor accuracy of around 4%.[7] In our auditing setup, we use a learning rate of 0.01. We swept over $\{0.1, 0.01, 0.001\}$; the resulting ρ_{LB} values were all small and did not differ meaningfully across the sweep, so we report results for 0.01 throughout.

Since unlearning just involves rewinding and Gaussian noise addition, we use $L = 20,000$ runs here as unlearning is quite cheap here.

Observations. In Figure 1b the audited bound rises monotonically with the target but stays orders of magnitude below it, reaching about 0.2 against claimed ρ up to 500; it is small but nonzero, and a non-rejection is not evidence that R2D unlearns. Figure 1a shows uniform splitting yielding larger bounds than adversarial, as for the uncertified methods of Section 7.2. Figure 2a shows the pairwise baseline beating the joint sign-vector predictor of Section 5 where the bound is small; this is specific to ρ_{LB} and does not carry over to the pure-DP bound $\varepsilon_{\mathrm{LB}}$ of Appendix C (see Figure 2b)

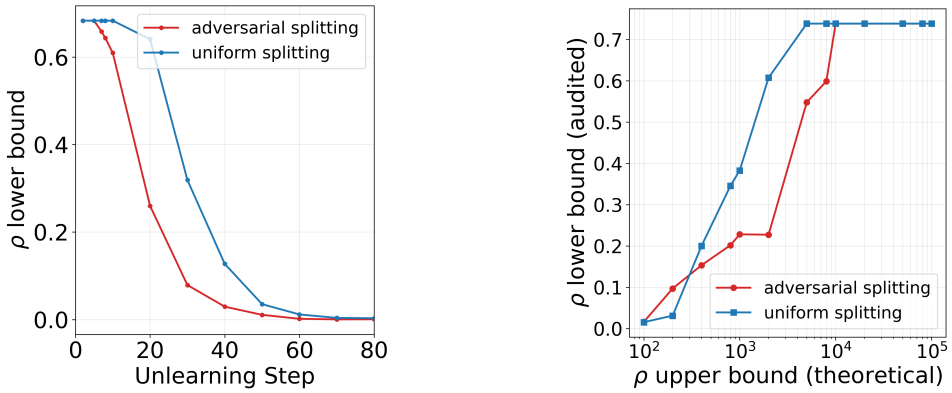
J Base training configurations for uncertified methods

Shakespeare. The base model is trained for 15 epochs using SGD with learning rate 0.1, batch size 256, and gradient clipping of 1.0, with no weight decay or learning-rate scheduling. The resulting accuracy is roughly 0.53, in line with (McMahan et al., 2017). Unlearning runs for a total of 8 epochs; the retain-set fine-tuning phase uses learning rate 0.1.

CIFAR-100. For Hessian-based unlearning, we use a cosine scheduler with learning rate 0.1 following Zhang et al. (2025), training for a total of 400 epochs. For the remaining three uncertified methods (gradient difference, gradient ascent, retain fine-tuning), the training configuration matches Appendix I.1; the retain-set fine-tuning phase reverts to a one-cycle schedule with maximum learning rate 0.1. For gradient ascent, gradient difference and retain finetuning, our training configuration is identical to the one described in Appendix I.1.3.

[6] To compute σ, one needs the smoothness constant L and a uniform bound on the gradient G, both estimated following Mu & Klabjan (2025).

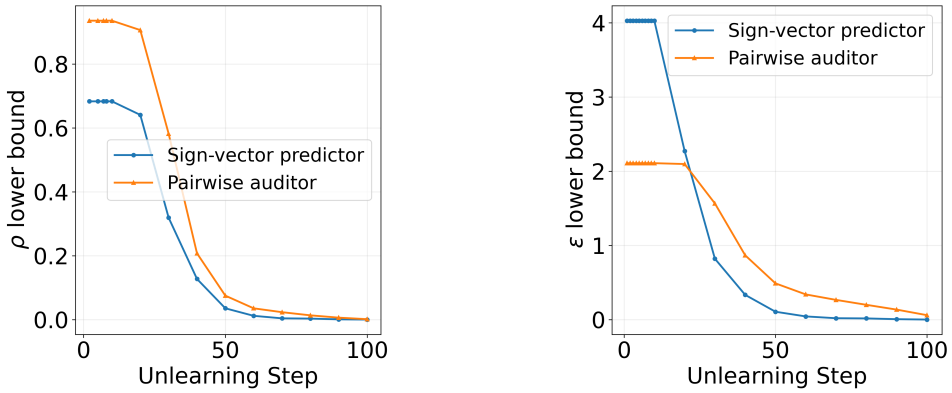
[7] Mu & Klabjan (2025) instead evaluate binary tasks—eICU with an MLP (test AUC 0.73) and Lacuna-100 with a ResNet-18 (test AUC 0.98)—using large mini-batches, outside the full-batch hypothesis. The small ρ_{LB} we report for R2D should be read with this weak base model in mind.



(a) Model Clipping diagnostic

(b) R2D across target values

Figure 1: diagnostic audits for model clipping and rewind-to-delete on CIFAR-100. The certified configurations have $\rho_{\mathrm{LB}} = \varepsilon^{(c)}(\rho_{\mathrm{LB}}) = 0$; these panels provide additional comparisons and intermediate behavior.



(a) ρ-zCDP lower bound

(b) ε lower bound

Figure 2: Diagnostic ρ-zcdp and ε lower bounds for model clipping for pairwise auditor vs sign vector predictor

J.1 Hessian-based unlearning

Update rule. Let g_f denote the forget-set gradient and H the (regularised) Hessian. The unlearning step is

$$\theta \;\mapsto\; \theta - H^{-1} g_f + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I}), \tag{35}$$

where σ is the final noise scale and the Gaussian perturbation is added to every parameter.

LiSSA approximation. As in Zhang et al. (2025), we approximate the inverse-Hessian–vector product $H^{-1}g$ via the LiSSA algorithm of Agarwal et al. (2017), avoiding explicit construction of the Hessian. The LiSSA approximation is governed by four hyperparameters — depth parameters s_1 and s_2, the `scale`, and the retain batch size — together with a weight decay applied during the Hessian computation. We describe each below before giving the per-dataset values.

Depth parameters s_1 and s_2. The parameter s_1 controls the number of independent LiSSA estimation runs whose results are averaged to reduce variance, while s_2 controls the depth of the Neumann-series recursion within each run, governing convergence to the true Newton direction.

Scale. The `scale` parameter must strictly upper-bound the largest eigenvalue of the regularised Hessian to ensure that the Neumann series converges. Too small a value causes divergence, while too large a value slows convergence and undershoots the Newton step.

Unlearning batch size. The unlearning batch size controls the size of the random retain mini-batches sampled at each recursion step, trading off gradient noise against computational cost per LiSSA iteration.

Weight decay. The weight decay enters as a squared L_2 regularisation term added to the loss during the Hessian computation, effectively shifting the Hessian by $\lambda \boldsymbol{I}$ and improving its conditioning for inversion. Crucially, it is applied only in the Hessian computation and not in the gradient g, so that the Newton update approximates $H^{-1}g$ where H is the regularised Hessian but g is the unregularised retain gradient.

Per-dataset values. For both datasets we set the weight decay to 5×10^{-4}. For CIFAR-100 we use $s_1 = 10$, $s_2 = 1000$, `scale` $= 1000$, and an unlearning batch size of 10. For Shakespeare we use $s_1 = 5$, $s_2 = 700$, `scale` $= 5000$, and an unlearning batch size of 256 (matching the training batch size); s_1 and s_2 are chosen slightly smaller for Shakespeare for computational efficiency.

Final noise. After the Hessian-based step, Gaussian noise of standard deviation 10^{-3} is added to every parameter, identically in both setups. The audit obtained by varying this final noise scale is reported in Appendix L.3.

J.2 Gradient difference (GD)

Update rule. Letting g_t^r and g_t^f denote the retain- and forget-loss gradients at step t, the GD update for the first E_f epochs is

$$x_{t+1} = x_t - \eta_t\big(g_t^r - \lambda\, \mathbf{1}\{t \equiv 0 \pmod{q+1}\}\, g_t^f\big), \tag{36}$$

so that the interleaved step (with the forget-gradient term active) fires once per $q+1$ updates and the remaining q updates within each cycle reduce to plain retain descent $x_{t+1} = x_t - \eta_t g_t^r$. After E_f epochs, only the latter (plain retain descent) is used.

CIFAR-100. We fix $\lambda = 0.75$, vary $q \in \{1, 2, 4\}$, and set $E_f = 10$. The learning rate is held constant at 10^{-3} during the interleaved phase, and reverts to a one-cycle schedule with maximum 0.1 during the retain-set fine-tuning phase.

Shakespeare. We vary $q \in \{1, 2, 4\}$, $E_f \in \{5, 7\}$, and $\lambda \in \{0.5, 1.0, 1.5\}$. The learning rate is held constant at 0.05 during the interleaved phase and held constant at 0.1 during the retain fine-tuning phase.

J.3 Gradient ascent

Update rule. The unlearning phase performs gradient ascent on the forget set,

$$x_{t+1} = x_t + \eta_t\, g_t^f, \tag{37}$$

where g_t^f is the gradient of the forget loss at step t. After a fixed number of epochs (2 for both CIFAR-100 and Shakespeare), the algorithm switches to plain retain-set fine-tuning.

CIFAR-100. The ascent learning rate is set to 10^{-3} during the unlearning phase, reverting to a one-cycle schedule with maximum 0.1 during the retain-set fine-tuning phase.

Shakespeare. The ascent learning rate is held constant at 0.05 during the unlearning phase and held at 0.1 during the retain fine-tuning phase, identical to the rate for gradient difference.

J.4 RETAIN FINE-TUNE

The unlearning phase consists entirely of retain fine-tuning, equivalent to a special case of model clipping with $C = C_2 = \infty$ and $\sigma = 0$. Learning-rate schedules match the corresponding retain-set fine-tuning phases described in Appendix J.

