THE CALIBRATION SET IS IN THE CHECKPOINT: HOW RELEASE GEOMETRY GOVERNS MEMBERSHIP LEAKAGE IN POST-TRAINING QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantizing a model often uses a small calibration set after training is complete. This set can remain encoded in the released checkpoint. In SmoothQuant-style methods, comparing the released checkpoint with its public base model recovers activation statistics computed from the calibration data. A maximum can copy a calibration record's activation exactly into the checkpoint, while averaging removes this exact constraint but can keep other membership signal. We show that leakage grows on a common scale, the number of released channels per calibration record (d/k), while the released statistic sets its shape. On OPT-125M, the reference SmoothQuant implementation and NVIDIA Model Optimizer's stock exporter each identify about 68% of calibration records at measured false-positive rates below 10^{-4}. Changing only the statistic in a Qwen2.5-7B exporter changes detection at low false-positive rates by an order of magnitude at essentially unchanged perplexity, and the same effect appears in isotonic probability calibration. Post-training calibration can therefore itself be a privacy-relevant release mechanism.

1 INTRODUCTION

Most membership-inference work targets training-set inclusion. Yet many deployment steps still read data afterwards: post-training quantization (PTQ) and probability calibration use a small calibration set and then discard it. Because they take no gradient steps, they can appear privacy-neutral. But the artifact they produce still depends on that data, and it often travels beyond the boundary that protected the calibration set.

SmoothQuant-style PTQ is a clear case. For each channel it measures a statistic of the calibration activations, such as the largest activation, and folds it into the model's normalization parameters; comparing the released checkpoint with its public base model recovers that statistic in closed form (§2). Suppose a candidate record's activation on some channel is 7. If the candidate supplied the channel's maximum (we say it *owns* the channel), the checkpoint stores exactly 7: the maximum copies the record into the release. If the exporter instead averaged the two largest values, it would store $(7 + y)/2$ for another record's value y, and knowing the candidate would no longer fix the stored value.

This is the paper's central distinction. A maximum can preserve a record exactly, so an exact match tells the attacker the candidate was used; we call such a test a *certificate*, and a statistic that copies one record's value a *selector*. A *mixture* such as a top-2 mean blends records and rarely preserves any one of them, but it still shifts the release, so the attacker can rank candidates without an exact match. The two kinds of leakage look very different: from $m = 1$ to $m = 2$ in a top-m mean, AUC barely moves while detection at low false-positive rates falls fourfold (Fig. 2), and isotonic probability calibration identifies a third of its calibration records by exact match at a membership AUC of only 0.67 (§4). High dimension amplifies both: a record supplies any one of d channels with probability $1/k$, but d/k of them on average. In short, scale is d/k; geometry determines shape: exact constraints for a maximum, jumps for a top-m mean, a smooth shift for a mean.

We make three contributions. (1) We recover calibration statistics from real PTQ checkpoints and use them for membership inference on four stacks (§2, §4). (2) We give a geometric criterion for

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

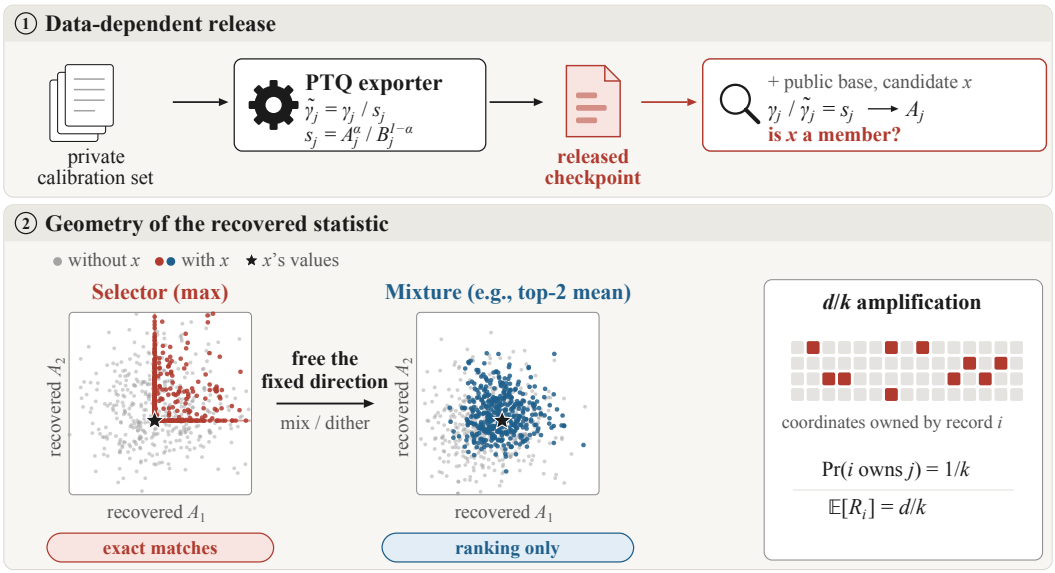

Figure 1: **Overview.** ① A PTQ exporter fuses a calibration statistic A_j into the released gains, which an adversary with the public base inverts in closed form. ② On two simulated coordinates ($k = 6$), statistics computed with candidate x concentrate on the rays where x owns a coordinate under a maximum (certificates), but form a shifted full-dimensional cloud under a top-2 mean (ranking only; signed ranges behave alike). Mixing or dithering frees the fixed direction (§5). A record owns d/k of d coordinates on average (grid: $d = 60$, $k = 6$).

exact membership certificates, namely that fixing a candidate must remove degrees of freedom from the release, and derive a Poisson model of membership for releases built from each channel's largest values, which yields a likelihood-ratio attack and predicts its ROC from public data (§3). (3) We show that maxima and means share the signal scale d/k but differ in shape, and that this distinction predicts leakage across real exporters and probability calibration and explains why blurring a maximum removes exact matches long before it removes membership signal (§3–§5). App. A maps every claim to its evidence.

2 The checkpoint releases a calibration statistic

For a linear layer preceded by LayerNorm, canonical SmoothQuant (Xiao et al., 2023) computes a calibration-set activation magnitude A_j, a public weight magnitude B_j, and

$$s_j = A_j^{\alpha} / B_j^{1-\alpha}, \qquad \widetilde{\gamma}_j = \gamma_j / s_j, \tag{1}$$

where γ_j and $\widetilde{\gamma}_j$ are the public-base and released normalization gains. Consequently,

$$s_j = \frac{\gamma_j}{\widetilde{\gamma}_j}, \qquad A_j = \left[\frac{\gamma_j}{\widetilde{\gamma}_j} B_j^{1-\alpha}\right]^{1/\alpha} = B_j \left(\frac{\gamma_j}{\widetilde{\gamma}_j}\right)^2 \quad (\alpha = 0.5). \tag{2}$$

The smoothing division lands in floating-point normalization parameters; the ratio therefore passes through later integer weight quantization intact.

Proposition 1 (Checkpoint recovery)**.** *On nonzero, unclamped coordinates, the base-to-released gain ratio recovers s_j and hence A_j under the canonical transformation, up to storage error. If an exporter also fuses one positive multiplier per site, recovery is up to that multiplier; a site-normalized membership score cancels it.*

Threat model. The adversary holds a released smoothed/quantized checkpoint and its public base model, knows the exporter configuration as published with the artifact, and can evaluate a candidate

Table 1: What each audited PTQ stack fuses into the released checkpoint, from source (App. H gives where each statistic is stored).

Stack (source audit)	Fused activation statistic	Type	Attacked here
MIT SmoothQuant (reference)	per-channel $\max_t \lvert a_{tj}\rvert$	selector	yes
Intel Neural Compressor (PyTorch)	per-channel $\max_t \lvert a_{tj}\rvert$	selector	–
NVIDIA Model Optimizer	per-channel $\max_t \lvert a_{tj}\rvert$, exported as a tensor	selector	yes
torchao (prototype SmoothQuant)	running per-channel $\max_t \lvert a_{tj}\rvert$, stored	selector	–
LLM Compressor	per-channel signed range	two-endpoint mixture	yes
Intel ONNX SmoothQuant	pooled 99.999th percentile of $\lvert a\rvert$	near-top order statistic	yes
ONNX Runtime static quantization	per-tensor min/max or histogram threshold	selector or histogram	–
AWQ	per-channel mean of $\lvert a\rvert$, searched exponent	regular	–

Table 2: Calibration membership from released checkpoints on four PTQ stacks (512 calibration records; protocol gives draws and nonmembers per draw; *attacks selected before any artifact of that experiment existed). Rows within a stack change only the calibration statistic. †On a separate 178,234-nonmember benchmark (§4).

Stack	Model	Calibration statistic	Recovery ε	Protocol	AUC	TPR@1%	TPR at zero obs. FP
MIT SmoothQuant (reference)	OPT-125M	absolute max	2.7×10^{-7}	3 draws, 7,213	0.994	0.920	0.68 at 7×10^{-5} †
NVIDIA Model Optimizer 0.47, stock	OPT-125M	absolute max	2.9×10^{-4}	3 draws, 178,234*	0.995	0.923	0.42
LLM Compressor 0.13, stock	Qwen2.5-7B	signed range	4.7×10^{-3}	3 draws, 2,000*	0.771	0.104	0.007
same path, one modifier	Qwen2.5-7B	absolute max	4.4×10^{-3}	3 draws, 2,000*	0.863	0.234	0.081
same path, one modifier	Qwen2.5-7B	top-16 mean	4.5×10^{-3}	3 draws, 2,000*	0.824	0.109	0.016
Intel ONNX SQ + ORT INT8	OPT-125M	99.999th pct.	1.9×10^{-7}	3 draws, 1,000	0.940	0.396	0.159
same stack	Llama-3.2-1B	99.999th pct.	7.3×10^{-7}	3 draws, 1,000*	0.869	0.219	0.047

record, but sees no calibration text. We consider the setting where calibration uses sensitive inputs and the resulting checkpoint crosses a broader access boundary, such as an internal registry or a vendor or edge transfer. This setting is supported by current deployment guidance: LLM Compressor's INT8 recipe recommends calibration data that "closely matches the type of data used in deployment", for a fine-tuned model "a sample of your training data", starting from 512 samples (Red Hat AI and vLLM Project, 2026b). The remaining attack inputs are often public as well: among 642 public Hub artifacts carrying a per-channel smoothing scale, the parent checkpoint is public for 56–83% and the smoothing strength α is stated for 99%; inexact candidates and domain-matched private calibration data (Enron email) leave the leak intact (App. H).

Can the statistic be recovered from a saved checkpoint? Yes, on every stack we ran, to within storage precision (Table 2). The MIT implementation fuses the per-channel maximum and LLM Compressor 0.13.0 the signed range $R_j = \max a_j - \min a_j$ (with weight factor $2B_j$), both through the gains; NVIDIA Model Optimizer 0.47 stores the scale outright, as `pre_quant_scale` $= 1/\max_t \lvert a_{tj}\rvert$ under its default $\alpha = 1$; and Intel's ONNX SmoothQuant fuses a pooled 99.999th percentile. Our source audit finds the per-channel maximum, a selector, in four of the eight stacks examined, and floating-point gains or an explicitly stored scale in every audited pipeline, so the statistic remains after integer weight storage (Table 1, App. H).

Table 2 previews the end-to-end result of §4: every stack we attack leaks membership, and through one production exporter, the calibration statistic alone changes detection at low false-positive rates by an order of magnitude.

3 Two geometries of membership leakage

Three results organize the leakage. Theorem 1 determines when a release can identify a record exactly, Theorem 2 predicts the rest of the signal for releases built from each channel's largest values, and Theorem 3 shows that maxima and means share one signal scale.