J.5 DISTILLATION-BASED UNLEARNING

The four distillation-based unlearners audited in Section 7 share a teacher–student template. Write $f_t := \mathcal{A}(\mathcal{D})$ for the original trained model, which plays the role of the *teacher*, and f_s for the *student*: the model currently being unlearned, initialised at f_t and returned as the unlearned model f_u once the objective below has been optimised. In the notation of the problem setup, these methods instantiate $f_u = \mathcal{U}\big(\mathcal{A}(\mathcal{D}), \mathcal{D}, \mathcal{D}_f\big)$ with $\mathcal{A}(\mathcal{D}) = f_t$, and the student is the iterate $\mathcal{U}$ maintains along the way. Write

$$p_\tau(f; x) := \mathrm{softmax}(f(x)/\tau)$$

for the temperature-τ softened predictive distribution. The KL terms in their objectives have the form

$$\mathcal{L}_{\mathrm{KL}}(f_s) = \sigma \underbrace{\mathbb{E}_{x\sim\mathcal{D}_f} \mathrm{KL}(q(x) \,\|\, p_\tau(f_s; x))}_{\text{forget term}} + \lambda \underbrace{\mathbb{E}_{x\sim\mathcal{D}_r} \mathrm{KL}(p_\tau(f_t; x) \,\|\, p_\tau(f_s; x))}_{\text{retain term}}, \tag{38}$$

where $\sigma = +1$ for BadTeacher and DELETE and $\sigma = -1$ for SCRUB and SCRUB+R. The latter two also add a retain-set task loss. The methods differ in the forget-set target $q(x)$ and in whether the retain term is present. For the audit, the relevant distinction is how the forget objective uses the teacher's logits at the same input x: it may ignore them, drive the student away from them, or define a target to match. Table 11 summarises these roles.

BadTeacher (Chundawat et al., 2023). A second, randomly initialised *incompetent* teacher f_b supplies the forget-set target, so $q(x) = p_\tau(f_b; x)$ in Equation (38) with $\lambda > 0$. Because f_b is never trained, its predictions are close to uniform and, crucially, independent of x's training history: the student is pulled towards the same noise target whether or not x was ever fitted.

SCRUB (Kurmanji et al., 2023). A single teacher is used and the forget term is *maximised* rather than minimised, giving the max–min objective

$$\min_{f_s} \left\{ \lambda\, \mathbb{E}_{x\sim\mathcal{D}_r} \mathrm{KL}(p_\tau(f_t; x) \,\|\, p_\tau(f_s; x)) + \alpha\, \mathbb{E}_{(x,y)\sim\mathcal{D}_r} \ell(f_s(x), y) - \mathbb{E}_{x\sim\mathcal{D}_f} \mathrm{KL}(p_\tau(f_t; x) \,\|\, p_\tau(f_s; x)) \right\}, \tag{39}$$

where ℓ is the task loss. Optimisation alternates: each epoch performs one *max-step* ascending the forget term, followed by a *min-step* descending the retain and task terms, with the max-steps switched off after a fixed number of epochs so that the final phase restores utility. Unlike BadTeacher, SCRUB uses $q(x) = p_\tau(f_t; x)$ in the forget KL term, but maximises that term instead of matching $q(x)$. The direction the student is pushed therefore depends on the teacher; the divergence is unbounded above and is stopped by the schedule rather than by convergence. We believe pushing away from the teacher in this way could remove the per-example structure our attack exploits, although the objective does not guarantee that the final student's outputs lose it.

SCRUB+R (Kurmanji et al., 2023). Identical objective and schedule to SCRUB; the two differ only in which iterate is returned. SCRUB returns the final iterate, whereas SCRUB+R *rewinds*: it retains the checkpoint from each epoch, computes on the forget set an error-based selection statistic, and returns the earliest checkpoint whose forget-set error first reaches the level expected of a model that never saw $\mathcal{D}_f$. The motivation is that running the max-step to completion can overshoot—driving forget-set error *above* the retrained reference and thereby reintroducing distinguishability in the opposite direction—so rewinding selects the crossing point instead. SCRUB+R therefore coincides with SCRUB exactly when the selection rule picks the final epoch.

DELETE (Zhou et al., 2025). Neither a second teacher nor $\mathcal{D}_r$ is used: $\lambda = 0$ in Equation (38), and the forget-set target is obtained by masking the teacher's logits on the forgotten classes and renormalising,

$$q(x)_c = \frac{\mathbf{1}\{c \notin \mathcal{C}_f\} \exp(f_t(x)_c/\tau)}{\sum_{c' \notin \mathcal{C}_f} \exp(f_t(x)_{c'}/\tau)}, \tag{40}$$

Method	Second teacher	Uses $\mathcal{D}_r$	Forget-set target $q(x)$	Role of $f_t(x)$ on $\mathcal{D}_f$
BadTeacher	Yes	Yes	$p_\tau(f_b; x)$, f_b random	None
SCRUB	No	Yes	$p_\tau(f_t; x)$, KL maximised	Pushes student away
SCRUB+R	No	Yes	$p_\tau(f_t; x)$, KL maximised	Pushes student away
DELETE	No	No	masked, renormalised f_t logits	Defines target to match

Table 11: Mechanisms of the four distillation-based unlearners. The last column describes how the forget-set objective uses the original teacher's logits at the same input. It excludes the teacher's role in student initialisation and the retain-set objective. SCRUB+R uses SCRUB's forget objective but selects a rewound checkpoint. See Appendix L.1 for the audited bounds across learning rates.

where $\mathcal{C}_f$ is the set of forgotten classes. A single objective on $\mathcal{D}_f$ thus supplies both forgetting (mass removed from $\mathcal{C}_f$) and retention (the relative ordering of the surviving classes is inherited from f_t). The target is a deterministic transform of the teacher's own logits at x, so it could carry the teacher's per-example structure into the student; we believe this is one reason our attack is most successful against DELETE (Appendix L.1), though the audit does not isolate this mechanism.

Fixed configuration. All four are run on CIFAR-100 under the adversarial split of Appendix H.1, with a base model trained for 200 epochs under a cosine learning-rate schedule. BadTeacher is run for one unlearning epoch and DELETE for 20, following the configurations of their respective papers. SCRUB is audited at its final iterate and SCRUB+R at the checkpoint its rewind rule selects. The unlearning learning rate is the only quantity swept; that sweep and its analysis are in Appendix L.1.

J.6 Unlearning algorithms on TOFU with Llama-3.2-1B-Instruct

All methods start from the *same* fine-tuned checkpoint for a given run and receive the run's forget set $\mathcal{D}_f(S)$ (the $S_j = +1$ batches, 200 QA pairs) and the fixed retain set $\mathcal{D}_r$ (3,600 QA pairs). We follow the OpenUnlearning reference implementations (Dorna et al., 2025) and their default hyperparameters, except where noted. Unless stated otherwise every method uses AdamW at a learning rate of 1×10^{-5} decayed linearly to zero over its own step budget, weight decay 0.01, gradient-norm clipping at 1.0, micro-batch 2 with 8 accumulation steps (effective batch 16), and maximum sequence length 512. With 200 forget examples this gives 13 optimizer steps per epoch. Where a retain term is present, retain examples are drawn by walking $\mathcal{D}_r$ cyclically, one micro-batch per forget micro-batch, so that the retain gradient is not dominated by whichever examples happen to sit at the front of the set.

The objectives differ not only in functional form but also in how token losses are reduced. Let B be a micro-batch of examples $z = (q, a)$ with $|a_z|$ answer tokens, and write

$$\mathrm{nll}_\theta(z) = -\sum_{t=1}^{|a_z|} \log p_\theta(a_t \mid q, a_{<t}), \qquad \overline{\mathrm{nll}}_\theta(z) = \frac{\mathrm{nll}_\theta(z)}{|a_z|} \tag{41}$$

for the sequence sum and its length-normalized counterpart. We denote by

$$\mathcal{T}_\theta(B) = \frac{\sum_{z \in B} \mathrm{nll}_\theta(z)}{\sum_{z \in B} |a_z|} \tag{42}$$

the token-weighted batch mean returned by the causal language model. In general, $\mathcal{T}_\theta(B) \neq |B|^{-1} \sum_{z \in B} \overline{\mathrm{nll}}_\theta(z)$: the former gives longer answers greater weight. Prompt and padding positions are masked with `-100`, so only answer tokens contribute to any loss. These reductions are part of the implemented objectives, not tunable hyperparameters.

Gradient ascent. Plain ascent on the forget set, i.e. descent on the negated forget loss,

$$\mathcal{L}_{\mathrm{GA}}(\theta; B_f) = -\mathcal{T}_\theta(B_f). \tag{43}$$

Following upstream, which inherits from the plain fine-tuning recipe and exposes no retain weight, there is no retain term and no reference model. The objective is unbounded below and therefore diverges if run long; this is the method rather than a defect, and is precisely the failure mode the saturating objectives below are designed to avoid. We use 2 epochs (26 optimizer steps), matching the

ascent phase used elsewhere in this paper, where OpenUnlearning defaults to 10. Even at 2 epochs the model collapses in every run: the median final forget NLL is 9.61 per token against $\ln|\mathcal{V}| = 11.76$ for a uniform predictor, i.e. the model has become close to a random-token generator, and the damage is global rather than targeted (never-trained candidates score just as badly). This is why ρ_{LB} must be read beside the utility metrics: gradient ascent's near-chance attack accuracy reflects a destroyed model, not successful forgetting.

Gradient difference. Ascent on the forget set with a descent term on the retain set,

$$\mathcal{L}_{\mathrm{GD}}(\theta; B_f, B_r) = -\gamma\mathcal{T}_\theta(B_f) + \alpha\mathcal{T}_\theta(B_r), \tag{44}$$

at upstream's defaults $\gamma = \alpha = 1.0$, for 5 epochs (65 steps). This is the general gradient-difference construction audited in Section 7 at its always-interleaved, equal-weight configuration ($q = 0$, $\lambda = 1$), run for the full 5 epochs rather than an initial E_f epochs of a longer schedule. The retain term counteracts the forget-set ascent. The resulting updates are well behaved in practice: the final forget NLL has median 3.77 (range 2.74–5.33) and no run is flagged as diverged.

Negative preference optimization (NPO). NPO (Zhang et al., 2024b) replaces ascent with a DPO-style objective against a frozen reference π_{ref}, taken to be the trained model at the start of unlearning,

$$\mathcal{L}_{\mathrm{NPO}}(\theta; B_f, B_r) = \frac{2\gamma}{\beta|B_f|}\sum_{z\in B_f}\mathrm{softplus}(\beta\,[\log\pi_\theta(a \mid q) - \log\pi_{\mathrm{ref}}(a \mid q)]) + \alpha\mathcal{T}_\theta(B_r), \tag{45}$$

with $\beta = 0.1$, $\gamma = 1.0$, and a sequence-*summed* log-likelihood. Thus the nonlinear loss is formed per example and only then averaged over the micro-batch. The sigmoid saturates, so per-example gradients stay bounded and the model degrades far more gracefully than under plain ascent. We set the retain weight $\alpha = 1.0$, which corresponds to OpenUnlearning's `NPO_RT` variant and is the configuration that preserves usable utility ($\alpha = 0$ recovers pure NPO). We run 5 epochs (65 steps). NPO-RT does not collapse: its post-unlearning loss on the candidate pool is $\sim$2.9 per token.