Exact identification needs the candidate to force the output onto a smaller set. If including a candidate x pins part of the release, as the maximum pins an owned channel to 7, the possible outputs lie on a lower-dimensional set, and an independent nonmember lands there with probability zero. A test that accepts that set is a certificate: it accepts some members at *population* FPR zero, which is stronger than observing no false positive in a finite sample. If another record still contributes

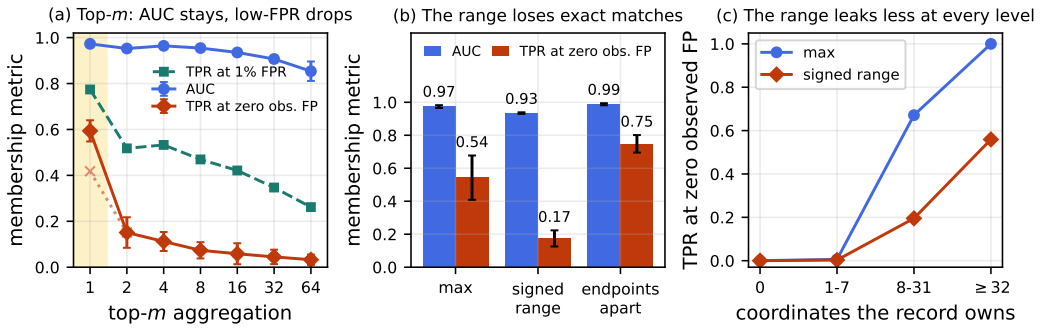

Figure 2: **Ranking quality and low-FPR detection come apart.** (a) AUC declines gradually across top-m, while TPR at zero observed FP drops sharply from the selector ($m = 1$, shaded) to averages of two or more records; the dotted line holds one attack fixed. (b) The maximum (as in SmoothQuant), the signed range (as in LLM Compressor) and the two endpoints released separately: the range is a function of the endpoints but loses most of their TPR at zero observed FP. (c) TPR at zero observed FP by the number of coordinates a member owns: the mixture leaks less at every ownership level.

freely, as y does in $(7 + y)/2$, the output stays full-dimensional. Membership then only shifts its distribution, which supports ranking. Maxima, top-m means and signed ranges all have the form Theorem 1 assumes: once the records are ordered, the release is an affine function of them.

Theorem 1 (Exact identification requires rank loss)**.** *Suppose that on each cell σ of a finite partition (for selectors, a strict ordering of the records) the membership release is $Y = L_\sigma Z + b_\sigma(x) \in \mathbb{R}^d$, where the other records' features Z have a density, and that the nonmember release law $P_{0,x}$ has a density on $\mathbb{R}^d$ that is positive almost everywhere on the full-rank membership images. Then the largest power at population FPR zero is $\pi_x = \Pr_x(\rho_\sigma > 0)$ with $\rho_\sigma = d - \operatorname{rank}(L_\sigma)$: x has a certificate exactly when fixing x lowers the local dimension of the release.*

Isotonic calibration stores a variable number of knots, so it gets its own lemma.

Lemma 1 (Knot certificates)**.** *If nonmember base scores have an atomless distribution, a nonmember's score equals one of the finitely many released knots with probability zero, while a member whose score is kept as a knot matches it exactly.*

The theorem and the lemma sort common releases by when exact identification is possible:

Release	Exact identification when	Signal otherwise
maximum	candidate supplies any coordinate	exclusions (candidate above release)
top-m mean, $m \geq 2$	generically never	ranking
signed range	candidate supplies both endpoints	ranking
isotonic calibration	candidate's score is kept as a knot	weak ranking
sigmoid, temperature scaling	never	none detected

Rounding turns exact matching into matching within a tolerance: a stored 7 matches every candidate within ϵ of 7, and a nonmember falls in that band with probability of order ϵ^ρ (Prop. 2, App. B).

Changing only the statistic changes leakage as predicted. We fix one OPT-125M capture ($d = 18{,}432$ channels) and select attacks on development draws only. Replacing the maximum by the top-2 mean barely moves AUC (0.97 to 0.95) but cuts TPR at zero observed FP from 0.59 to 0.15 (Fig. 2a). The signed range, a deterministic function of the two endpoints, keeps 0.17 of the endpoints' 0.75, because few members own both endpoints of a coordinate (Fig. 2b). Shuffled-channel and excluded-record controls stay at chance (App. E).

Tail position predicts the remaining signal. For releases built from each channel's largest few values, one change of scale makes every channel look the same. Measure a value v on channel j by

its *tail position* $e = k(1-F_j(v))$, the expected number of calibration records above v; a candidate at tail position s_j has value $F_j^{-1}(1 - s_j/k)$. On this scale the attacker needs only where the candidate sits in each channel's tail: every continuous channel gives the same pair of member and nonmember release distributions, which we call the membership experiment.

Theorem 2 (Poisson limit for order-statistic membership)**.** *Let records be i.i.d. with continuous channel marginals F_j, and let the release on channel j be a function of its r largest values. (a) On one channel the membership experiment is the same for every F_j and every k: the records' tail positions are independent and uniform on $(0, k)$, the nonmember release depends on the r smallest of k of them, and the member release on the r smallest of $k-1$ of them together with s_j. (b) As $k \to \infty$ it converges in total variation distance, at rate $O(1/k)$ for fixed r, to a Poisson model: the r smallest points of a unit-rate Poisson process on $(0, \infty)$, without or with an added point at s_j. (c) On channels that are independent within records, the experiment is the product of the per-channel experiments, and a small multiplicative change of the release adds noise of known width on the log tail-position scale (Prop. 4, App. B).*

For a maximum, membership adds the candidate as one more possible extreme point. This creates two signals: an exact match when the candidate supplies the maximum, and an exclusion when the candidate exceeds the released value. In the example, the likelihood ratio is infinite on a match of 7, zero above 7 and one otherwise. The best test has power $\min(1, \delta_x + \alpha)$, where $\delta_x = 1 - \prod_j (1 - e^{-s_j})$ is the probability that the candidate owns some channel (Cor. 2, App. B). For a top-m mean the exact match becomes finite jumps in the likelihood ratio, and a small change to the maximum spreads it over a narrow band (part (c)).

Because the experiment depends on the candidate only through its tail positions, public reference data predict which records are most exposed and how well the attack works. A record's expected number of owned channels, $\lambda_x = \sum_j e^{-s_j}$, averages d/k and ranks records by exposure (Spearman 0.94 with realized ownership, App. E). The likelihood ratio of this Poisson model gives the Neyman–Pearson optimal attack for the model, and simulating the model at each candidate's tail positions predicts the attack's ROC. The prediction is accurate for maxima and small changes to them. Correlation between channels makes it optimistic for larger top-m means (App. E).

Scale is d/k; geometry determines shape. At the two extremes, maxima and means leak in very different ways, but share the same natural signal scale, d/k.

Theorem 3 (Maxima and means share a d/k signal scale)**.** *(i) If the k records are i.i.d. with a joint density on $\mathbb{R}^d$, the likelihood ratio of a candidate given the released maximum is infinite on an exact match, zero if the candidate exceeds the release on some channel, and otherwise $c(y) = f_{k-1}(y)/f_k(y)$, the same for every candidate. For exchangeable records, a release of D values each supplied by one record gives a record D/k of them in expectation. (ii) If records are $N(\mu, \Sigma)$, membership shifts the released mean by $(x-\mu)/k$: up to a variance factor $(k-1)/k$, a Gaussian shift of strength $\mu_x^2 = (x-\mu)^\top \Sigma^{-1}(x-\mu)/k$, with $\mathbb{E}\,\mu_X^2 = d/k$ (Azize & Basu, 2025).*

For a maximum, a larger d/k means that a record controls more released coordinates: because the records are interchangeable, each record owns each channel with probability $1/k$, and Fig. 3 shows the count tracking d/k across three orders of magnitude. For a mean, a larger d/k means a larger standardized membership shift.

The statistic sets how that signal appears in the ROC: exact matches and exclusions for maxima, finite jumps for top-m means, and a smooth shift for means. Figure 4 shows the three shapes at one d/k. All three rank members well, but at a false-positive rate of 10^{-4} the maximum still detects 75% of members, against 2% for the top-2 mean and 1% for the mean. For the mean, the prediction includes a factor $1 - d/n$ for a covariance estimated from n references (App. B). Across numbers of channels and references, the predicted AUC is within 0.03 of the realized one for the mean and for top-m means with $m = 2$–16 (App. E).

These results give an audit rule: ask whether a single calibration record can set a released value. If it can, expect exact or near-exact matches. If each released value still depends on other calibration records, exact matching generically disappears and membership remains as a distributional shift. Either way, more released channels per calibration record amplify the membership signal.

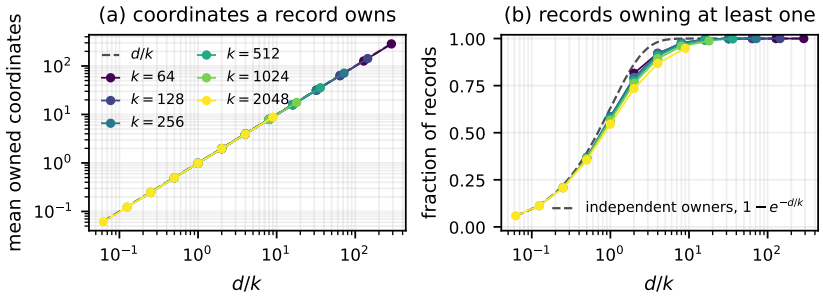

Figure 3: **Ownership scales as d/k.** Twenty random calibration sets per cell on a grid of $k \in [64, 2048]$ and $d \in [128, 18{,}432]$ OPT-125M channels. (a) A record owns d/k coordinates on average for every k. (b) The fraction of records owning at least one coordinate follows d/k closely, slightly below the independent-owner curve.

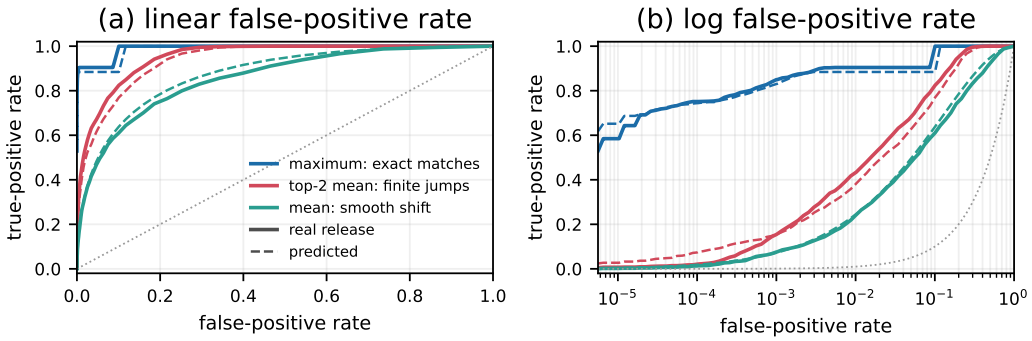

Figure 4: **One scale, three shapes.** Maximum, top-2 mean and mean of the same OPT-125M calibration draws ($d/k = 4$; three draws, 178,234 held-out nonmembers), each attacked by its exact likelihood ratio (solid) and compared with the theory's prediction (dashed; numbers in Table 15). (a) All three rank members well. (b) On a log scale, the maximum keeps substantial power at the smallest false-positive rates, where the top-2 mean and the mean have little.

4 Does this happen in real exporters?

Does a real exporter expose members at very low false-positive rates? Yes, for both the reference implementation and a stock production exporter that fuse the maximum. We ran the pinned MIT SmoothQuant implementation unchanged, and NVIDIA Model Optimizer's INT8 SmoothQuant unmodified with its own calibration loop in native float16, on the same three 512-record OPT-125M draws from the Pile validation set. Attacks were fixed on development data, and thresholds come from a disjoint null split.

Both exporters let the attacker detect about 68% of members at nominal FPR 10^{-4} (measured 7.3×10^{-5} and 9.5×10^{-5}), and about 41% with no false positive among the 178,234 nonmembers; all controls stay at chance (Table 2, App. D).

Does changing only the released statistic change the attack? LLM Compressor 0.13.0 fuses the signed range. We selected attacks on Qwen2.5-1.5B, then exported three calibration draws of Qwen2.5-7B through the same exporter with three calibration statistics: the stock range, the absolute maximum, and a top-16 mean of per-record maxima (App. D).