SimNPO. SimNPO (Fan et al., 2024) is a reference-free variant in which the ratio against π_{ref} is replaced by the length-normalized NLL, so no frozen copy of the model is required,

$$\mathcal{L}_{\mathrm{SimNPO}}(\theta; B_f, B_r) = -\frac{2\gamma}{\beta|B_f|}\sum_{z\in B_f}\log\sigma\big(\beta\,[\overline{\mathrm{nll}}_\theta(z) - \delta]\big) + \alpha\mathcal{T}_\theta(B_r). \tag{46}$$

Consequently the forget terms of GA/GD, NPO, and SimNPO are not on a common scale, and their hyperparameters (γ, β) are not comparable across families. We use upstream's defaults, which differ sharply from NPO-RT's: $\beta = 4.5$, $\delta = 0$, $\gamma = 0.125$, $\alpha = 1.0$, for 5 epochs (65 steps). Dropping the reference model saves $\sim$2.3 GiB and roughly 15% of the unlearning time. At these defaults the objective is effectively saturated on our data and SimNPO leaves the model almost unchanged: the final forget NLL has median 1.43, against 1.29 for the checkpoint before unlearning.

Retain fine-tune. Fine-tuning the trained checkpoint on $\mathcal{D}_r$ alone for 2 epochs (450 optimizer steps, the forget set playing no part), equivalent to OpenUnlearning's `finetune` baseline and to the retain fine-tuning method audited elsewhere in this paper. Note that this is distinct from the retain-only *reference* model used for TOFU forget quality, which is trained from the base checkpoint and therefore never sees the forget data at all; `retain_ft` starts from a checkpoint that has already been exposed to it.

Implementation note. The loss reductions match OpenUnlearning; our pipeline instead walks the retain set with a persistent cyclic cursor rather than randomly sampling retain examples.

Table 12: Unlearning hyperparameters on TOFU. All methods use AdamW, learning rate 1×10^{-5} decayed linearly to zero, weight decay 0.01, gradient clipping 1.0, effective batch size 16, and maximum sequence length 512. "Forget reduction" specifies how answer-token NLLs are aggregated before the method's forget objective is applied; it is not an additional hyperparameter.

Method	Epochs	Steps	Retain term	Reference	Forget reduction	Method-specific
Gradient ascent	2	26	—	—	batch token mean	—
Gradient diff.	5	65	$\alpha = 1.0$	—	batch token mean	$\gamma = 1.0$
NPO-RT	5	65	$\alpha = 1.0$	frozen	seq. sum $\rightarrow$ ex. mean	$\beta = 0.1, \gamma = 1.0$
SimNPO	5	65	$\alpha = 1.0$	—	seq. mean $\rightarrow$ ex. mean	$\beta = 4.5, \delta = 0, \gamma = 0.125$
Retain fine-tune	2	450	sole objective	—	n/a	—

Accumulation. Micro-batch gradients are scaled by $1/A$ with $A = 8$ the accumulation count and summed until an optimizer step, giving an effective batch of 16 sequences. The optimizer also steps at the end of each epoch, so the final accumulation group may be short—with 200 forget examples at micro-batch 2 there are 100 micro-batches per epoch and $100 = 12 \times 8 + 4$—and its gradient is scaled by $1/A$ regardless of the group's actual size. Gradient-norm clipping at 1.0 is applied immediately before every step, so the fixed scaling does not change any update whose gradient is clipped, and AdamW's scale invariance damps whatever remains. It is in any case identical across the four forget-set methods and runs.

K Comparison with basic membership inference for CIFAR100

Unlearning algorithms are frequently evaluated using the accuracy of a basic membership-inference attack (MIA) to distinguish between the forget set and unseen test data, treating 50% accuracy as an indicator of successful unlearning. Because this baseline adversary cannot train multiple shadow models, it represents a significantly weaker threat model than our auditor. In this section, we evaluate our unlearning algorithms against this basic MIA framework using CIFAR-100.

Basic MIA attack. We use the basic membership-inference attack (MIA) of Kurmanji et al. (2023). The attack trains a logistic regression classifier on the audited model's per-example losses to distinguish training examples ("in") from unseen examples ("out"). Losses are clipped to $[-100, 100]$, and attack performance is measured using five-fold cross-validation, with the mean accuracy across the held-out folds reported. Following Kurmanji et al. (2023), we match the "in" and "out" populations in distribution; otherwise, the classifier could distinguish them based on distributional differences rather than membership. Specifically, in each run, the "in" population consists of the 2250 candidates sampled into the forget set, while the "out" population consists of the remaining 2250 candidates that were not sampled. Thus, the two populations have the same classes, data split, and preprocessing, and differ only in membership. Because they are equal in size, no reweighting is needed. Each reported figure aggregates all $\Gamma + L = 60$ runs in a sweep, which differ only in which half of the $m = 4500$ candidates is sampled.[8]

Comparison with our audit attack The two measurements assume very different adversaries. The basic attacker observes a single model and thresholds its per-example losses, with no shadow models and no calibration; our auditor observes $\Gamma = 50$ calibration runs of the whole train-and-unlearn pipeline, fits per-example in- and out-distributions from them, and evaluates a likelihood ratio over $L = 10$ further runs with $r = 3000$ guesses at confidence $1 - \zeta = 0.95$.

Different trained-model baselines. The pre-unlearning rows start from different base-training configurations. Gradient difference, gradient ascent, and retain fine-tuning use a one-cycle learning-rate schedule for 400 epochs; Hessian-based unlearning uses a cosine schedule for 400 epochs; and the class-based distillation methods use a cosine schedule for 200 epochs (Appendices I.1, J

[8] Pooling the calibration and evaluation runs is appropriate here because MIA accuracy is used only as a descriptive metric, not as a lower bound. By contrast, the independence between fitting and evaluation runs required in Section 5 is essential for the validity of the audit's lower bound; we make no analogous claim for attack accuracy.

Method	Unlearn LR	R/F/T (%)	Attack accuracy (%) mean	Δ	median	ρ_{LB}
Before unlearning		69.3/43.1/51.5	58.13 ±0.55	—	58.22	84.97
DELETE	1×10^{-4}	69.3/4.2/49.3	58.61 ±0.62	+0.47	58.66	111.55
	1×10^{-3}	69.3/1.6/49.1	58.53 ±0.68	+0.39	58.60	103.69
	3×10^{-3}	69.1/0.6/49.0	58.61 ±0.65	+0.47	58.61	89.44
BadTeacher	1×10^{-4}	68.4/30.4/50.3	58.67 ±0.52	+0.53	58.72	91.15
	1×10^{-3}	63.0/22.5/48.1	56.99 ±0.67	−1.14	56.94	46.40
	3×10^{-3}	44.6/7.7/38.1	51.95 ±0.87	−6.18	51.91	0.46
SCRUB	1×10^{-4}	69.5/13.7/49.9	57.80 ±0.59	−0.33	57.89	110.90
	5×10^{-4}	68.1/5.2/48.7	57.02 ±0.61	−1.12	57.07	82.26
	1×10^{-3}	48.8/2.0/36.8	53.99 ±2.57	−4.15	55.08	10.06
	3×10^{-3}	2.0/0.0/1.8	49.99 ±0.77	−8.14	50.09	0.0
SCRUB+R	1×10^{-4}	69.5/13.7/49.9	57.80 ±0.59	−0.33	57.89	110.90
	5×10^{-4}	67.9/6.4/48.7	57.24 ±0.57	−0.90	57.27	83.51
	1×10^{-3}	46.3/3.4/34.1	54.46 ±3.34	−3.67	56.26	11.39
	3×10^{-3}	1.3/0.0/1.2	50.02 ±0.77	−8.11	50.03	0.0

Table 13: Basic membership-inference attack accuracy against class-unlearned CIFAR-100 models, beside the audited zCDP lower bound on the same checkpoints. The s.d. is across runs; Δ is the change in the mean from the reference row, and R/F/T the checkpoint's mean retain / sampled-forget / test accuracy. The median is given because SCRUB and SCRUB+R at 1×10^{-3} are skewed, a minority of collapsed runs pulling their means below it. Lower bounds are reproduced from Table 17 for the unlearned rows; the pre-unlearning bound was measured on the trained checkpoints for this comparison and is not taken from that table.

and J.5). In the resulting trained checkpoints, retain and forget accuracy both exceed 90% for the gradient methods, while retain accuracy is about 69–72% for the cosine-trained models. Their forget accuracy also depends on the split: about 70% for the uniform Hessian case and 42–43% for the adversarial Hessian and class-based cases. These starting-point differences should not be attributed to the unlearning updates.

The metric often preserves direction but compresses magnitude. At 1×10^{-3}, both measurements rank the methods DELETE > BadTeacher > SCRUB+R > SCRUB. At 3×10^{-3}, their means have the same numerical ordering, but the lower bound ties SCRUB and SCRUB+R at zero and their MIA means differ by only 0.03 points. At 1×10^{-4} the rankings differ: BadTeacher has the highest MIA mean (58.67%), while DELETE has the highest lower bound (111.55). The more consistent pattern is compression in the scale of the differences. Across the table, ρ_{LB} ranges from 0 to 111.55, while attack accuracy varies only from 49.99% to 58.67%. Even among the nine configurations that preserve utility, the lower bound spans 46.40 to 111.55, whereas attack accuracy is confined to the narrow interval from 56.99% to 58.67%. This interval is only 1.68 percentage points wide and contains the 58.13% accuracy obtained against the model that was never unlearned. Thus, attack accuracy places configurations with substantially different privacy leakage on nearly the same scale: the 194-fold gap between BadTeacher and DELETE at 3×10^{-3} is compressed to 6.7 percentage points, and the 10-fold gap at 1×10^{-3} to just 4.5 percentage points.

The independently measured pre-unlearning bound also gives a directional check. Relative to its $\rho_{\mathrm{LB}} = 84.97$ and 58.13% MIA baseline, the bound and MIA mean move in the same direction for 12 of the 14 unlearned rows. At 1×10^{-4}, SCRUB and SCRUB+R are the exceptions: their bound rises to 110.90 while MIA falls by 0.33 points, less than the reported across-run s.d. The trained-model baseline therefore supports the broad trend, but not a strict row-by-row ordering.