Changing only the statistic changes the attack as predicted. The maximum raises AUC from 0.77 to 0.86 and TPR at zero observed FP from 0.007 to 0.081; the top-16 mean improves ranking but stays near the range's low-FPR detection. We apply two attacks to every arm (App. Table 8):

Table 3: Beyond PTQ: stock scikit-learn calibration of fixed classifiers, changing only the method (three draws; remaining deduplicated test records are nonmembers). TPR at zero obs. FP: exact knot matches (isotonic) or calibrated loss (others); AUC: best of the two attacks; ECE and accuracy on nonmembers.

Fixed base and calibration set	Stock calibrator	TPR at zero obs. FP	AUC	ECE	Accuracy
20 Newsgroups, linear SVM, $k = 1{,}024$	isotonic	0.341 ± 0.013	0.666	0.080	0.708
	sigmoid	0.000	0.499	0.127	0.715
	temperature	0.000	0.496	0.064	0.722
CIFAR-100, ResNet-56, $k = 1{,}024$	isotonic	0.655 ± 0.009	0.827	0.063	0.682
	sigmoid	0.000	0.514	0.225	0.712
	temperature	0.000	0.503	0.026	0.726
CIFAR-100, ResNet-56, $k = 5{,}000$	isotonic	0.379 ± 0.007	0.687	0.115	0.712
	temperature	0.000	0.503	0.025	0.725

a matching attack, which looks for exact matches, and a ranking attack, which scores the whole release distribution. Only the maximum gives the matching attack low-FPR power; the ranking attack finds similar, weak signal on all three. On $16{,}384$ fresh nonmembers the maximum detects 12% of members at a measured FPR of 0.087%, against 4.8% for the range at 0.14% (App. D). The three statistics give the same perplexity to within draw-to-draw variation.

The attack transfers to a third stack and model family. Intel's ONNX SmoothQuant with ONNX Runtime INT8 fuses a pooled 99.999th percentile, which lies among the top few pooled activations; the attack reaches AUC 0.94 on OPT-125M and 0.87 on Llama-3.2-1B (Table 2, App. D).

The mechanism also appears in stock probability calibration. We calibrate a TF-IDF linear SVM, trained on the official 20 Newsgroups training split (Lang, 1995), with scikit-learn (Pedregosa et al., 2011) on three draws of $k = 1{,}024$ test records, changing only the method. The other $6{,}260$ deduplicated test records are nonmembers (App. F). Isotonic regression (Zadrozny & Elkan, 2002) stores its knots as exact calibration scores; sigmoid (Platt, 1999) and temperature scaling (Guo et al., 2017) store only smooth fitted parameters.

Only 2% of calibration scores remain as knots, yet an exact knot match identifies 34% of members with no match among the $6{,}260$ nonmembers, at an AUC of only 0.67 (Table 3); the smooth calibrators are at chance. The scale law predicts both numbers (Theorem 3): the K knots give members $K/k = 0.41$ owned knots (measured 0.41), so $1 - e^{-K/k} = 0.34$ are certified (measured 0.34). A fixed public CIFAR-100 ResNet-56 replicates this, also with logits recomputed at another batch size (App. F).

5 Why simple fixes leave the leak in place

If a maximum leaks because it copies a record, the obvious fix is to blur it. Blurring removes exact matches quickly, but it removes membership signal much more slowly.

Small changes remove exact matches quickly. Return to the candidate with activation 7. If the exporter releases 7 plus a small change, the attacker can no longer demand an exact match, but a release close to 7 is still suspicious. The key quantity is how concentrated the change is near zero. If most changes are tiny, a released value close to the candidate remains unusually likely. At fixed RMS distortion D, more diffuse changes therefore hide exact matches better. We measure concentration by κ, the peak density of the change in units of its RMS size. The change may depend on the calibration set, as in top-2 mixing, or be added random noise, which we call a dither.

Corollary 1 (Small perturbations of the maximum)**.** *Under the regularity conditions of App. G, as the change shrinks, the most powerful test at level α for a candidate with value x has power $\beta^\star(\alpha) = \frac{\alpha\,\kappa}{D\,k f(x)}(1 + o(1)) + O(\alpha)$, where f is the channel's density.*

At fixed RMS distortion, the uniform change has the smallest possible peak concentration ($\kappa = 1/\sqrt{12}$). Top-2 mixing, the obvious fix, moves 7 a small fraction of the way toward the runner-up

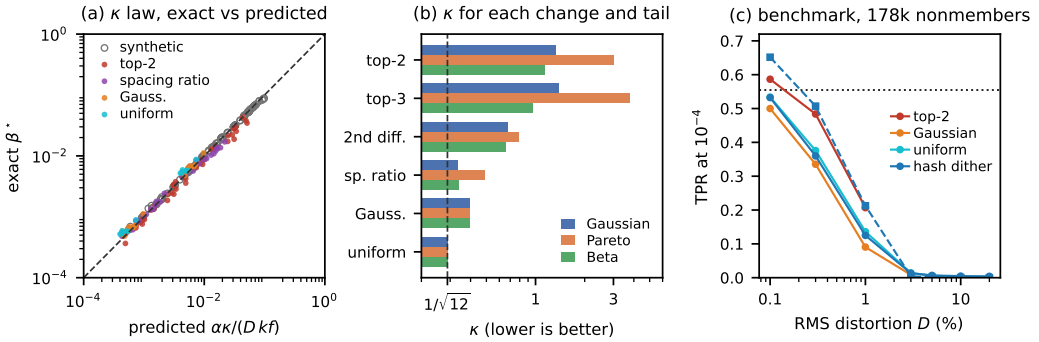

Figure 5: **Blurring the maximum.** (a) Exact Neyman–Pearson power against Corollary 1 (synthetic tails; eleven real OPT-125M channels). (b) κ for each change and tail family; the uniform attains $1/\sqrt{12}$. (c) Matching attack (solid) and, for the hash dither, the full attack, which adds distributional evidence (dashed); dotted: the unmodified maximum. TPR at 10^{-4} falls near 1% distortion, while exclusions at higher FPR persist much longer.

y. That gap is often small, so the change is concentrated near zero: κ is 1.13–3.05, against 0.40 for Gaussian noise. A deterministic hash-based dither closely matches uniform noise empirically (App. G).

On the Qwen export with fp32 gains, every 1% modification cuts matching-attack detection at 0.1% FPR from 0.60 to 0.01 at unchanged perplexity. The κ law matches exact power on eleven real channels (median ratio 1.01, Fig. 5a).

Exclusions persist much longer. A member never exceeds what the release allows; a nonmember often does. This creates a one-sided test: values that are too large rule out membership even after the exact match has been blurred. For a multiplicative dither of relative size η, the chance that a nonmember is excluded on a channel has an exact formula whatever the dependence across channels, and on Pareto-like tails it decays only like $1/\eta$, much more slowly than exact matching (Prop. 5, App. G).

With F_j estimated from public reference ranks, this law predicts 16.2, 12.5 and 2.6 exclusions per nonmember at 0%, 1% and 10% distortion on the Qwen2.5-1.5B export, against 15.9, 12.0 and 2.4 observed. The likelihood-ratio attack built on Theorem 2 therefore still detects 82% of members at 0.1% FPR after a 1% change and keeps AUC 0.79 after a 10% change. On the OPT-125M benchmark, the Poisson model predicts this attack's ROC within 0.08 at every operating point (Table 4, App. Table 13). On all $86{,}016$ channels a nonmember would still be excluded on about 27 channels at 10% distortion and on 5 under a $\pm 50\%$ dither.

Small changes to the statistic trade exact matches for a smoother ROC. Bounding membership leakage requires randomized, formally private statistics.

6 Discussion, Related Work, and Limitations

Implications and guidance. Privacy analysis of released models should cover every data-dependent stage that shapes the artifact, especially statistics that one record can set, such as maxima, other top values and isotonic knots. Because AUC captures only part of the leakage, audits should report low-FPR detection, rank records by their exposure λ_x, and use likelihood-ratio attacks when the released statistic is known; bounding leakage needs a formal mechanism such as differentially private statistics.

Related work. Membership inference targets training-set inclusion (Shokri et al., 2017; Yeom et al., 2018). Likelihood-ratio formulations go back to Sablayrolles et al. (2019), difficulty calibration reduces false positives (Watson et al., 2022), and Carlini et al. (2022) argue for evaluation at

Table 4: Modifications of the maximum at 1% RMS distortion (three draws; means). The matching attack scores only exact-match evidence on owned coordinates; the full attack adds distributional evidence; the exact-LR attack uses the per-channel likelihood ratios of Theorem 2, and the predicted row simulates the Poisson model at each candidate's tail positions. Top: OPT-125M capture, $2{,}044$ channels, $k = 512$, attacker knows the algorithm. Bottom: stock-path LLM Compressor exports of Qwen2.5-1.5B with fp32 gain storage, 512 members and $2{,}000$ nonmembers. The $D = 10\%$ column tests Proposition 5 (App. G). Full tables in App. G.

Setting	Metric	max	top-2 mix	hash dither	uniform noise	Gaussian noise	hash, $D = 10\%$
OPT capture, 178,234 held-out	TPR@10^{-4}, full attack	0.643	0.331	0.213	0.192	0.159	0.000
	TPR@10^{-4}, exact-LR attack	0.751	0.364	0.208	0.205	0.170	0.003
	TPR@10^{-4}, predicted (Thm. 2)	0.744	0.292	0.199	0.207	0.176	0.010
	AUC, exact-LR attack	0.995	0.987	0.978	0.979	0.972	0.698
Qwen2.5-1.5B, stock path, fp32 gains	TPR@0.1%, matching attack	0.597	0.011	0.012	–	–	0.012
	TPR@0.1%, full attack	0.965	0.525	0.848	–	–	0.013
	TPR@0.1%, exact-LR attack	0.949	0.845	0.822	–	–	0.054
	AUC, exact-LR attack	1.000	0.998	0.997	–	–	0.789
	WikiText-2 perplexity	10.609	10.610	10.610	–	–	10.600

very low FPR. LLM work studies extraction and memorization (Carlini et al., 2021; 2023; Mattern et al., 2023; Duan et al., 2024). Azize & Basu (2025) show that the leakage of a released mean about one record depends on that record's distance from the mean after accounting for covariance (its Mahalanobis distance), the Gaussian end of Theorem 3; ours are the exact-matching end and the shared d/k scale. Data curation also leaks membership without a gradient step (Wahdany et al., 2026). Membership inference from released statistics goes back to genomic summaries (Homer et al., 2008; Sankararaman et al., 2009) and continuous biomedical summary statistics (Backes et al., 2016), with recent work revisiting attacker assumptions (Berrang et al., 2026) and general statistical releases (Oakley et al., 2026). Stored model statistics also support data reconstruction: DeepInversion synthesizes training-like images from a checkpoint's BatchNorm statistics (Yin et al., 2020); here the statistics come from a separate calibration set after training and support record-level membership tests. We build on f-DP trade-offs (Dong et al., 2022; Leemann et al., 2023), optimal noise (Geng & Viswanath, 2016; Geng et al., 2019) and extreme-value limits (Resnick, 1987; de Haan & Ferreira, 2006) to derive the membership experiment that post-training releases create, its likelihood-ratio tests and a scale law that predicts real checkpoints. The choice of calibration set affects PTQ quality (Williams & Aletras, 2024); we study what the checkpoint keeps about those records. Prior PTQ privacy work studies how compression changes membership leakage about the model's *training* data (Li et al., 2026; Zhang et al., 2025; Aubinais et al., 2026) or exploits runtime dynamic quantization (Foerster et al., 2026); we take the calibration records themselves as the private population and ask whether their statistics are encoded in the static artifact. Post-hoc probability calibration (Platt, 1999; Zadrozny & Elkan, 2002; Guo et al., 2017) is another data-dependent release step; private isotonic regression (Ghazi et al., 2022) gives a formally private alternative, whereas we study the ordinary non-private release.

Limitations. Our attacks require the public base model and knowledge of the calibration transform. How often deployed calibration sets contain sensitive records remains to be measured. Exact certificates become finite-precision matching tests in stored checkpoints, and the product likelihood approximation is less accurate for correlated top-m releases. The modifications of §5 reduce leakage empirically and carry no formal privacy guarantee.