Retain fine-tuning, gradient ascent, and gradient difference. Tables 14 and 15 extend the comparison to the gradient-based methods. The class-centric forget pool corresponds to the adversarial split (A) in Table 18; the uniform pool corresponds to U. Both ascent rows use $B = 1$. The attack accuracy uses the matched in/out forget populations defined above.

Method	R/F/T (%)	Attack accuracy (%) mean	Δ	median	ρ_{LB}
Before unlearning	97.8/91.7/52.1	77.73 ±0.50	—	77.74	370
Retain FT	93.9/2.8/51.3	56.41 ±0.74	−21.32	56.48	38.6
Grad. ascent	93.8/2.7/51.3	56.46 ±0.78	−21.27	56.50	38.56
GD ($q = 1$)	91.8/2.3/51.0	49.99 ±0.60	−27.74	50.09	0.0
GD ($q = 2$)	92.2/2.9/51.4	50.85 ±1.17	−26.88	51.13	2.02
GD ($q = 4$)	92.7/2.8/51.5	52.16 ±0.90	−25.58	52.44	11.63

Table 14: Basic MIA accuracy (%) for gradient-based methods on the class-centric (adversarial) CIFAR-100 split. Δ is the change in mean attack accuracy from each sweep's own pre-unlearning mean. The mean is shown with its across-run s.d.; the median is also across runs. Unlearned-model bounds are from Table 18; the trained-model bound was measured for this comparison.

Method	R/F/T (%)	Attack accuracy (%) mean	Δ	median	ρ_{LB}
Before unlearning	97.4/97.3/52.3	73.44 ±0.57	—	73.41	—
Retain FT	92.7/65.3/53.9	56.32 ±0.64	−17.15	56.36	37.5
Grad. ascent	92.6/65.3/53.9	56.32 ±0.68	−17.16	56.18	35.68
GD ($q = 1$)	91.5/65.5/54.1	56.37 ±0.67	−17.09	56.41	38.36
GD ($q = 2$)	91.5/66.1/54.0	56.58 ±0.64	−16.89	56.64	42.45
GD ($q = 4$)	91.5/66.3/54.1	56.64 ±0.67	−16.84	56.63	46.17

Table 15: Basic MIA accuracy (%) on the uniform CIFAR-100 split. Columns are as in Table 14. Unlearned-checkpoint bounds are from Table 18; no pre-unlearning bound is reported for this sweep.

For the class-centric split, the trained model's 370 bound and 77.73% mean attack accuracy both fall after every unlearning method. Across both tables, the largest absolute mean–median gap is 0.28 points, so none of these trends depends on a large mean–median gap. Attack accuracy and the audit also agree on the direction of the GD q sweep: both rise with q, most visibly under the adversarial split. Under uniform splitting, the GD bounds rise from 38.36 to 46.17 while MIA moves only 0.27 percentage points, less than the run-to-run s.d. in any of those MIA rows; under the adversarial split, the bounds rise from 0 to 11.63 while MIA moves 2.17 points. Thus, the GD results extend the compressed-scale pattern above without requiring a separate method-by-method interpretation.

Hessian-based unlearning. The Newton-style update of Zhang et al. (2025) adds Gaussian noise of scale σ to every parameter as its final step, which gives a sweep over which the audited bound moves far further than any hyperparameter of the distillation methods does. Table 16 reports the same attack across it, beside the bounds of Table 20 and Table 5.

Both measurements fall as noise is added, so the trained-model rows are also directionally consistent with the attack. Under uniform splitting, the independently measured pre-unlearning bound of 95.68 accompanies 59.21% MIA accuracy; at $\sigma = 10^{-2}$ these become 0.011 and 52.45%. Across the three unlearned noise settings alone, ρ_{LB} falls from 66.943 to 0.011, a factor of some six thousand, while MIA falls only 5.85 points. Under adversarial splitting, the trained-model bound and MIA mean fall from 93.84 and 59.30% to 2.8 and 56.57% at $\sigma = 10^{-3}$; the latter MIA mean has a large 2.86-point run-to-run s.d. At that noise scale the adversarial bound is far below the uniform 66.943, while MIA differs by less than two points (56.57% versus 58.30%). Thus the trained baselines support the same broad direction, although the scales differ and the adversarial estimate is noisy. At $\sigma = 10^{-2}$ the lower bound is near zero, while MIA still reads 52.45% against a chance line of 50%.

Split	σ	R/F/T (%)	Attack accuracy (%) mean	Δ	median	ρ_{LB}
Uniform	*before unlearning*	70.3/70.1/51.4	59.21 $\pm$0.62	—	59.27	95.68
	1×10^{-3}	69.5/67.0/50.6	58.30 $\pm$0.72	-0.92	58.50	66.943
	5×10^{-3}	53.2/51.6/41.9	55.69 $\pm$0.80	-3.53	55.61	12.276
	1×10^{-2}	21.5/21.1/18.7	52.45 $\pm$0.83	-6.76	52.47	0.011
Adversarial	*before unlearning*	72.4/41.7/50.9	59.30 $\pm$0.66	—	59.22	93.84
	1×10^{-3}	54.6/8.9/37.7	56.57 $\pm$2.86	-2.73	57.91	2.8

Table 16: Basic membership-inference attack accuracy against Hessian-based unlearning on CIFAR-100, as the final Gaussian noise scale σ is varied. Columns are as in Table 13, with Δ measured from each split's own pre-unlearning row. Unlearned-checkpoint bounds are from Table 20 (uniform) and Table 5 (adversarial, $\sigma = 10^{-3}$); the pre-unlearning bounds were measured on the trained checkpoints for this comparison, rather than taken from those tables. The adversarial $\sigma = 10^{-3}$ row has a far larger s.d. than the rest, so its median is the figure to read.

	DELETE			**BadTeacher**		
Unlearn LR	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	R/F/T (%)	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	R/F/T (%)
1×10^{-4}	111.55	164.56	69.3/4.2/49.3	91.15	138.91	68.4/30.4/50.3
1×10^{-3}	103.69	154.74	69.3/1.6/49.1	46.40	80.08	63.0/22.5/48.1
3×10^{-3}	89.44	136.73	69.1/0.6/49.0	0.46	3.36	44.6/7.7/38.1
	SCRUB			**SCRUB+R**		
Unlearn LR	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	R/F/T (%)	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	R/F/T (%)
1×10^{-4}	110.90	163.75	69.5/13.7/49.9	110.90	163.75	69.5/13.7/49.9
5×10^{-4}	82.26	127.56	68.1/5.2/48.7	83.51	129.16	67.9/6.4/48.7
1×10^{-3}	10.06	25.20	48.8/2.0/36.8	11.39	27.55	46.3/3.4/34.1
3×10^{-3}	0.0	0.0	2.0/0.0/1.8	0.0	0.0	1.3/0.0/1.2

Table 17: Mean-aggregation audit lower bounds for distillation-based (DELETE, BadTeacher) and interleaved ascent–descent (SCRUB, SCRUB+R) unlearning on CIFAR-100, together with retain-, forget-, and test-set accuracy. Each pair shares a single unlearning-rate column. Objectives and the fixed per-method configuration (unlearning epochs, checkpoint selection) are given in Appendix J.5. Forward conversions are reported only as interpretive scales, not confidence lower bounds. SCRUB+R rows are identical to the corresponding SCRUB row wherever the rewind rule keeps the final epoch.

L Hyperparameter sweeps for unlearning algorithms

This appendix studies how the audit responds to controlled changes in the strength of several uncertified unlearning procedures. Algorithm definitions and implementation details remain in Appendix I; here we focus on the observed audit trends.

L.1 Distillation-based unlearning algorithms

We audit the four distillation-based unlearners of Appendix J.5—BadTeacher (Chundawat et al., 2023), SCRUB and its rewound variant SCRUB+R (Kurmanji et al., 2023), and DELETE (Zhou et al., 2025)—as the unlearning learning rate is varied. Objectives, the rewind rule, and the fixed per-method configuration are given there; here we report only the sweep and what it shows.

Sweep protocol. The unlearning learning rate is the sole quantity varied, over $\{1 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}\}$, with every other setting held at the values of Appendix J.5. We use the batchwise likelihood-ratio auditor from Section 5 and report the mean-aggregation bound, together with retain-, forget-, and test-set accuracy at each configuration.

Monotonicity in the unlearning rate. For all four methods the audited bound decreases monotonically as the unlearning learning rate increases. This is the behaviour a sound audit should exhibit. The unlearning rate controls how far the student is driven from the teacher, so the sweep interpolates between two degenerate endpoints: at rate zero the student *is* the teacher and the forget set remains maximally distinguishable, while at a large enough rate the student retains nothing of the teacher and the audit must return zero—as it does for SCRUB and SCRUB+R at 3×10^{-3}.

Relative comparison. How successful our attack is against each method changes across the sweep. At the mildest rate 1×10^{-4} all four bounds lie in $[91, 112]$: too little unlearning has occurred for the attack to succeed differently against them. Differences appear only once unlearning is strong enough to matter. At 1×10^{-3} the attack is least successful against SCRUB (10.1) and SCRUB+R (11.4), followed by BadTeacher (46.4) and DELETE (103.7); at 3×10^{-3} it finds SCRUB = SCRUB+R (0.0), then BadTeacher (0.5), and remains most successful against DELETE (89.4). We believe this order reflects how each forget objective uses the trained model's logits, summarised in the last column of Table 11. BadTeacher's forget target is independent of $f_t(x)$, and SCRUB and SCRUB+R maximise divergence from $f_t(x)$ rather than match it; either could remove the per-example confidence patterns the attack exploits. For DELETE, among surviving classes $c, c' \notin \mathcal{C}_f$, Equation (40) gives $q(x)_c/q(x)_{c'} = \exp((f_t(x)_c - f_t(x)_{c'})/\tau)$: the target retains the teacher's relative odds exactly, and so could carry those patterns into the student.

The comparison with utility is what makes this more than a list of bounds. Every method against which the attack is unsuccessful pays for it in utility. At 3×10^{-3}, SCRUB and SCRUB+R reach $\rho_{\mathrm{LB}} = 0$ only once the model has collapsed, at $2.0/0.0/1.8$ and $1.3/0.0/1.2$; BadTeacher's 0.46 comes at $44.6/7.7/38.1$. DELETE alone reaches $69.1/0.6/49.0$—forget accuracy under 1% with retain and test untouched—and it is at the same time the method against which the attack is most successful, by two orders of magnitude. Low forget-set accuracy therefore certifies nothing on its own.

Appendix K runs the basic membership-inference metric of Kurmanji et al. (2023) over these same checkpoints. It recovers the ordering above, on a scale that separates the four by a few points of attack accuracy.

SCRUB versus SCRUB+R. The rewind step changes the audited bound negligibly: the two agree exactly at 1×10^{-4} and at 3×10^{-3}, and differ by less than 1.4 in ρ_{LB} at the intermediate rates. This is expected from the selection rule of Appendix J.5, which chooses the checkpoint whose forget-set *error* first matches the retrained reference. But a scalar error rate does not constrain the per-example confidence distribution on $\mathcal{D}_f$, and it is that distribution the auditor reads. An iterate chosen to look retrained by its error can remain fully distinguishable by its confidences. The rewind therefore addresses overshoot in a statistic the audit does not use.

L.2 Gradient difference

The GD update rule and dataset-specific configurations are given in Appendix J.2. For CIFAR-100, we use the representative configuration $q = 2$, $\lambda = 0.75$, and $E_f = 10$ in the main paper. We vary q in gradient difference and also compare it with retain fine-tuning and two epochs of gradient ascent in Table 18.

All methods yield large lower bounds under the uniform split. Under the adversarial split, the GD bound increases monotonically with q, and the representative $q = 2$ configuration yields a substantially smaller bound than either retain fine-tuning or gradient ascent.

Table 19 reports the corresponding Shakespeare sweeps over q, λ, and E_f. The main-paper comparison uses $q = 1$, $\lambda = 1.5$, and $E_f = 4$. On Shakespeare, the lower bound increases mildly with q $(4.802 \to 5.207 \to 5.245)$, again indicating that less frequent forget updates leave a stronger signal. Holding $q = 1$, increasing either λ or E_f reduces the lower bound, and increasing both drives it to 0.009. Thus the audit changes monotonically with the strength and duration of the forget updates under this fixed configuration.

Setting	Params	$\rho_{\mathrm{LB}}\,[\varepsilon^{(c)}(\rho_{\mathrm{LB}})]$ U	$\rho_{\mathrm{LB}}\,[\varepsilon^{(c)}(\rho_{\mathrm{LB}})]$ A	R/F/T accuracy U	R/F/T accuracy A
GD ($\lambda = 0.75$)	$q = 1$, $E_f = 10$	38.36 [68.9]	0.0 [0.0]	92/66/54	92/2.2/51
	$q = 2$, $E_f = 10$	42.45 [74.6]	2.02 [8.5]	92/66/54	92/2.9/51
	$q = 4$, $E_f = 10$	46.17 [79.8]	11.63 [28.0]	91/67/54	93/2.7/51
Other methods	Retain fine-tune	37.5 [67.7]	38.6 [69.2]	92.7/65.3/53.9	93.9/2.8/51.3
	Gradient ascent ($E_f = 2$)	35.68 [65.1]	38.56 [69.1]	93/65/54	94/2.7/51

Table 18: CIFAR-100 comparison of gradient-based unlearning methods. Entries are $\rho_{\mathrm{LB}}[\varepsilon^{(c)}(\rho_{\mathrm{LB}})]$, with the bracketed forward conversion shown only as an interpretive scale. U and A denote uniform and adversarial splits; R/F/T denotes retain-, forget-, and test-set accuracy.

Setting	Parameters	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$
Varying q	$q = 1$, $\lambda = 0.5$, $E_f = 4$	4.802	15.04
	$q = 4$, $\lambda = 0.5$, $E_f = 4$	5.207	15.89
	$q = 9$, $\lambda = 0.5$, $E_f = 4$	5.245	15.97
Varying (λ, E_f); $q = 1$	$\lambda = 0.5$, $E_f = 4$	4.802	15.04
	$\lambda = 0.5$, $E_f = 7$	3.532	12.23
	$\lambda = 1.5$, $E_f = 4$	3.444	12.02
	$\lambda = 1.5$, $E_f = 7$	0.009	0.33

Table 19: Shakespeare GD hyperparameter sweep. The converted $\varepsilon^{(c)}(\rho_{\mathrm{LB}})$ values are interpretive scales rather than confidence lower bounds on ε. Here E_f is the number of forget epochs within the eight unlearning epochs. The main-paper comparison uses $q = 1$, $\lambda = 1.5$, and $E_f = 4$. Across nearly every setting, retain and forget accuracy average $\sim$0.56 and test accuracy $\sim$0.53.

L.3 Hessian-based unlearning

The Hessian update, LiSSA approximation, and default hyperparameters are given in Appendix J.1. To isolate the effect of the final perturbation, we repeat the CIFAR-100 uniform-split audit while changing only the final Gaussian noise scale. Increasing the noise decreases the audit lower bound monotonically, but it also substantially reduces retain and test accuracy, exposing the associated privacy–utility tradeoff. A detailed comparison in presented in Table 20.

Limitations of output-based auditing. Our auditor is black-box and operates only on model outputs, so it may miss residual information encoded in internal representations. For example, the Prototypical Relearning Attack of Ha et al. (2025) constructs forgotten-class prototypes in representation space and shows that class structure can remain recoverable even when output-level forgetting metrics appear successful. That attack requires access to internal representations and therefore assumes a stronger access model than our auditor. We view the two approaches as complementary.

M Utility metrics and unlearning audits on TOFU

Metrics. Let $z = (q, a)$ be a QA pair, $\tilde{a}$ its paraphrased answer, and $\{\hat{a}^{(1)}, \ldots, \hat{a}^{(k)}\}$ its perturbed answers. Following Maini et al. (2024), we use the length-normalized answer probability and the truth ratio

$$R_{\mathrm{truth}}(z) = \frac{\frac{1}{k}\sum_{j=1}^{k} P_\theta(\hat{a}^{(j)} \mid q)^{1/|\hat{a}^{(j)}|}}{P_\theta(\tilde{a} \mid q)^{1/|\tilde{a}|}}. \tag{47}$$

TOFU's forget-quality score is the two-sample Kolmogorov–Smirnov (KS) p-value comparing the truth-ratio distribution of the unlearned model with that of a reference model fine-tuned from the base checkpoint on $\mathcal{D}_r$ alone. A small p rejects equality of these distributions but does not identify the direction: relative to the reference, a lower truth ratio is consistent with residual memorization,

Final noise σ	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	R/F/T
10^{-3} (original)	66.943	107.661	0.69/0.67/0.51
5×10^{-3}	12.276	29.086	0.53/0.51/0.42
10^{-2}	0.011	0.384	0.22/0.21/0.19

Table 20: Hessian-based zCDP audit under a sweep of the final noise scale on the uniform split. The forward conversion is shown only as an interpretive scale; R/F/T denotes retain-, forget-, and test-set accuracy.

whereas a higher one is consistent with over-forgetting. A large p is only a non-rejection, not evidence of equality.

We also report an adapted utility score: the harmonic mean of $\{P_\theta, \max(0, 1 - R_{\mathrm{truth}})\}$ over the retain, real-authors, and world-facts splits. Unlike TOFU's Model Utility, it omits ROUGE and drops zero terms; it is therefore comparable only within Table 21.

Relation to certified unlearning. The KS test cannot certify Definition 3.1: it compares a scalar post-processing against one retrained reference, and a non-rejection does not establish Rényi closeness. Conversely, a rejection rules out only that reference, whereas the definition permits any retain-only reference. Our audit instead uses identifiability of the hidden sign vector to bound ρ for every such reference (Lemma 4.2).

Evaluation convention. Superscripts $+$ and $-$ denote the selected ($S_j = +1$) and unselected ($S_j = -1$) halves of the 400 candidate rows. Only the $+$ rows were trained on and subsequently unlearned. The $-$ rows are an off-target diagnostic: a difference there cannot be caused by unlearning those rows, although the unlearned and reference models have different training histories. We report the median p^+ and mean p^- across $L = 10$ runs as descriptive summaries, not combined hypothesis tests. For comparability across batchings, Table 21 fixes $r/m = 0.5$; the headline experiment instead uses $B = 1$, $m = 400$, and $r/m = 0.25$.

Results. NPO-RT has the largest median p^+ (0.47–0.67), so the KS test does not reject equality with the retrained reference in any batching. Nevertheless, the audit recovers 83–94% of its reported membership guesses and establishes $\rho_{\mathrm{LB}} = 0.605$–$9.808$. Passing the conventional KS threshold therefore does not imply removal of the auditable membership signal.

Across these five methods, ranking by mean p^- reproduces the ordering of the reported ρ_{LB} values in all four batchings: GA yields no positive bound in three settings and the smallest bound at $B = 1$, followed by GD, NPO-RT, and the censored RetainFT/SimNPO tie. This is a descriptive regularity over five methods, not a fitted relationship. The p^+ ranking does not agree: at $B = 1$, for example, RetainFT has $p^+ = 0.026$ and $\rho_{\mathrm{LB}} = 31.428$, whereas NPO-RT has $p^+ = 0.466$ and $\rho_{\mathrm{LB}} = 9.808$. Together with the truth-ratio, utility, and loss diagnostics, this is consistent with smaller bounds arising alongside greater off-target model change rather than with better forget-set behavior alone.

Gradient ascent is the limiting case. At $B = 1$ its overlap is 106.7 against the random-guess baseline 100, giving the table's smallest bound ($\rho_{\mathrm{LB}} = 0.004$). That bound alone does not establish unlearning, and the definition itself imposes no utility requirement. The behavioral evidence instead shows severe collapse: $R^+_{\mathrm{truth}} = 1.78$–$1.96$ and median forget NLL 9.61 per token against 11.76 for a uniform predictor (Appendix J.6). Gradient difference exhibits less severe off-target change and retains nonzero, but smallest non-degenerate, bounds in every batching.

Finally, RetainFT and SimNPO attain $\bar{V} = r$ in every evaluation run and batching: every *reported* guess is correct. Their bounds therefore reach the largest value measurable at each r and are censored; the audit cannot order the two methods beyond this ceiling. Their behavioral diagnostics distinguish them but do not give a common ordering: RetainFT has higher adapted utility, while SimNPO has higher p^+ in every batching.