7 Conclusion

Model privacy extends past training: a data-dependent post-training transformation is itself a release mechanism. The number of released channels per record, d/k, sets the scale of membership leakage, and the geometry sets its shape: exact constraints for a maximum, jumps for a top-m mean, a smooth shift for a mean. Modifying a maximum blurs matches much faster than exclusions. Calibration records can remain identifiable after they are discarded because their values are encoded in the checkpoint. Privacy audits should examine the statistics kept from calibration as well as the model's training procedure.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods: they helped write and edit the code for the quantization pipelines, attacks and experiments. We also used generative AI tools to draft and edit parts of the paper for readability. The authors reviewed and tested all AI-assisted code and checked all AI-assisted text against the results and proofs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments use public benchmark data (Pile-10k, the Pile validation set, WikiText-2, 20 Newsgroups, CIFAR-100, and the AESLC Enron email corpus as a deployment proxy) and public model checkpoints; no private data were used or exposed. The paper shows a privacy risk in a widely used quantization pattern and evaluates it on artifacts we produced ourselves; we make no claim that any real private calibration set has been recovered. The mechanism is a property of published algorithms rather than an implementation defect, and the mitigations discussed are algorithmic choices exporters already expose.

REPRODUCIBILITY STATEMENT

Attack choices and hyperparameters were fixed on development data before evaluation; target artifacts and held-out test sets were unseen during selection. App. A maps every claim to its protocol and results and lists the software versions, exporter commits, models and data; proofs are in App. B. The code, protocols and a command for each table and figure will be released with the paper.

REFERENCES

Eric Aubinais, Philippe Formont, Pablo Piantanida, and Elisabeth Gassiat. An indicator of membership inference security in post-training quantized models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 300 of *Proceedings of Machine Learning Research*, pp. 649–657, 2026.

Achraf Azize and Debabrota Basu. Some targets are harder to identify than others: Quantifying the target-dependent membership leakage. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pp. 397–405, 2025.

Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. Membership privacy in MicroRNA-based studies. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 319–330, 2016. doi: 10.1145/2976749.2978355.

Pascal Berrang, Mark Ryan, and Kiera Wooldridge. Revisiting assumptions for membership inference on summary statistics. *Proceedings on Privacy Enhancing Technologies*, 2026(4):542–560, 2026. doi: 10.56553/popets-2026-0134.

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

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

J. Lawrence Carter and Mark N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences*, 18(2):143–154, 1979. doi: 10.1016/0022-0000(79)90044-8.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer, 2006.

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.

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

Hanna Foerster, Ilia Shumailov, Cheng Zhang, Yiren Zhao, Jamie Hayes, and Robert Mullins. Quantamination: Dynamic quantization can leak your data across the batch. *arXiv preprint arXiv:2604.26505*, 2026. ICML 2026 Workshop on Technical AI Governance Research.

Quan Geng and Pramod Viswanath. The optimal noise-adding mechanism in differential privacy. *IEEE Transactions on Information Theory*, 62(2):925–951, 2016. doi: 10.1109/TIT.2015.2504967.

Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Optimal noise-adding mechanism in additive differential privacy. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *PMLR*, pp. 11–20, 2019.

Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Private isotonic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 8577–8590, 2022. doi: 10.52202/068431-0624.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

L. R. Haff. An identity for the Wishart distribution with applications. *Journal of Multivariate Analysis*, 9(4):531–544, 1979. doi: 10.1016/0047-259X(79)90056-3.

Johan Håstad, Russell Impagliazzo, Leonid A. Levin, and Michael Luby. A pseudorandom generator from any one-way function. *SIAM Journal on Computing*, 28(4):1364–1396, 1999. doi: 10.1137/S0097539793244708.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):e1000167, 2008. doi: 10.1371/journal.pgen.1000167.

Hugging Face. Quantization. `https://huggingface.co/docs/optimum/concept_guides/quantization`, 2026. Optimum concept guide, accessed 2026-09-26.

Russell Impagliazzo, Leonid A. Levin, and Michael Luby. Pseudo-random generation from one-way functions. In *ACM Symposium on Theory of Computing (STOC)*, pp. 12–24, 1989. doi: 10.1145/73007.73009.

Intel. Intel Neural Compressor. `https://github.com/intel/neural-compressor`, 2026. Commit 6007bbb22242, source audited 2026-09-20.

Ken Lang. NewsWeeder: Learning to filter netnews. In *International Conference on Machine Learning (ICML)*, pp. 331–339, 1995. doi: 10.1016/B978-1-55860-377-6.50048-7.

Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Gaussian membership inference privacy. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 73866–73878, 2023. doi: 10.52202/075280-3231.

Na Li, Yansong Gao, Hongsheng Hu, Boyu Kuang, and Anmin Fu. CompLeak: Deep learning model compression exacerbates privacy leakage. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 6545–6564, 2026.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 6, pp. 87–100, 2024.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, 2023. doi: 10.18653/v1/2023.findings-acl.719.

NVIDIA Corporation. NVIDIA Model Optimizer. `https://github.com/NVIDIA/Model-Optimizer`, 2026a. Commit b311c054de40, source audited 2026-09-20.

NVIDIA Corporation. PyTorch quantization. `https://nvidia.github.io/Model-Optimizer/guides/_pytorch_quantization.html`, 2026b. NVIDIA Model Optimizer documentation, accessed 2026-09-26.

Lisa Oakley, Sam Stites, Cameron Moy, Steven Holtzen, Alina Oprea, and Marco Gaboardi. A Bayesian approach to membership inference for statistical release. *arXiv preprint arXiv:2605.30203*, 2026.

ONNX Runtime developers. ONNX Runtime. `https://github.com/microsoft/onnxruntime`, 2026. Static quantization; commit 09dfa6ad06ed, source audited 2026-09-20.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Red Hat AI and vLLM Project. LLM Compressor. `https://github.com/vllm-project/llm-compressor`, 2026a. Version 0.13.0.

Red Hat AI and vLLM Project. int8 weight and activation quantization. `https://github.com/vllm-project/llm-compressor/tree/0.13.0/examples/quantization_w8a8_int8`, 2026b. LLM Compressor 0.13.0, accessed 2026-09-26.

Sidney I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer, New York, NY, 1987. doi: 10.1007/978-0-387-75953-1.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5558–5567, 2019.

Sriram Sankararaman, Guillaume Obozinski, Michael I. Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9):965–967, 2009. doi: 10.1038/ng.436.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

torchao maintainers and contributors. torchao: PyTorch native quantization and sparsity for training and inference. `https://github.com/pytorch/ao`, 2026. Prototype SmoothQuant; commit df97218e877f, source audited 2026-09-20.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Dariush Wahdany, Matthew Jagielski, Adam Dziedzic, and Franziska Boenisch. Curation leaks: Membership inference attacks against data curation for machine learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations (ICLR)*, 2022.

Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 10100–10118, 2024. doi: 10.18653/v1/2024.acl-long.544.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099, 2023.

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. doi: 10.1109/CSF.2018.00027.

Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via DeepInversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8712–8721, 2020. doi: 10.1109/CVPR42600.2020.00874.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002. doi: 10.1145/775047.775151.

Chenxiang Zhang, Tongxi Qu, Zhong Li, Tian Zhang, Jun Pang, and Sjouke Mauw. Bits for privacy: Evaluating post-training quantization via membership inference. In *IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 853–860, 2025. doi: 10.1109/TrustCom66490.2025.00099.

A Setup and reproducibility

Table 5 maps each main-text claim to the appendix that holds its protocol and results.

Table 5: Where each claim's evidence lives.

Claim	Main text	Appendix
The checkpoint reveals the calibration statistic	§2, Prop. 1	C, D
Which stacks fuse which statistic	§2, Table 1	H
The attacker's inputs are often public	§2	H
Exact identification requires rank loss	§3, Thm. 1	B
Maximum vs top-m vs range on one capture	§3, Fig. 2	E
Ownership scales as d/k and is predictable	§3, Thm. 3, Fig. 3	B, E
Poisson limit and its likelihood ratios	§3, Thm. 2	B, E
Scale d/k for maxima and means	§3, Thm. 3, Fig. 4	B, E
Two exporters expose members at 10^{-4} FPR	§4, Table 2	D
Changing only the statistic changes the attack (7B)	§4, Table 2	D
Third stack and model family	§4, Table 2	D
Isotonic vs smooth probability calibration	§4, Table 3	F
Small changes remove exact matches quickly	§5, Cor. 1, Fig. 5	G
Exclusions persist much longer	§5, Table 4	E, G

Protocol. Attack families, hyperparameters, tolerances and channel subsets were fixed on development data before evaluation; target artifacts and held-out test sets were unseen during selection. Thresholds are rank p-values on null-calibration records disjoint from the held-out nonmembers, so each operating point reports a nominal level and the realized held-out rate. Each experiment uses

three independent calibration draws (mean $\pm$ SD across draws); AUCs grant half credit to ties. Every released artifact is saved, reloaded in a separate process and attacked from its stored tensors and the public base alone.

Software, models and data. MIT SmoothQuant at commit `c61476d728e4`; LLM Compressor 0.13.0 with compressed-tensors 0.18.0; NVIDIA Model Optimizer 0.47.0; Intel `onnx-neural-compressor` 1.0 with ONNX Runtime 1.30.0; scikit-learn 1.9.1; PyTorch 2.13–2.14 and Transformers 5.14–5.16. Models are the public OPT-125M, Qwen2.5-1.5B, Qwen2.5-7B and Llama-3.2-1B checkpoints and a public CIFAR-100 ResNet-56; data are Pile-10k, the Pile validation set, WikiText-2, 20 Newsgroups, CIFAR-100 and AESLC. The full protocols and implementation will be released with the code.

B PROOFS

Zero-FPR power is singular mass. Write $P_{1,x} = \pi_x P_{1,x}^{\perp} + (1-\pi_x) P_{1,x}^{\mathrm{ac}}$ with $P_{1,x}^{\perp} \perp P_{0,x}$ and $P_{1,x}^{\mathrm{ac}} \ll P_{0,x}$ (Lebesgue decomposition). A level-0 test is a region R with $P_{0,x}(R) = 0$; its power $P_{1,x}(R)$ is maximized by the support of the singular part, so the supremum power over tests with population FPR zero equals π_x. For a selector, on the event that x owns coordinate j the release equals $a_j(x)$ exactly, which a fresh nonmember matches with probability zero under a continuous population; deleting the owner changes a maximum by the winner–runner-up gap and deleting any other record changes nothing.

Both conditions on $P_{0,x}$ matter. Clipping loses rank but also gives nonmembers an atom on the clip value, so rank loss then certifies nothing; the knots of Lemma 1 are fixed by the calibration set, independent of a fresh nonmember, and finite precision turns every certificate into a collision probability (Prop. 2), which TPR at zero observed FP estimates.

Proof of Theorem 1. By the singular-mass identity above it suffices to characterize the singular cells of the affine release map. On cell σ, a rank-$r < d$ affine map sends the nuisance law to an r-dimensional affine set, which has zero d-dimensional Lebesgue measure and hence zero probability under the assumed non-membership density. If L_σ has full row rank, split the nuisance coordinates into the kernel of L_σ and a d-dimensional complement. Integrating the nuisance density over the kernel gives a Lebesgue density for the pushforward; positivity of the non-membership density on that support gives domination. Summing over finitely many cells makes the singular mass exactly the probability of the rank-deficient cells.