Table 21: Audit and behavioral metrics for five TOFU unlearning methods on Llama-3.2-1B-Instruct, using $r/m = 0.5$ and $L = 10$ evaluation runs. $\bar{V}$ is mean overlap (random-guess baseline $r/2$), and ρ_{LB} is the zCDP lower bound at $\zeta = 0.05$. Superscripts $+$ and $-$ identify the selected and unselected (off-target) audit slices. The adapted utility score is defined above and is comparable only within this table.

	Audit		Behavioral diagnostics			
Method	$\bar{V}$	ρ_{LB}	med. p^+	R^+_{truth}	mean p^-	Adapt. util.
$B = 20$, $m = 20$, $r = 10$						
Before Unlearning	10.0	1.021	0.012	0.796	1.00	0.340
Retain FT	10.0	1.021	5.2e−3	0.788	0.99	0.381
SimNPO	10.0	1.021	0.016	0.802	1.00	0.355
NPO-RT	9.4	0.605	0.671	0.982	0.63	0.328
Grad. diff.	9.0	0.436	0.028	1.244	0.16	0.361
Grad. ascent	5.3	0.0	3.0e−7	1.780	0.00	0.302
$B = 10$, $m = 40$, $r = 20$						
Before Unlearning	20.0	2.499	0.026	0.801	0.95	0.339
Retain FT	20.0	2.449	0.017	0.793	0.86	0.380
SimNPO	20.0	2.449	0.052	0.808	0.91	0.354
NPO-RT	18.7	1.467	0.587	0.969	0.48	0.343
Grad. diff.	16.8	0.701	0.010	1.207	0.09	0.378
Grad. ascent	10.0	0.0	4.4e−7	1.952	0.00	0.297
$B = 4$, $m = 100$, $r = 50$						
Before Unlearning	50	7.050	0.082	0.823	0.87	0.339
Retain FT	50.0	7.050	0.068	0.816	0.73	0.381
SimNPO	50.0	7.050	0.142	0.829	0.89	0.354
NPO-RT	44.4	3.185	0.628	0.964	0.62	0.349
Grad. diff.	43.5	2.803	0.010	1.238	0.06	0.378
Grad. ascent	25.7	0.0	3.9e−9	1.930	0.00	0.277
$B = 1$, $m = 400$, $r = 200$						
Before Unlearning	200.0	31.428	0.052	0.816	0.76	0.339
Retain FT	200.0	31.428	0.026	0.810	0.61	0.380
SimNPO	200.0	31.428	0.112	0.822	0.85	0.354
NPO-RT	165.4	9.808	0.466	0.983	0.53	0.346
Grad. diff.	157.5	7.226	8.1e−3	1.357	0.04	0.372
Grad. ascent	106.7	0.004	5.3e−7	1.961	0.00	0.312

N Convex unlearning: auditor and hyper-parameter details

Definition N.1 ((ε, δ)-certified convex unlearning)**.** Given a dataset $\mathcal{D}$ of size n_{pt} and a forget set $\mathcal{D}_f \subseteq \mathcal{D}$ of size m_{pt}, let $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{D}_f$, so that $|\mathcal{D}_r| = n_{\mathrm{ret}} := n_{\mathrm{pt}} - m_{\mathrm{pt}}$. An unlearning algorithm $\mathcal{U}$ is an (ε, δ)-certified convex unlearning algorithm for $\mathcal{A}$ if $\mathcal{U}(\mathcal{D}_f, \mathcal{A}(\mathcal{D}), S(\mathcal{D}))$ and $\mathcal{U}(\emptyset, \mathcal{A}(\mathcal{D}_r), S(\mathcal{D}_r))$ are (ε, δ)-indistinguishable.

Definition N.2 (ρ-zCDP-certified convex unlearning)**.** Given $\mathcal{D}, \mathcal{D}_f$, and $\mathcal{D}_r$ as above, define

$$P_f := \mathrm{Law}(\mathcal{U}(\mathcal{D}_f, \mathcal{A}(\mathcal{D}), S(\mathcal{D}))),$$
$$P_r := \mathrm{Law}(\mathcal{U}(\emptyset, \mathcal{A}(\mathcal{D}_r), S(\mathcal{D}_r))).$$

We say that $\mathcal{U}$ is ρ-zCDP-certified convex unlearning for $\mathcal{A}$ if, for every $\gamma > 1$,

$$D_\gamma(P_f \| P_r) \le \rho\gamma \qquad \text{and} \qquad D_\gamma(P_r \| P_f) \le \rho\gamma.$$

This appendix gives the auditor pipeline under both privacy notions, the Perturbed Newton unlearning algorithm of Sekhari et al. (2021, Algorithm 1), the cubic and logistic loss constructions, and the hyper-parameter sweep used in Section 7.

N.1 AUDITOR PIPELINE

We adapt the pairwise auditor of Section 5 (Nasr et al., 2021-style) to the convex setting, with two changes.[9] First, in line with Definitions N.1 and N.2, the auditor distinguishes the unlearned model produced from a non-empty forget set, $\mathcal{U}(\mathcal{D}_f, \mathcal{A}(\mathcal{D}), S(\mathcal{D}))$, against the model produced from an empty forget set, $\mathcal{U}(\emptyset, \mathcal{A}(\mathcal{D}_r), S(\mathcal{D}_r))$, rather than two unlearned models from different non-empty forget sets. Second, since the unlearned models live in a low-dimensional Euclidean parameter space, we instantiate the pairwise auditor's predictor directly on the model weights rather than on per-example logit scores: we train $\Gamma = 50$ independent models under each setting and fit Gaussian distributions $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_f, \Sigma_f)$ to the resulting weight vectors.

Evaluation. The Γ fitting runs are used only to construct the predictor. At evaluation we perform $2L$ fresh independent executions—L of $\mathcal{U}(\mathcal{D}_f, \mathcal{A}(\mathcal{D}), S(\mathcal{D}))$ and L of $\mathcal{U}(\emptyset, \mathcal{A}(\mathcal{D}_r), S(\mathcal{D}_r))$—and assign each resulting weight vector W to whichever fitted Gaussian it is closer to, i.e.

$$g(W) = \mathbf{1}[\log \mathcal{N}(W; \mu_f, \Sigma_f) > \log \mathcal{N}(W; \mu_r, \Sigma_r)]\,,$$

predicting the non-empty-forget hypothesis when $g(W) = 1$. Because the evaluation runs are independent of the fitting runs, the resulting TPR and FPR estimates are binomial conditional on the fitted predictor, which is what the Clopper–Pearson intervals of Appendix D.4 require. Neither privacy lower bound assumes that P_f or P_r is Gaussian; the fits enter only through g, and any measurable g would yield a valid bound.

For the pure-DP audit, we set $\delta = 0$ in Equation (21) and plug in the resulting false-positive and false-negative rates to obtain $\varepsilon_{\mathrm{LB}}^{(p)}$. As explained in Appendix D.4, this is a one-sided $(1-\zeta)$ lower confidence bound on the smallest valid pure-DP parameter ε.

Direct zCDP pairwise bound. The zCDP analysis simplifies in this setting. The general unlearning audit compares two unlearned outputs through a common reference and therefore incurs the weak-triangle bound in Lemma 4.1. Here, by contrast, Definition N.2 directly compares the two distributions observed by the auditor. Thus

$$D_\gamma(P_f \| P_r),\ D_\gamma(P_r \| P_f) \le \varepsilon_\gamma^{\mathrm{conv}}(\rho) := \rho\gamma, \qquad \gamma > 1, \tag{48}$$

with no transitivity loss.

Let A be the event that the pairwise predictor outputs the non-empty-forget hypothesis, so that

$$P_f(A) = \mathrm{TPR}, \qquad P_r(A) = \mathrm{FPR}.$$

Applying the Rényi change-of-measure inequality to Equation (48) gives, for every $\gamma > 1$,

$$\mathrm{TPR} \le \exp\left(\frac{\gamma-1}{\gamma}\rho\gamma\right) \mathrm{FPR}^{(\gamma-1)/\gamma} = e^{(\gamma-1)\rho}\mathrm{FPR}^{(\gamma-1)/\gamma}.$$

Rearranging yields the population lower bound

$$\rho \ge \frac{\log \mathrm{TPR}}{\gamma-1} - \frac{\log \mathrm{FPR}}{\gamma}. \tag{49}$$

Reversing P_f, P_r and applying the same argument to A^c gives the analogous TNR/FNR expression. Let $\mathrm{TPR}^{\mathrm{low}}, \mathrm{TNR}^{\mathrm{low}}$ and $\mathrm{FPR}^{\mathrm{high}}, \mathrm{FNR}^{\mathrm{high}}$ be the simultaneous one-sided Clopper–Pearson endpoints defined in Appendix D.4. The direct convex pairwise auditor reports

$$\boxed{\rho_{\mathrm{LB,conv}}^{(p)} := \sup_{\gamma>1} \max\left\{0, \frac{\log \mathrm{TPR}^{\mathrm{low}}}{\gamma-1} - \frac{\log \mathrm{FPR}^{\mathrm{high}}}{\gamma}, \frac{\log \mathrm{TNR}^{\mathrm{low}}}{\gamma-1} - \frac{\log \mathrm{FNR}^{\mathrm{high}}}{\gamma}\right\}.} \tag{50}$$

[9] *Notation.* In the tables and running text of this appendix and of Section 8, the unsubscripted symbols $\rho_{\mathrm{LB}}^{(p)}$, ρ_{LB}, $\varepsilon_{\mathrm{LB}}^{(p)}$ and $\varepsilon_{\mathrm{LB}}$ always abbreviate the *direct* convex bounds, written $\rho_{\mathrm{LB,conv}}^{(p)}$ and $\rho_{\mathrm{LB,conv}}$ where they are derived below. These use the curve $\varepsilon_\gamma^{\mathrm{loc}} = \rho\gamma$ of Definition N.2 and so carry no transitivity penalty; they are not the weak-triangle quantities of Lemma 4.1 reported in Section 7. The superscript (p) marks the pairwise auditor; unmarked symbols are the $m = 2$, $r = 2$ Instantiation II auditor.

A term with a zero lower confidence bound contributes no positive lower bound. On the simultaneous Clopper–Pearson coverage event, the substitutions in Equation (50) can only decrease the two population expressions. The same event holds for every γ, so the supremum requires no further multiple-testing correction. Hence $\rho^{(p)}_{\mathrm{LB,conv}}$ is a one-sided $(1-\zeta)$ lower confidence bound on the smallest valid direct zCDP parameter.

Direct zCDP overlap bound. The pairwise Clopper–Pearson bound above is expressed directly in terms of the event A and does not require the mechanism $\mathcal{M}$ from the general overlap audit. We use that mechanism only for the second, joint-sign-vector bound reported in Table 22. To make this connection explicit, let

$$s_f := (+1,-1), \qquad s_r := (-1,+1),$$

encode the non-empty- and empty-forget hypotheses, respectively. Conditional on the independently fitted classifier, draw $W_{s_f} \sim P_f$ or $W_{s_r} \sim P_r$, apply the binary model-weight predictor g described above, and encode its output using the same two sign vectors. This defines the post-processed prediction mechanism

$$\mathcal{M}_{\mathrm{conv}}(s) := \begin{cases} s_f, & g(W_s) \text{ predicts the non-empty-forget hypothesis,} \\ s_r, & g(W_s) \text{ predicts the empty-forget hypothesis.} \end{cases}$$

This is exactly the $m = 2, r = 2$ specialization of Algorithm 1: its overlap score is 2 when the binary prediction is correct and 0 otherwise. For the reported pure-DP bound, we use $\delta = 0$; because Definition N.1 directly compares P_f and P_r, the overlap-score audit applies without the factor-of-two reference reduction from Lemma C.2. Under Definition N.2, data processing gives, in both directions,

$$D_\gamma(\mathcal{M}_{\mathrm{conv}}(s_f) \| \mathcal{M}_{\mathrm{conv}}(s_r)) \le D_\gamma(P_f \| P_r) \le \rho\gamma,$$

and analogously after reversing s_f, s_r. Consequently, this separately reported convex overlap audit is obtained from Lemma 4.2 by substituting $\varepsilon^{\mathrm{loc}}_\gamma = \rho\gamma$, rather than the weak-triangle expression from Lemma 4.1. Inverting the resulting monotone tail bound gives $\rho_{\mathrm{LB,conv}}$ without a transitivity penalty.

Alongside these zCDP lower bounds, Table 22 reports direct pure-DP $(\varepsilon, 0)$ lower bounds computed from the same evaluation samples. Rows are indexed by the nominal $(\varepsilon, 10^{-3})$ guarantee used to calibrate the Gaussian noise. For this mechanism, Sekhari et al. (2021, Lemma 3) gives the pre-noise center-separation bound

$$\Delta_{\mathrm{PN}} := \frac{2M m_{\mathrm{pt}}^2 G^2}{\mu^3 n_{\mathrm{pt}}^2} = 0.0649, \qquad \rho = \frac{\Delta_{\mathrm{PN}}^2}{2\sigma^2}.$$

The values $\varepsilon^{(c)}(\rho_{\mathrm{LB}})$ use the fixed conversion slack 10^{-3} and are included only for interpretation; they are not lower confidence bounds on ε. All reported lower bounds have simultaneous confidence level $1 - \zeta = 0.95$.

Why MSE is excluded. Mean squared error is somewhat degenerate in this context because its Hessian is constant. Consequently, in Algorithm 1 of Sekhari et al. (2021), there is zero noise addition at the final step. Also, the Newton-corrected iterate coincides with the model obtained by retraining on the retain set as the pre-noise center separation bound $\Delta_{\mathrm{PN}} = 0$.

N.2 Cubic loss in one dimension

We instantiate the cubic loss in dimension $d = 1$, so $w, z \in \mathbb{R}$. The pointwise loss is

$$\ell(w; z) = \tfrac{\lambda_0}{2} w^2 + \tfrac{M}{6} w^3 - zw, \tag{51}$$

constrained to the box $|w| \le B$, on which the loss is μ-strongly convex with $\mu = \lambda_0 - MB$. We place the n_{ret} retain points at $z = 0$ and the m_{pt} forget points at $z = -R$, where $R > 0$ is a data-radius parameter. Under this construction, the retain-only ERM has closed-form solution $\hat{w}_r = 0$, while the forget points exert a constant gradient pull of magnitude R, pushing the unlearned weight $\hat{w}_f$ as far from $\hat{w}_r$ as the box constraint allows. In our experiments we set $\lambda_0 = 0.2$, $M = 0.001$, $B = 2.5$, and $R = 10.0$, with a full dataset of $n_{\mathrm{pt}} = 1050$ points of which $m_{\mathrm{pt}} = 50$ form the forget set, leaving $n_{\mathrm{ret}} = 1000$ retain points. The full hyper-parameter sweep over (λ_0, M, B, R) is given below.

Table 22: Direct and zCDP audit results for Perturbed Newton unlearning on the convex cubic loss. The analytic ρ column uses the Lemma 3 bound $\Delta_{\mathrm{PN}} = 0.0649$, while the $\varepsilon_{\mathrm{LB}}$ columns use $\delta = 0$. The bounds with superscript (p) come from the pairwise Clopper–Pearson auditor; the unmarked bounds come from our $m = 2, r = 2$ Instantiation II auditor. All four are the direct convex bounds; see Appendix N.1.

ε	ρ	$\rho_{\mathrm{LB}}^{(p)}$	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}}^{(p)})$	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	$\varepsilon_{\mathrm{LB}}^{(p)}$	$\varepsilon_{\mathrm{LB}}$
0.1	0.0017	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.5	0.0235	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0754	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.2394	0.0050	0.0015	0.2354	0.1156	0.1317	0.0036
5	1.051	0.0225	0.0077	0.5687	0.3045	0.2540	0.0169
10	3.032	0.0666	0.0228	1.069	0.5726	0.4329	0.0485
20	8.214	0.1831	0.0613	1.936	1.018	0.7202	0.1306
50	27.79	0.6147	0.1870	4.012	1.961	1.368	0.4257
100	65.51	1.416	0.3640	6.745	2.919	2.205	0.9591
200	147.8	2.973	0.5612	10.90	3.795	3.573	2.007
400	322.5	5.778	0.6711	17.07	4.234	6.023	4.146
600	502.9	8.262	0.6852	21.91	4.288	8.356	6.482
800	685.6	10.16	0.6867	25.38	4.294	10.21	9.421

N.3 Hyper-parameter selection

Let $\ell(w; z)$ denote the pointwise loss at sample z, and let $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ be the full training set, where $\mathcal{D}_r$ and $\mathcal{D}_f$ are the retain and forget sets respectively.

Let

$$\hat{w}_r \;=\; \arg\min_{w} \sum_{z \in \mathcal{D}_r} \ell(w; z)$$

denote the (noiseless) retain-only ERM, and let $\hat{w}_f$ denote the *unlearned* weight, i.e. the weight obtained by first running ERM on the full dataset $\mathcal{D}$ to produce a trained model and *then* applying the Hessian-based Newton unlearning steps of Sekhari et al. (2021, Eq. 8) on the forget set $\mathcal{D}_f$ (with the Hessian evaluated on the retain set). In particular, $\hat{w}_f$ is *not* the model produced by training alone; it is the post-unlearning weight before addition of noise. Sekhari et al. (2021, Lemma 3) bounds

$$\left\|\hat{w}_r - \hat{w}_f\right\| \;\le\; \frac{2M m_{\mathrm{pt}}^2 G^2}{\mu^3 n_{\mathrm{pt}}^2},$$

when $\ell(\cdot)$ is μ-strongly convex, G-Lipschitz, and M-Hessian-Lipschitz in w. Here $n_{\mathrm{pt}} = |\mathcal{D}|$ is the size of the full pre-deletion dataset and $m_{\mathrm{pt}} = |\mathcal{D}_f|$ the number of deleted points, matching the convention of Sekhari et al. (2021); we write G for the Lipschitz constant, reserving L for the number of auditor runs. Our objective is to choose hyper-parameters that make the empirical $\|\hat{w}_r - \hat{w}_f\|$ as close to this theoretical bound as possible, i.e. that drive the ratio

$$R_{\mathrm{tight}} \;:=\; \frac{\|\hat{w}_r - \hat{w}_f\|_{\text{empirical}}}{2M m_{\mathrm{pt}}^2 G^2/(\mu^3 n_{\mathrm{pt}}^2)}$$

toward 1, since a value close to 1 indicates that Lemma 3 is empirically tight on the chosen problem instance. While the cubic loss typically yields values of R_{tight} around 0.22, the logistic loss yields values of R_{tight} around 3×10^{-5}, which possibly explains why the lower bound is so low.

N.4 Placement of retain and forget points

N.4.1 Cubic loss.

We instantiate the cubic loss in dimension $d = 1$, so $w, z \in \mathbb{R}$ throughout this subsection. We construct $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ as follows. The retain set $\mathcal{D}_r$ consists of n_{ret} points placed at the origin,

$$\mathcal{D}_r = \big\{(z_i, y_i)\big\}_{i=1}^{n_{\mathrm{ret}}}, \qquad z_i = 0, \quad y_i = 0,$$

and the forget set $\mathcal{D}_f$ consists of m_{pt} points placed at $-R$,

$$\mathcal{D}_f = \{(z_j, y_j)\}_{j=1}^{m_{\mathrm{pt}}}, \qquad z_j = -R, \quad y_j = 0,$$

where $R > 0$ is the data-radius parameter. All labels are zero because the cubic loss

$$f(w, z) \;=\; \tfrac{\lambda_0}{2}\, w^2 \;+\; \tfrac{M}{6}\, w^3 \;-\; z\, w$$

does not depend on labels. Under this construction, the retain-only ERM admits the closed-form solution $\hat{w}_r = 0$, since the retain gradient $\lambda_0 w + \frac{M}{2} w^2$ vanishes at $w = 0$.

The retain points are placed at the origin so that $\hat{w}_r = 0$ provides a clean reference. The forget points are placed at $-R$ so as to maximise their gradient contribution to the full-data ERM: from

$$\left.\frac{\partial f(w, z)}{\partial w}\right|_{z=-R} \;=\; \lambda_0\, w + \tfrac{M}{2}\, w^2 + R,$$

the forget set exerts a constant pull of magnitude R on w toward $-B$, pushing the full-data optimum as far from $\hat{w}_r = 0$ as the box constraint $|w| \le B$ allows, and producing the tightest empirical Lemma 3 ratio.

For the cubic loss, we sweep $M \in \{0.001, 0.002, 0.005\}$, $B \in \{2.5, 3.0, 3.5\}$, $R \in \{4, 5, 6, 7, 8, 10, 12\}$, and $\lambda_0 \in \{0.10, 0.12, 0.15, 0.18, 0.20\}$, and obtain the best (largest) ratio of $R_{\mathrm{tight}} = 0.22$ at $M = 0.001$, $B = 2.5$, $\lambda_0 = 0.2$ and $R = 10.0$. Because for the cubic loss (instantiated below in dimension $d = 1$) we assume $|w|$ is bounded by B, which yields the strong-convexity parameter $\mu = \lambda_0 - MB$, we additionally require the optimum after loss minimisation to lie inside the interval $[-B, B]$; otherwise the first-order optimality condition $\nabla \ell = 0$ may fail to hold.