For a maximum, every coordinate owned by fixed x is constant in the nuisance variables, while every unowned coordinate selects a distinct nuisance feature coordinate; the latter rows contain a full-rank diagonal minor. For a top-m mean with $m \geq 2$, each output retains at least one nuisance summand and these feature coordinates are distinct across output coordinates, giving a full-rank diagonal minor. A range coordinate loses all nuisance dependence only when x owns both endpoints; otherwise an upper or lower nuisance endpoint supplies its pivot. For the projective case, let $y_s = \log A_s \in \mathbb{R}^{d_s}$ and $P_s = I - d_s^{-1}\mathbf{1}\mathbf{1}^\top$. Multiplication by an unknown positive site scalar adds $c_s\mathbf{1}$ to y_s, so the observable is $P_s y_s$. Let J_s be the set of r_s coordinates owned by x. When $r_s \geq 1$, choose $j_0 \in J_s$; the $r_s - 1$ functionals $(e_j - e_{j_0})^\top y_s$, $j \in J_s \setminus \{j_0\}$, annihilate $\mathbf{1}$, are linearly independent, and are fixed by the candidate on the cell. All unowned coordinates retain their nuisance pivots. Thus the quotient loses exactly $(r_s - 1)_+$ nuisance dimensions. The full-dimensional density and unconstrained-coordinate assumptions exclude additional deterministic rank loss. □

Proof of Lemma 1. Each knot is the score of a calibration record, so a member kept as a knot matches it exactly; a fresh nonmember's score is independent of the finitely many knots and atomless, so it hits one with probability zero.

Finite precision and unknown multipliers. Records are token sequences and stored releases are rounded, so in general $\mathrm{FPR}_x = P_{0,x}(\mathcal{M}_x)$ and $\mathrm{TPR}_x = P_{1,x}(\mathcal{M}_x)$ for a certificate set $\mathcal{M}_x$: the collision mass $P_{0,x}(\mathcal{M}_x)$ governs the discrete analogue of a certificate.

Proposition 2 (Affine certificate tubes and nuisance dimension)**.** *(i) Let the singular membership component for x lie on finitely many affine sets of codimension $\rho_\ell \geq 1$ inside a compact window, let the ideal nonmember density be at most B on their 2ϵ-neighbourhoods, and let both hypotheses incur observation error at most ϵ. The test accepting the ϵ-neighbourhood retains all singular mass and has $\mathrm{FPR} \leq B \sum_\ell K_\ell \epsilon^{\rho_\ell}$. (ii) If the certificate sets depend on a parameter $\theta \in \Theta \subset \mathbb{R}^q$ unknown to the adversary, and the union $\bigcup_\theta \mathcal{M}_\ell(\theta)$ within the window is covered by finitely many images of L-Lipschitz maps on bounded domains of dimension at most $n = d - \rho_\ell + q$, then for $q < \rho_\ell$ the test that must accept every θ has $\mathrm{FPR} = O(\epsilon^{\rho_\ell - q})$ under the same density bound, and no decay rate is implied when $q \geq \rho_\ell$: a q-dimensional hidden hyperparameter removes at most q certificate constraints.*

For SmoothQuant, $\log s_j = \alpha(\log A_j + \log B_j) - \log B_j$, so a member owning r coordinates of a site retains $r - 1$ equalities free of a hidden α ($q = 1$), the same count as an unknown per-site scale; in practice α is published with the recipe (App. H).

Proof of Proposition 2. (i) An observed nonmember enters the ϵ-acceptance tube only if its ideal release lies in the 2ϵ-tube of some $\mathcal{M}_\ell$. A bounded portion of a codimension-ρ_ℓ affine set has 2ϵ-tube volume at most $K_\ell \epsilon^{\rho_\ell}$; the density bound and a union bound over ℓ give the FPR bound, and every singular membership point is within ϵ of its set by assumption, so no singular mass is lost. (ii) Cover each bounded n-dimensional parameter domain by $O(\epsilon^{-n})$ balls of radius ϵ/L; their images lie in $O(\epsilon^{-n})$ balls of radius ϵ, so the 2ϵ-tube of the union lies in $O(\epsilon^{-n})$ balls of radius 3ϵ, of total volume $O(\epsilon^{-n}\epsilon^d) = O(\epsilon^{\rho_\ell - q})$. A test that must accept every θ accepts the ϵ-tube of the union, whence the bound. The Lipschitz condition holds for an unknown positive site multiplier in a bounded range and for $\alpha \in [0, 1]$. □

Ownership.

Proposition 3 (Ownership scales as d/k)**.** *(i) If the k records are exchangeable and each coordinate has a unique owner, then $\Pr(I_j = i) = 1/k$ for every j regardless of cross-coordinate dependence, so $\mathbb{E}[R_i] = d/k$ exactly. (ii) Coverage obeys $\Pr(R_i > 0) \geq \mathbb{E}[R_i]^2/\mathbb{E}[R_i^2]$ with $\mathbb{E}[R_i^2] = d/k + 2\sum_{j<l} \Pr(I_j = I_l = i)$; independent owners give $1 - (1 - 1/k)^d \to 1 - e^{-d/k}$. (iii) For i.i.d. records with per-coordinate CDF F_j, a fixed target x owns $\lambda_x = \sum_j F_j(m_j(x))^{k-1}$ coordinates in expectation.*

Proof of Proposition 3. (i) Fix j. Unique ownership means exactly one index I_j attains the coordinate. Exchangeability of the k records makes the law of $(m_{1j}, \ldots, m_{kj})$ invariant under permutations, so $\Pr(I_j = i)$ does not depend on i; these k probabilities sum to one, hence each equals $1/k$ whatever the dependence across coordinates. Summing $\mathbb{E}[\mathbf{1}\{I_j = i\}]$ over j gives $\mathbb{E}[R_i] = d/k$. (ii) $R_i^2 = \sum_j \mathbf{1}\{I_j = i\} + 2\sum_{j<l} \mathbf{1}\{I_j = I_l = i\}$ gives the second-moment identity, and the Paley–Zygmund inequality at threshold zero gives $\Pr(R_i > 0) \geq \mathbb{E}[R_i]^2/\mathbb{E}[R_i^2]$. Under independent owners R_i is $\mathrm{Binomial}(d, 1/k)$. (iii) Conditional on x, coordinate j is owned by x iff all $k - 1$ other i.i.d. records fall strictly below $m_j(x)$, an event of probability $F_j(m_j(x))^{k-1}$ for continuous F_j; linearity of expectation gives λ_x without any cross-coordinate assumption. □

Poisson limit for order-statistic releases. Records are i.i.d. with continuous channel marginals F_j; write $S_j = 1 - F_j$ and $e_j(v) = kS_j(v)$ for the tail position of a value on channel j. A candidate is specified by fixed tail positions s_j, i.e. it is the sequence $x_{k,j} = F_j^{-1}(1 - s_j/k)$; a fixed ordinary value has $kS_j(x) \to \infty$ and carries no information in the limit. The release on channel j is $T_j = \psi(Z_{(1),j}, \ldots, Z_{(r),j}; V_j)$ with V_j i.i.d. noise independent of the data. *Asymptotic tail independence* means $k \Pr(S_i(X_i) \leq a/k,\ S_j(X_j) \leq b/k) \to 0$ for $i \neq j$ and all $a, b > 0$ (Resnick, 1987; de Haan & Ferreira, 2006).

Proposition 4 (Multiplicative changes and tail dependence)**.** *(i) Under the von Mises condition $zS_j'(z)/S_j(z) \to -1/\gamma_j$ for a tail of index γ_j, a multiplicative change of relative size η adds, in the limit of Theorem 2(b), log-tail noise of width $w_j = \gamma_j^{-1} \log \frac{1+\eta/2}{1-\eta/2}$. (ii) Under tail dependence the release laws converge weakly to those of a Poisson process whose intensity is the records' exponent measure, with or without the candidate's point s.*

Proof of Theorem 2 and Proposition 4. (a) e_j is a decreasing bijection on the support, so observing the release is equivalent to observing it on the tail scale, and $U_i = S_j(X_{ij})$ are i.i.d. uniform whatever F_j is: without noise the nonmember observation is a function of the r smallest of $kU_1, \dots, kU_k$, and the member observation the same function of the r smallest of $kU_1, \dots, kU_{k-1}$ and s_j. (b) The r smallest of k scaled uniforms have density $\frac{k!}{(k-r)!\,k^r}(1 - e_r/k)^{k-r}$ on $0 < e_1 < \cdots < e_r < k$, which converges pointwise to the Poisson density e^{-e_r}; by Scheffé's lemma the laws converge in total variation, and expanding the logarithm gives $O(1/k)$ for fixed r (for $r = 1$ the nonmember distance is $0.271/k$ and the member distance $0.231/k$ at $s = 1$, computed exactly). The member law is the same with $k-1$ uniforms plus the fixed point s_j, and total variation is preserved by the common measurable map ψ, so the whole experiment converges in total variation: every trade-off curve converges uniformly (van der Vaart, 1998, Ch. 9). (c) For channels independent within records, the joint laws are products, and total variation of a product is at most the sum over the finitely many channels. Under the von Mises condition $zS_j'(z)/S_j(z) \to -1/\gamma_j$, $S_j(z(1+\eta v))/S_j(z) \to (1+\eta v)^{-1/\gamma_j}$ with convergent densities, so a multiplicative change adds $N_j = -\gamma_j^{-1}\log(1+\eta V_j)$ to the log tail position, supported on an interval of width w_j; with V_j having a density, Scheffé again gives total-variation convergence. Storage rounding at relative spacing ρ acts the same way with $\eta = \rho$. For Proposition 4(ii), the records' vectors of tail positions converge as a point process on $[0,\infty]^J \setminus \{\infty\}$ to a Poisson process whose intensity is the exponent measure ν, $k\Pr(k(1-F(X)) \in \cdot) \to \nu$ (Resnick, 1987); the candidate adds the fixed point s, and continuous mapping gives weak convergence of both release laws. Weak convergence alone does not yield convergence of the experiment, which is why Proposition 4(ii) is stated for the release laws only; asymptotic independence places ν on the axes, recovering (c).

Finite k for the maximum. By Theorem 3(i) the likelihood ratio on the consistent region is $c(y) = f_{k-1}(y)/f_k(y)$; for one channel it is $(k-1)/(kF(y)) = (k-1)/(k-e)$ on the tail scale, so for independent channels $\log c(Y) = \sum_j \log\frac{k-1}{k-E_j} \approx \sum_j (E_j - 1)/k$, with standard deviation $\sqrt{d}/k$ under the nonmember law (0.091 simulated against 0.088 for $d = 2{,}044$, $k = 512$). Candidates scored against one release share $c(y)$; only the fixed-target experiment sees it.

Corollary 2 (The maximum in the Poisson limit)**.** *In the limit of Theorem 2(b)–(c) without change, the best test of the hypothesis pair for a candidate has power* $\min(1, \delta_x + \alpha)$, $\delta_x = 1 - \prod_j(1 - e^{-s_j}) \approx 1 - e^{-\lambda_x}$*: the trade-off curve of* $(0, \delta_x)$*-DP for this pair.*

Proof. In the limit experiment, on one channel the nonmember release has tail position $E \sim \mathrm{Exp}(1)$ and the member release $\min(s, E)$: the member law equals the nonmember law on $(0, s)$, has an atom of mass e^{-s} at s, and no mass above s. The likelihood ratio is therefore ∞ at s, 1 on $(0, s)$ and 0 above, and the best test has power $\min(1, \alpha + e^{-s})$, the trade-off of $(0, e^{-s})$-DP. Across independent channels the atom set (owning some channel) has nonmember mass 0 and member mass $\delta_x = 1 - \prod_j(1 - e^{-s_j})$; off it the likelihood ratio is a product of zeros and ones, equal to one exactly when the candidate is excluded nowhere, an event of probability $\prod_j(1 - e^{-s_j}) = 1 - \delta_x$ under both hypotheses. Hence power $\min(1, \delta_x + \alpha)$. This is a property of the pair of limiting laws for this candidate, not a differential-privacy guarantee over neighbouring datasets. Since $e^{-s_j} \to F_j(x_j)^{k-1}$, $\sum_j e^{-s_j}$ is the exposure λ_x of Proposition 3.

Per-channel likelihood ratios. Let $\phi(u) = e^{u - e^u}$ and $\Phi_L(u) = 1 - \exp(-e^u)$ be the density and distribution of $\log E$, and q the density of the log-tail noise N. The observation is $\ell = \log e_j(T_j)$, with

$$p_0(\ell) = \int \phi(\ell - n) q(n)\, dn, \qquad p_1(\ell) = \int_{\ell - n < \log s} \phi(\ell - n) q(n)\, dn + e^{-s} q(\ell - \log s).$$

For N uniform on $[-w/2, w/2]$ (a uniform or hash dither to first order in η, plus rounding), $p_0(\ell) = [\Phi_L(\ell + \frac{w}{2}) - \Phi_L(\ell - \frac{w}{2})]/w$ and $p_1(\ell) = [\Phi_L(\min(\ell + \frac{w}{2}, \log s)) - \Phi_L(\min(\ell - \frac{w}{2}, \log s))]/w + e^{-s}\mathbf{1}\{|\ell - \log s| \le \frac{w}{2}\}/w$; Gaussian noise is integrated by 24-point Gauss–Hermite quadrature. The exact-LR attack scores a candidate by $\sum_j \log(p_1/p_0)$ over channels with $s_j < 25$ (others carry no information), with tail positions from the Pareto tails fitted on public references. Small w recovers Corollary 1: the atom spreads into a density e^{-s}/w that the Neyman–Pearson region captures at rate κ.

Top-2 mixing and the signed range. With $v(e) = Ce^{-\gamma}$ on a Pareto tail, top-2 mixing releases $t = (1-\eta)v(e_1)+\eta v(e_2)$ from the two smallest tail positions. With $\tau = e(t)$, $a = \tau(1-\eta)^{1/\gamma}$ and $e_2(e_1)$ the solution of the release equation, $p_0(t) = \int_a^\tau g(e_1)\,de_1$ with $g = e^{-e_2}e_2^{1+\gamma}/(\gamma C\eta)$, and the member density sums three cases: the candidate owns the maximum, $e^{-e_2(s)}e_2(s)^{1+\gamma}/(\gamma C\eta)$ for $a < s \le \tau$; it is the runner-up, $e^{-s}{e_1^*}^{1+\gamma}/(\gamma C(1-\eta))$ for $s > \tau$, with $e_1^* = ((\tau^{-\gamma}-\eta s^{-\gamma})/(1-\eta))^{-1/\gamma}$; or it is below both, $\int_{\max(a,e_1^*)}^\tau g$. The stock signed range $r = H + M$ adds the largest record maximum H and the largest negated minimum M, two asymptotically independent tails. With $f(v) = e^{-e(v)}e(v)/(\gamma v)$ and $f^<(v) = f(v)\mathbf{1}\{v > x\}$, $p_0(r) = \int f_H(h)f_M(r-h)\,dh$ and $p_1(r) = e^{-s_H-s_M}\delta(r-x_H-x_M)+e^{-s_H}f_M^<(r-x_H)+e^{-s_M}f_H^<(r-x_M)+\int f_H^<(h)f_M^<(r-h)\,dh$: the atom, owning both endpoints, is the rank loss of Theorem 1, and owning one endpoint gives only a shifted density. All integrals are one-dimensional, and each case carries its predicted member mass under H_0 in simulation.

Proof of Theorem 3. (i) Let M_j be the channelwise maximum of $k-1$ i.i.d. records, with any joint law that has a density. If the candidate is a member and its value lies below the release y on every channel, the release is the maximum of the other $k-1$ records, so $P_1(dy \cap \{x < y\}) = f_{k-1}(y)\,dy\,\mathbf{1}\{x < y\}$, while $P_0(dy) = f_k(y)\,dy$; the ratio $c(y) = f_{k-1}(y)/f_k(y)$ does not involve x. If $x_j > y_j$ for some j the member probability is zero, and $x_j = y_j$ has nonmember probability zero. Expected ownership d/k is exchangeability: each record is equally likely to own each channel. (ii) For $N(\mu,\Sigma)$ records the released mean is exactly $(x + S_{k-1})/k$ for a member and S_k/k for a nonmember, with $S_n \sim N(n\mu, n\Sigma)$, giving $N(\mu+(x-\mu)/k,\ (k-1)\Sigma/k^2)$ and $N(\mu,\Sigma/k)$. They differ by the shift $(x-\mu)/k$ and the covariance factor $(k-1)/k$; the latter alone moves the Hellinger distance by $O(\sqrt{d}/k)$, and in whitened coordinates the shift has squared length $(x-\mu)^\top\Sigma^{-1}(x-\mu)/k$, whose expectation over $X \sim N(\mu,\Sigma)$ is $\mathrm{tr}(I)/k = d/k$ (Azize & Basu, 2025).

Plug-in covariance. With the sample covariance of n Gaussian references, the plug-in attack has strength $\mu_x^2(1-d/n)$ to first order. To see this, whiten so that $\Sigma = I$ and let $\widehat{\Sigma} = W/m$ with $W \sim W_d(m, I)$ independent of the candidate y and of the release. The score $y^\top\widehat{\Sigma}^{-1}(T-\mu)$ has member-minus-nonmember mean $y^\top\widehat{\Sigma}^{-1}y/k$ and variance $y^\top\widehat{\Sigma}^{-2}y/k$ (up to $(k-1)/k$). With $\mathbb{E}\,W^{-1} = I/(m-d-1)$ and $\mathbb{E}\,W^{-2} = (m-1)I/((m-d)(m-d-1)(m-d-3))$ (Haff, 1979), the squared mean signal over the mean noise variance is $\frac{\|y\|^2}{k}\cdot\frac{(m-d)(m-d-3)}{(m-1)(m-d-1)}$, which is $\mu_x^2(1-d/n)$ to first order as $d, n \to \infty$; both quadratic forms concentrate there, so the realized ratio converges to the same value.

C Recovering the statistic

LLM Compressor. The stock SmoothQuant modifier with the W8A8 round-to-nearest exporter uses the channel range $R_j = \max a_j - \min a_j$ and weight factor $2B_j$, so its unclamped scale at $\alpha = 0.5$ is $s_j = \sqrt{R_j/(2B_j)}$ and recovery uses $R_j = 2B_j(\gamma_j/\widetilde{\gamma}_j)^2$. With h_{ij} and ℓ_{ij} record i's largest and smallest activations, removing record i changes the range by $(h_{ij} - \max_{r\neq i} h_{rj})_+ + (\min_{r\neq i}\ell_{rj} - \ell_{ij})_+$, so only an endpoint's owner has any influence. The range attack inserts the candidate into bootstrap backgrounds of $k-1$ reference records and compares the release with the distribution of releases that include the candidate.

Unknown per-site multiplier. If an exporter also fuses one positive multiplier per site, $r_j = s_j q_s$, the recovered statistic is $q_s^{1/\alpha}A_j$; a site-normalized score cancels q_s exactly. On the OPT-125M capture 93.6% of members keep a positive projective codimension $\sum_s (r_{is}-1)_+$ (median 15), against 100% owning at least one maximum in known units.

D PTQ experiments

Reference SmoothQuant. The pinned implementation's smoothing and simulated W8A8 quantization ($\alpha = 0.5$, per-tensor weights and activations) run unchanged on three 512-record OPT-125M draws; the state dict is saved and attacked from a separate process. With $2{,}048$ null-calibration

records and 7,213 held-out nonmembers the site-normalized score reaches AUC 0.994 ± 0.001 and TPR 0.920 ± 0.007 at nominal 1% FPR (realized 0.93%).

Large-null benchmark. The deduplicated Pile validation set (208,954 records) is split by record into 2,048 references, 16,384 null-calibration records, an 8,192-record member pool from which each draw takes 512 members, and 178,234 untouched nonmembers. Per-record per-channel maxima are captured once through the public base; each draw exports a real artifact. At nominal 10^{-4}, the smallest level the null split expresses, the realized held-out rate is 7.3×10^{-5} (21, 17 and 1 false positives over the draws). TPR at zero observed false positives is 0.41, with a one-sided 95% bound of 1.7×10^{-5} on the corresponding FPR; output-loss, channel-shuffled and excluded-pool controls give AUC 0.49–0.52.

Table 6: Large-null benchmark (mean over three draws; FP counts on 178,234 held-out nonmembers).

Gains	Score	AUC	TPR@10^{-2} (FP/N)	TPR@10^{-3} (FP/N)	TPR@10^{-4} (FP/N)
fp16	matching	0.8341	0.443 (1639/178234)	0.310 (176/178234)	0.198 (8/178234)
fp16	site-normalized	0.9949	0.923 (1549/178234)	0.840 (145/178234)	0.677 (14/178234)
fp32	matching	0.9095	0.598 (1572/178234)	0.459 (134/178234)	0.380 (29/178234)
fp32	site-normalized	0.9947	0.919 (1509/178234)	0.843 (160/178234)	0.679 (13/178234)

NVIDIA Model Optimizer. Model Optimizer 0.47.0 runs as documented: its INT8 SmoothQuant preset with the default $\alpha = 1$, then its Hugging Face export, which writes the per-channel W8A8 checkpoint that TensorRT-LLM and vLLM load. Every smoothed linear layer stores INT8 weights and $\texttt{pre_quant_scale}_j = 1/A_j$ with $A_j = \max_t |a_{tj}|$, so recovery is a reciprocal (median relative error 2.9×10^{-4} in float16, 1.7×10^{-4} in float32). Members, null-calibration records, held-out nonmembers and the attack are those of the large-null benchmark; controls stay at AUC 0.49–0.51.

Table 7: Stock NVIDIA Model Optimizer INT8 SmoothQuant exports of OPT-125M (three draws, mean; FP is the mean false-positive count among 178,234 held-out nonmembers at each nominal level; thresholds from 16,384 null-calibration records).

Export dtype	Score	AUC	TPR@10^{-2} (FP)	TPR@10^{-3} (FP)	TPR@10^{-4} (FP)	TPR@0 obs. FP
float16	site-normalized	0.9948	0.923 (1578)	0.844 (174)	0.688 (17)	0.418
float16	matching	0.8335	0.395 (1584)	0.222 (147)	0.153 (26)	0.104
float32	site-normalized	0.9947	0.919 (1473)	0.844 (165)	0.683 (13)	0.415
float32	matching	0.8368	0.455 (1517)	0.313 (146)	0.201 (9)	0.156

LLM Compressor on Qwen2.5-7B. Attacks were selected on Qwen2.5-1.5B exports and applied unchanged to three new 512-record draws of Qwen2.5-7B, exported through the same stock path with three calibration statistics: the stock signed range, the absolute maximum, and a top-16 mean of per-record maxima, the last two by replacing only the collected statistic before the unchanged smoothing. Reloaded W8A8 artifacts have WikiText-2 perplexity 8.54 ± 0.16 (range), 8.69 ± 0.18 (maximum) and 8.70 ± 0.16 (top-16 mean). A second evaluation scores the same artifacts on 16,384 fresh Pile validation nonmembers with thresholds set on 4,096 null-calibration records (Table 9).

Intel ONNX SmoothQuant with ONNX Runtime INT8. Intel's calibrator fuses the 99.999th percentile of $|a|$ pooled over all calibration tokens at each weighted MatMul/Gemm input (at 2.6×10^5 tokens, between the third- and fourth-largest values), and ONNX Runtime static quantization writes INT8 operators. We replace only the stock calibrator's materialization of every activation by a streaming per-channel top-16, which reproduces the stock statistic bitwise on all 49 tensors. The statistic is recovered from the saved model to 1.9×10^{-7} on OPT-125M and 7.3×10^{-7} on Llama-3.2-1B, where a Gaussian likelihood of the release conditioned on the candidate (the composite likelihood) reaches AUC 0.940 ± 0.002 and 0.869 ± 0.010 (Tables 10 and 11; the Llama attack is copied from the OPT selection).

Table 8: 7B comparison with attacks selected before evaluation, each attack applied to every arm (three draws; entries are mean ± SD across draws). The matching attack looks for exact matches; the ranking attack uses the full release distribution.