N.4.2 Logistic loss (boundary construction).

The retain set consists of n_{ret} points with i.i.d. Gaussian features and labels generated from a fixed direction θ^*,

$$\mathcal{D}_r = \{(x_i, y_i)\}_{i=1}^{n_{\mathrm{ret}}}, \qquad x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0,\ \tfrac{1}{d} I_d\big)\,, \quad y_i \mid x_i \ \sim\ \mathrm{Bernoulli}\big(f_1(x_i^\top \theta^*)\big),$$

where $f_1(z) = \eta \cdot \mathbf{1}[z \ge 0] + (1 - \eta) \cdot \mathbf{1}[z < 0]$ is the label-flip function with flip probability $1 - \eta$. The forget set places all m_{pt} points at the same fixed boundary location with a fixed label,

$$\mathcal{D}_f = \{(x^*, y^*)\}_{j=1}^{m_{\mathrm{pt}}}, \qquad x^* = -\operatorname{sign}(\theta^*) \cdot \rho_f, \quad y^* = 1,$$

where $\rho_f > 0$ is the forget-point norm.

This construction is chosen for two reasons. First, concentrating all forget points at a single location x^* makes every forget sample exert an identical, unidirectional gradient contribution, maximising the aggregate influence of $\mathcal{D}_f$ on the full-data optimum $\hat{w}_f$ relative to the retain-only optimum $\hat{w}_r$. Second, placing x^* in the direction opposite to θ^* at distance ρ_f from the origin ensures that the forget signal directly opposes the retain signal, creating the largest possible shift between $\hat{w}_f$ and $\hat{w}_r$, and hence the tightest empirical Lemma 3 ratio R_{tight}. However, the sweep for the logistic loss yielded a maximum R_{tight} of only 3×10^{-5}, which explains the poor audit lower bounds, at parameters $d = 2$, $\eta = 0.8$, $\kappa = 0.7$, and $\rho_f = 5.0$.

Remark N.3 (Why we report $m = 2$ candidate forget batches). The convex audits above use the $m = 2$, $r = 2$ specialization of Algorithm 1. We also ran the audit with $m > 2$, taking $d \ge m$ and placing the forget batches along distinct coordinates so that each batch exerts an independent gradient pull on the optimum. This did not yield larger lower bounds, so we report the $m = 2$ case throughout.

O Output perturbation: mechanism and point placement

O.1 Mechanism

Given a trained model $\hat{x}$, the output-perturbation mechanism produces the unlearned model

$$x_u \;=\; \Pi_{C_0}(\hat{x}) \;+\; \xi_0, \qquad \xi_0 \sim \mathcal{N}(0,\, \sigma^2 \mathbb{I}),$$

Table 23: Direct and zCDP audit results for output perturbation. The analytic ρ column uses the projection-diameter bound $\Delta_{\mathrm{OP}} = 2C_0 = 0.200$, while the $\varepsilon_{\mathrm{LB}}$ columns use $\delta = 0$. The bounds with superscript (p) come from the pairwise Clopper–Pearson auditor; the unmarked bounds come from our $m = 2, r = 2$ Instantiation II auditor. All four are the direct convex bounds; see Appendix N.1.

ε	ρ	$\rho_{\mathrm{LB}}^{(p)}$	ρ_{LB}	$\varepsilon^{(c)}(\rho_{\mathrm{LB}}^{(p)})$	$\varepsilon^{(c)}(\rho_{\mathrm{LB}})$	$\varepsilon_{\mathrm{LB}}^{(p)}$	$\varepsilon_{\mathrm{LB}}$
0.1	0.0017	0.0000	0.0000	0.0048	0.0027	0.0042	0.0001
0.5	0.0235	0.0061	0.0021	0.2659	0.1414	0.1330	0.0049
1	0.0754	0.0272	0.0092	0.6346	0.3383	0.2816	0.0201
2	0.2394	0.1000	0.0334	1.356	0.7152	0.5419	0.0709
5	1.051	0.4659	0.1442	3.388	1.681	1.198	0.3196
10	3.032	1.310	0.3421	6.422	2.812	2.121	0.8815
20	8.214	3.254	0.5822	11.58	3.881	3.815	2.201
50	27.79	8.850	0.6857	23.00	4.290	8.943	6.828
100	65.51	10.17	0.6867	25.39	4.294	10.21	11.11
200	147.8	10.17	0.6867	25.39	4.294	10.21	11.11
400	322.5	10.17	0.6867	25.39	4.294	10.21	11.11
600	502.9	10.17	0.6867	25.39	4.294	10.21	11.11
800	685.6	10.17	0.6867	25.39	4.294	10.21	11.11

where Π_{C_0} denotes Euclidean projection onto an ℓ_2-ball of radius $C_0 = 0.100$. Any two projected centers therefore differ by at most

$$\Delta_{\mathrm{OP}} := 2C_0 = 0.200.$$

We calibrate σ so that adding Gaussian noise to centers separated by at most Δ_{OP} satisfies the nominal $(\varepsilon, 10^{-3})$ guarantee. Unlike the mechanism-specific Perturbed Newton center-separation bound Δ_{PN}, this bound follows directly from the projection geometry and does not invoke Lemma 3 of Sekhari et al. (2021). The auditing pipeline is identical to that of the cubic loss in Appendix N.1 (Gaussian fits on model parameters, Clopper–Pearson lower bound), with an MSE loss and no regulariser.

O.2 Audit results

Table 23 reports both the direct pure-DP $(\varepsilon, 0)$ lower bounds and the zCDP lower bounds obtained from the same evaluation samples. Rows are indexed by the nominal $(\varepsilon, 10^{-3})$ guarantee used to calibrate the Gaussian noise; this is a row label rather than an audited value. The second column is the analytic zCDP upper bound $\rho = \Delta_{\mathrm{OP}}^2/(2\sigma^2)$. Both auditors use the direct convex Rényi curve $\varepsilon_\gamma^{\mathrm{conv}}(\rho) = \rho\gamma$ and therefore incur no transitivity loss. The values $\varepsilon^{(c)}(\rho_{\mathrm{LB}})$ use conversion slack 10^{-3} and are included only for interpretation; they are not lower confidence bounds on ε. All lower bounds have simultaneous confidence level $1 - \zeta = 0.95$. Entries are displayed to four significant figures, but conversions use the unrounded ρ lower bounds; consequently, a displayed 0.0000 can have a small positive conversion.

O.3 Placement of retain and forget points

The retain set consists of n_{ret} points with i.i.d. Gaussian features and continuous labels generated from a fixed direction θ^*,

$$\mathcal{D}_r = \{(x_i, y_i)\}_{i=1}^{n_{\mathrm{ret}}}, \qquad x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0,\ \tfrac{1}{d} I_d\big), \quad y_i \mid x_i = f_{\mathrm{lin}}(x_i^\top \theta^*) - \tfrac{1}{2} + \varepsilon_i,$$

where

$$f_{\mathrm{lin}}(z) \ = \ \mathrm{clip}\big(\tfrac{1}{2} + \eta z,\ \tfrac{1}{2} - r,\ \tfrac{1}{2} + r\big)$$

is a clipped linear link with slope η and half-range r, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is observation noise (clipped so that $y_i \in [-r, r]$) with $\sigma = 0.05$.

The forget set consists of m_{pt} i.i.d. points generated from the opposite direction

$$\theta_f \ = \ -\theta^* \cdot \frac{\rho_f}{\|\theta^*\|},$$

i.e.

$$\mathcal{D}_f = \{(x_j, y_j)\}_{j=1}^{m_{\mathrm{pt}}}, \qquad x_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0,\ \tfrac{1}{d} I_d\big), \quad y_j \mid x_j = f_{\mathrm{lin}}(x_j^\top \theta_f) - \tfrac{1}{2} + \varepsilon_j,$$

with $\|\theta_f\| = \rho_f \gg \|\theta^*\|$. Because θ_f points in the direction opposite to θ^*, the forget labels are systematically negatively correlated with those in $\mathcal{D}_r$, so ERM on $\mathcal{D}_r \cup \mathcal{D}_f$ is pulled strongly away from θ^*, maximising the gap between the unlearned model trained with $\mathcal{D}_f$ and the one trained with the empty forget set, and therefore tightening the resulting audit lower bounds.

The full dataset comprises $n_{\mathrm{pt}} = 1000$ points, of which $m_{\mathrm{pt}} = 400$ form the forget set, leaving $n_{\mathrm{ret}} = 600$ retain points, with $r = 10$, $\eta = 0.5$, $\kappa = 0.9$, and $\rho_f = 10$.

Remark O.1. Output perturbation yields a tighter audit than Perturbed Newton in this experiment because its projection-diameter bound $\Delta_{\mathrm{OP}} = 2C_0$ is closer to the observed pre-noise separation than the conservative Lemma 3 bound Δ_{PN} used for Perturbed Newton.

P COMPUTATIONAL RESOURCES

All experiments were performed on NVIDIA RTX 4000 GPUs for Shakespeare or CIFAR100. Each individual run (one train + unlearn cycle for a given configuration) took roughly 45 minutes to 1 hour for Shakespeare or CIFAR100 dataset. Since we audit several unlearning algorithms with multiple calibration and evaluation runs each, the full set of experiments reported in the paper took on the order of two weeks of wall-clock time.

The TOFU experiments were performed on NVIDIA A100-SXM4-80GB GPUs. A single run here comprises one fine-tune of Llama-3.2-1B-Instruct on $\mathcal{D}(S)$ followed by all five unlearning algorithms branching from that shared checkpoint, so the expensive stage is paid once rather than five times. A run took a median of 52 minutes for a calibration run and 62 minutes for an evaluation run, which additionally scores the TOFU utility suite, at a peak of 15.0 GiB of GPU memory — the peak is set by NPO-RT's frozen reference copy; fine-tuning alone peaks at 11.6 GiB. The fine-tune dominates the wall clock at a median of 17.2 minutes ($1{,}190$ optimizer steps), while the unlearning stages themselves range from 22 s (gradient ascent, 2 epochs on 200 examples) to 394 s (retain fine-tuning, 450 steps on $\mathcal{D}_r$).