Attack (same for every arm)	Calibration statistic	AUC	TPR@1%	TPR at zero obs. FP
matching ($\tau = 0.01$)	absolute max	0.863 ± 0.016	0.234 ± 0.012	0.081 ± 0.014
	top-16 mean	0.511 ± 0.015	0.017 ± 0.008	0.001 ± 0.001
	signed range	0.466 ± 0.010	0.002 ± 0.002	0.001 ± 0.001
ranking (composite likelihood)	absolute max	0.737 ± 0.028	0.089 ± 0.016	0.009 ± 0.006
	top-16 mean	0.824 ± 0.043	0.109 ± 0.056	0.016 ± 0.006
	signed range	0.667 ± 0.027	0.058 ± 0.025	0.009 ± 0.006

Table 9: 7B arms on 16,384 fresh nonmembers (three draws; thresholds set at nominal levels on a disjoint 4,096-record null pool; realized FPR is the held-out mean; FP: held-out false positives per draw).

Calibration statistic	AUC	nominal FPR	TPR (mean ± SD)	per-draw TPR 95% Wilson	realized FPR	FP per draw
absolute max	0.861 ± 0.015	10^{-3}	0.120 ± 0.008	$[0.089, 0.161]$	8.7×10^{-4}	$18, 18, 7$
		2.4×10^{-4}	0.082 ± 0.021	$[0.048, 0.135]$	2.4×10^{-4}	$3, 5, 4$
signed range (stock)	0.772 ± 0.010	10^{-3}	0.048 ± 0.011	$[0.024, 0.082]$	1.4×10^{-3}	$30, 24, 16$
		2.4×10^{-4}	0.031 ± 0.016	$[0.007, 0.067]$	5.7×10^{-4}	$13, 5, 10$
top-16 mean	0.824 ± 0.040	10^{-3}	0.023 ± 0.005	$[0.009, 0.045]$	1.1×10^{-3}	$25, 21, 7$
		2.4×10^{-4}	0.007 ± 0.003	$[0.001, 0.023]$	1.8×10^{-4}	$2, 5, 2$

Table 10: Third stack, memory-only streaming calibrator (bitwise-identical statistic), $k = 512$ (three test draws; 512 members, 1,000 nonmembers).

Arm	score	AUC	TPR@1%	TPR@0%	recovery ε (p50)
folded into LayerNormalization	shadow_ll	0.940 ± 0.002	0.396 ± 0.079	0.159 ± 0.113	$1.9e-07$
explicit smooth-scale Mul	shadow_ll	0.940 ± 0.002	0.396 ± 0.079	0.159 ± 0.113	$1.9e-07$

Table 11: Llama-3.2-1B replication of the third stack at $k = 512$ (three test draws; 512 members, 1,000 nonmembers; attack fixed from the OPT development draw; mean ± draw SD).

Base	score	AUC	TPR@1%	TPR@0.1%	TPR@0%
Llama-3.2-1B	shadow_ll	0.869 ± 0.010	0.219 ± 0.057	0.098 ± 0.054	0.047 ± 0.019

E RELEASE-GEOMETRY EXPERIMENTS

Maximum, top-m means and signed range. One OPT-125M capture of Pile-10k records stores each record's per-channel minimum and maximum. The attacks, an ownership count and the composite likelihood over bootstrap backgrounds of reference records, were tuned on development records. Only the released statistic changes across rows (Table 12): the signed range is a deterministic function of the separately released endpoints yet keeps 0.17 of their 0.75 TPR at zero observed FP, because only 1.8% of members own both endpoints anywhere.

Table 12: Released functionals compared on one capture, record pool, and attack family (three draws; 512 members, 5,165 nonmembers). TPR@0% is separation in the evaluated sample.

Released functional	AUC	TPR@1%	TPR@0%
$\max_i \lvert a_{ij} \rvert$ (SmoothQuant)	0.974 ± 0.008	0.737	0.543 ± 0.134
$\max_i h_{ij} - \min_i l_{ij}$ (LLM Compressor)	0.935 ± 0.005	0.456	0.174 ± 0.049
$(\max_i h_{ij},\ \min_i l_{ij})$ released apart	0.988 ± 0.006	0.858	0.748 ± 0.053

Ownership and exposure. Mean ownership is 0.99, 3.95 and 15.8 at $d/k = 1, 4, 16$ for every k (Fig. 3), and coverage at $d/k = 1$ is 0.55–0.58 against 0.63 for independent owners, above the Paley–Zygmund bound of Proposition 3(ii) in every cell. One attack with a fixed matching tolerance, applied to every cell, has AUC 0.87–0.95 across k at $d/k = 4$; at fixed d/k its TPR falls with d, because the fixed tolerance admits coincidental matches at a rate that grows with d (Prop. 2). For exposure prediction, λ_x is computed from 2,048 public references with the posterior mean $\prod_{i<k-1}(r+1+i)/(n+2+i)$ of $F_j(m_j(x))^{k-1}$ given the strict reference rank r of n; it predicts realized ownership with Spearman 0.937, and detection at the 1% operating point rises from 0.09 in the least-exposed quintile to 0.99–1.0 in the top three.

The Poisson model predicts real attacks. Tail positions come from per-channel Pareto tails fitted on public references; the exact-LR attack sums per-channel log likelihood ratios over channels with $s_j < 25$. Table 13 compares the attack on the real releases, its prediction from the Poisson limit (each candidate's own tail positions, simulated releases, no labels) and our earlier full attack: predictions are within 0.08 at every operating point. For top-m means the likelihood ratio depends only on (γ, m, s, t) and is tabulated per tail-index bin; Table 14 shows the prediction on the benchmark and on a copy with every channel permuted independently across records, which keeps each channel's law and removes cross-channel dependence. The product approximation overstates top-m power below 1% FPR on real data and holds at every level on the permuted copy; the dependence behind this is small in the extremes (a pair of channels reaches the top $1/k$ together 1.3 times as often as under independence) and larger in the bulk (mean $|\rho| = 0.16$). On the Qwen2.5-1.5B export the exact-LR attack on the stock signed range reaches AUC 0.883, and the exclusion law predicts never-excluded fractions of 1.7, 3.1 and 26.9% at $D = 0, 1, 10\%$ (observed 2.2, 4.0, 30.9%). For the mean, across $d \in \{512, 1{,}024, 2{,}044\}$ channels and $n \in \{3{,}072, 6{,}144\}$ references, the prediction $\mu^2 = (d/k)(1 - d/n)$, with no fitted constant, matches the realized AUC to within 0.03 and TPR at 10^{-3} to within 0.01. Table 15 lists the numbers behind Figure 4.

Table 13: Likelihood-ratio attack of Theorem 2 ("real"), its prediction from the Poisson limit ("pred."), and the earlier full attack ("full") on the OPT-125M benchmark (2,044 channels, $k = 512$, three draws, 178,234 held-out nonmembers).

Release	D	AUC			TPR@10^{-2}			TPR@10^{-3}			TPR@10^{-4}		
		real	pred.	full	real	pred.	full	real	pred.	full	real	pred.	full
maximum	–	0.995	0.993	0.983	0.904	0.884	0.773	0.850	0.830	0.688	0.751	0.744	0.643
uniform noise	0.3%	0.991	0.989	0.992	0.819	0.798	0.826	0.662	0.609	0.672	0.490	0.447	0.498
uniform noise	1%	0.979	0.970	0.979	0.643	0.598	0.645	0.409	0.358	0.380	0.205	0.207	0.192
uniform noise	3%	0.909	0.906	0.870	0.296	0.275	0.046	0.107	0.113	0.006	0.025	0.049	0.001
uniform noise	10%	0.697	0.736	0.523	0.080	0.089	0.012	0.019	0.026	0.000	0.003	0.014	0.000
hash dither	1%	0.978	0.971	0.979	0.643	0.594	0.645	0.399	0.339	0.410	0.208	0.199	0.213
hash dither	10%	0.698	0.729	0.516	0.072	0.090	0.014	0.014	0.031	0.001	0.003	0.010	0.000
Gaussian noise	0.3%	0.990	0.985	0.990	0.811	0.751	0.809	0.643	0.570	0.641	0.467	0.422	0.465
Gaussian noise	1%	0.972	0.962	0.972	0.592	0.537	0.583	0.358	0.307	0.354	0.170	0.176	0.159
Gaussian noise	3%	0.884	0.875	0.751	0.255	0.228	0.018	0.074	0.097	0.003	0.012	0.041	0.001
top-2 mixing	0.1%	0.995	0.992	0.995	0.906	0.883	0.903	0.814	0.775	0.814	0.708	0.688	0.673
top-2 mixing	0.3%	0.993	0.990	0.993	0.886	0.862	0.880	0.755	0.706	0.742	0.622	0.560	0.610
top-2 mixing	1%	0.987	0.979	0.982	0.750	0.692	0.715	0.535	0.463	0.502	0.364	0.292	0.331

Table 14: Top-m means on the OPT-125M benchmark: tabulated likelihood-ratio attack (real) and its prediction from the Poisson limit, on the benchmark and with each channel permuted independently across records ($k = 512$, three draws, 178,234 held-out nonmembers).

Data	m	AUC		TPR@10^{-2}		TPR@10^{-3}		TPR@10^{-4}	
		real	pred.	real	pred.	real	pred.	real	pred.
benchmark	2	0.918	0.908	0.318	0.302	0.043	0.123	0.008	0.048
benchmark	4	0.885	0.881	0.230	0.259	0.016	0.096	0.005	0.032
benchmark	8	0.861	0.871	0.201	0.266	0.018	0.105	0.007	0.038
benchmark	16	0.847	0.876	0.175	0.273	0.020	0.103	0.004	0.038
channels permuted	4	0.921	0.911	0.301	0.243	0.073	0.083	0.012	0.018
channels permuted	16	0.893	0.913	0.289	0.309	0.096	0.094	0.028	0.030

Table 15: Numbers behind Figure 4: releases of the same OPT-125M calibration draws ($d/k = 4$; three draws, 178,234 held-out nonmembers), each attacked by its exact likelihood ratio and compared with the theory's prediction.

Release	Shape	Attack	AUC		TPR@10^{-2}		TPR@10^{-3}		TPR@10^{-4}	
			real	pred.	real	pred.	real	pred.	real	pred.
maximum	exact matches	exact LR (Thm. 2)	0.995	0.993	0.904	0.884	0.850	0.830	0.751	0.744
top-2 mean	finite jumps	exact LR (App. E)	0.952	0.940	0.432	0.380	0.154	0.156	0.021	0.074
mean	smooth shift	Gaussian LR, $n = 6{,}144$	0.858	0.876	0.240	0.244	0.076	0.072	0.013	0.018

F PROBABILITY CALIBRATION

A TF-IDF linear SVM is trained on the 20 Newsgroups training split with headers, footers and quotes removed. The test split, deduplicated by exact score vector (7,284 records), supplies three draws of $k = 1{,}024$ members; the other records are nonmembers. scikit-learn fits isotonic, sigmoid or temperature calibration of the fixed classifier on the same members, and only the fitted calibrators are released. The certificate attack counts classes whose isotonic knots contain the candidate's score; the loss attack scores the calibrated log-probability of the true label. Isotonic keeps 21.0 ± 0.9 knots per class; members own 0.41 knots on average, 34.1% own at least one and no nonmember does (a 95% bound of 4.8×10^{-4} on the nonmember match rate). The nearest nonmember score lies 2.2–7.0×10^{-7} (relative) from a knot, so matching tolerates recomputation error below that spacing. The CIFAR-100 replication uses the public ResNet-56 (test accuracy 0.726) with $k = 1{,}024$ and $5{,}000$: exact matching certifies 0.655 ± 0.009 and 0.379 ± 0.007 of members with no nonmember match, and with logits recomputed at batch size 1 (relative difference up to 8.9×10^{-5}) a fixed tolerance of 10^{-6} still certifies 0.364 and 0.179. Table 16 compares these rates with $1 - e^{-K/k}$ for K knots (Theorem 3(i)); the smooth calibrators stay at chance under both attacks.

Table 16: Scale law for isotonic calibration (three draws per setting).

Setting	knots/class	exposure K/k (pred./obs.)	certified $1 - e^{-K/k}$	certified (obs.)
20NG SVM k =1024	21.0	0.410 / 0.410	0.336	0.341
CIFAR-100 ResNet-56 k =1024	10.0	0.973 / –	0.622	0.655
CIFAR-100 ResNet-56 k =5000	22.8	0.455 / –	0.366	0.379

G SMALL CHANGES TO THE MAXIMUM

Setting and assumptions. Start from the maximum $Z_{[1]}$ and add a change ηG, where G is a function of the calibration set (for top-2 mixing, $G = Z_{[2]} - Z_{[1]}$) or independent noise, with RMS distortion $D = \eta(\mathbb{E}_0 G^2)^{1/2}$ and $\kappa = \|q_x\|_\infty (\mathbb{E}_0 G^2)^{1/2}$ for the density q_x defined below. Records' scalar summaries on one coordinate are i.i.d. with density f and CDF F under H_0; under H_1 one record equals the target x. $O = \{x > \max_{i \geq 2} X_i\}$, $p_x = F(x)^{k-1}$, and on O the release is $x + \eta Q_x$. (A1) f continuous and positive at x. (A2) Q_x has a density q_x. (A3) The null density $h_{0,\eta}$ of T_η converges to $h_0 = kfF^{k-1}$ uniformly near x. (A4) On O^c the release density r_η satisfies $r_\eta / h_{0,\eta} \leq L$.

Theorem 4 (Small perturbations, general form)**.** *Under (A1)–(A4), as $\eta \to 0$ and $\alpha/(\eta h_0(x)) \to s$, with $h_0 = kfF^{k-1}$ the null density of the maximum, the most powerful fixed-target test satisfies $\beta^\star(\alpha) \to p_x \, \mathcal{C}_{Q_x}(s)$, where $\mathcal{C}_Q(s) = \sup_{\mathrm{Leb}(A) \leq s} \Pr(Q \in A)$ is the concentration profile of Q. For bounded q_x and $\alpha/\eta \to 0$ this gives Corollary 1.*

Proof of Theorem 4. Write $h_{1,\eta} = p_x g_\eta + (1 - p_x) r_\eta$ with $g_\eta(t) = \eta^{-1} q_x((t-x)/\eta)$. For a test φ with $\int \varphi h_{0,\eta} \leq \alpha$: by (A4) $(1 - p_x) \int \varphi r_\eta \leq L\alpha$. On $N_\delta = \{|t - x| < \delta\}$, $\int_{N_\delta} \varphi \leq \alpha / \inf_{N_\delta} h_{0,\eta}$, and substituting $t = x + \eta u$ gives $\int_{N_\delta} \varphi g_\eta \leq \mathcal{C}_{Q_x}\big(\alpha/(\eta \inf_{N_\delta} h_{0,\eta})\big)$ (bathtub principle); outside N_δ the owned branch contributes at most $\Pr(|Q_x| \geq \delta/\eta) \to 0$. For the lower bound take $\varphi = \mathbf{1}\{(t - x)/\eta \in B\}$ with B a superlevel set of q_x of measure $\alpha/(\eta \sup_{N_\delta} h_{0,\eta})$. (A3), tightness,

continuity of $\mathcal{C}$ and $\delta \downarrow 0$ give the limit. The κ form follows from $\mathcal{C}_Q(s) = s\|q\|_\infty(1 + o(s))$ as $s \to 0$, $\eta = D/(\mathbb{E}_0 G^2)^{1/2}$, and $p_x/h_0(x) = 1/(kf(x))$. The $O(\alpha)$ remainder is the bulk evidence; the law is informative when $\kappa/(Dkf(x)) > L$, and certificates are operationally removed at $D_x^* = \kappa/(Lkf(x))$ ($L \approx 1.6$ on real OPT channels). The floor: a density bounded by M covers Lebesgue measure at least $1/M$, so its variance is at least $1/(12M^2)$ with equality for a uniform on an interval.

Relative units. For $T = Z_{[1]}(1 + \eta G)$ write the owned release as $x(1 + \eta Q_x)$; the proof above with $t = x(1 + \eta u)$ gives $\beta^\star(\alpha) = \alpha\,\kappa/(D\,x\,kf(x))\,(1 + o(1)) + O(\alpha)$ with D the relative RMS distortion and $\kappa = \|q_x\|_\infty(\mathbb{E}_0 G^2)^{1/2}$ in relative units. For the ideal dither $G = U - \frac{1}{2}$ with $U \sim$ Uniform$(0, 1)$ independent of the maximum, $\kappa = 1/\sqrt{12}$ exactly.

Comparing modifications. In exact Neyman–Pearson computations on Gaussian, Pareto(3) and Beta(2,2) tails ($k = 512$), κ is $1.34/3.05/1.13$ for top-2 mixing, 0.40 for Gaussian noise and $1/\sqrt{12} = 0.289$ for uniform noise, and exact power matches the prediction with median ratio 0.96–1.02; on eleven real OPT-125M channels the median ratio is 1.01 (IQR 0.88–1.10).

The hash dither. As an implementation of a uniform change without secret noise, U is a public-key universal hash of the nuisance maxima of ranks 3–16, truncated to 53 bits (Carter & Wegman, 1979). When those maxima carry enough min-entropy, the leftover hash lemma (Impagliazzo et al., 1989; Håstad et al., 1999) makes each owned offset nearly uniform; empirically the offsets are uniform (Kolmogorov–Smirnov distance ≤ 0.004), and the hash dither matches secret uniform noise to within 0.01 at every operating point (Table 13).

Exclusions. A member never exceeds what the release allows, so values above the release rule a candidate out.

Proposition 5 (Exclusions under a dither)**.** *For i.i.d. records and a dither $T_j = Z_{(1),j}(1+\eta V_j)$ with V_j uniform on $[-\frac{1}{2}, \frac{1}{2}]$, a member is never excluded, and a nonmember with values x is excluded on channel j with probability exactly $p_j(x) = \mathbb{E}_V\, F_j\big(x_j \frac{1-\eta/2}{1+\eta V}\big)^k$, whatever the dependence across channels. For a Pareto tail of index γ_j and a random record, $\mathbb{E}\, p_j \to 1/k$ as $\eta \to 0$ and $\mathbb{E}\, p_j \approx \gamma_j(1 - \eta/2)/(\eta k)$ once $\eta \gg \gamma_j$: exclusions then decay only like $1/\eta$.*

Proof. A member satisfies $x_j \leq Z_{(1),j} \leq T_j/(1 - \eta/2)$. For a nonmember, exclusion means $Z_{(1),j}(1 + \eta V) < x_j(1 - \eta/2)$, i.e. $Z_{(1),j} < x_j\rho(V)$ with $\rho(V) = (1 - \eta/2)/(1 + \eta V)$, of probability $F_j(x_j\rho(V))^k$ given V; channels are independent in the limit of Theorem 2. For a Pareto tail, $F_j(x\rho)^k \to \exp(-s\,\rho^{-1/\gamma})$, and a random record's tail position has density $1/k$, so its expected number of exclusions is $k^{-1}\sum_j c_j(\eta)$ with

$$c_j(\eta) = \mathbb{E}_V\, \rho(V)^{1/\gamma_j} = \frac{(1 - \eta/2) - (1 + \eta/2)\, r_\eta^{1/\gamma_j}}{\eta\,(1/\gamma_j - 1)}, \qquad r_\eta = \frac{1 - \eta/2}{1 + \eta/2},$$

which tends to 1 as $\eta \to 0$ (d/k exclusions for the maximum) and behaves like $\gamma_j(1 - \eta/2)/\eta$ for light tails. Exclusions that hold for every dither value have probability of order $r_\eta^{1/\gamma}$, which decays exponentially in η/γ; averaging over the dither gives the much slower $1/\eta$ decay.

H Surveys and robustness

Framework survey. Table 18 records, from source, where the statistic of Table 1 lands in each stack (NVIDIA Corporation, 2026a; Intel, 2026; torchao maintainers and contributors, 2026; ONNX Runtime developers, 2026; Red Hat AI and vLLM Project, 2026a; Lin et al., 2024); the AWQ paths were audited at llm-awq commit `d6e797a42b9e` and AutoAWQ commit `88e4c76b2075`. ONNX Runtime's MinMax calibrator without averaging is a per-tensor selector; its histogram calibrators are not.

Table 17: Production repair arms: stock-path LLM Compressor exports of Qwen2.5-1.5B with fp32 gain storage (three draws unless noted; WikiText-2 perplexity on 298,785 tokens; 8,180 scored channels; 512 members, 2,000 nonmembers). Exclusions: channels on which a nonmember exceeds the largest value the release allows (members: 0 in every arm).

Arm	PPL	cert. TPR@0.1% (spike)	AUC	TPR@1%	TPR@0.1% full attack	TPR@0 obs. FP	exclusions per nonmember	never excl.
max	10.609	0.597	1.000	0.985	0.965	0.682	15.9	0.022
top-2 mix ($\eta = 0.19$)	10.610	0.011	0.997	0.923	0.525	0.296	0.5	0.779
spacing-ratio dither	10.603	0.012	0.998	0.958	0.822	0.539	12.1	0.038
hash dither, $D = 1\%$	10.610	0.012	0.998	0.945	0.848	0.577	12.0	0.040
hash dither, $D = 10\%$	10.600	0.012	0.703	0.053	0.013	0.007	2.4	0.309
stock signed range	10.613	–	–	–	–	–	–	–

Table 18: Where each stack stores its calibration statistic, from source.

Stack	Where the fused statistic lands in the released artifact
MIT SmoothQuant (reference)	norm gain (and LayerNorm bias) divided by s_j; next linear columns multiplied
LLM Compressor	norm gain (and bias) divided by s_j; balance-layer columns multiplied
NVIDIA Model Optimizer	explicit `pre_quant_scale` tensor; inverse folded into weight columns
Intel Neural Compressor (PyTorch)	preceding norm weight and bias, or an explicit `input_scale` buffer
torchao (prototype SmoothQuant)	`act_pre_scale` $= 1/s_j$ stored on the quantized weight tensor
ONNX Runtime static quantization	scale and zero-point initializers in the graph
Intel ONNX SmoothQuant	explicit $\mathrm{Mul}(1/s_j)$ initializer, or folded into a single-consumer LayerNormalization
AWQ (llm-awq, AutoAWQ)	norm gain divided, linear columns multiplied; weights then packed to 4 bits

Threat-model survey. We read public metadata (model card, configuration, recipe, quantization config and tensor names; no weights) of 2,582 Hugging Face repositories found by name search and by library tag. 642 release a per-channel smoothing scale (a SmoothQuant recipe or an explicit `pre_quant_scale`). Among them 99% state the smoothing strength, the parent checkpoint is public for 56–83% and calibration data are named for 29% (Table 19). First-party guidance recommends calibration data matching deployment or training data (Table 20).

Table 19: Hugging Face Hub survey from public metadata: artifacts whose transform our recovery covers (counts in parentheses).

Condition (public metadata only)	per-channel smoothing scale ($n = 642$)	context: incl. static per-tensor scales ($n = 1192$)
parent named in card or config	57% (367)	71% (843)
and public, ungated	33% (215)	56% (671)
and public (gated included)	56% (360)	70% (832)
parent named, cached path or repo name	84% (541)	89% (1060)
and public (gated included)	83% (534)	88% (1049)
parent revision pinned	13% (85)	8% (100)
exporter family identifiable	100% (642)	97% (1155)
calibration-dependent transform identifiable	100% (642)	100% (1192)
recovery map determinable	100% (641)	100% (1191)
α stated	99% (633)	53% (633)
calibration data named	29% (185)	43% (508)
tokenizer shipped	100% (639)	99% (1180)

Table 20: First-party guidance on calibration data (verbatim; sources fetched 2026-09-26).

Source (fetched 2026-09-26)	Calibration data guidance (verbatim)	Samples	Transform
LLM Compressor 0.13.0 INT8 W8A8 example (Red Hat AI and vLLM Project, 2026b)	"it is very useful to use calibration data that closely matches the type of data used in deployment. If you have fine-tuned a model, using a sample of your training data is a good idea."	512 (start)	SmoothQuant in the recommended INT8 recipe
NVIDIA ModelOpt docs: PTQ (NVIDIA Corporation, 2026b)	"PTQ can be achieved with simple calibration on a small set of training or evaluation data (typically 128-512 samples)"	128–512	INT8 SmoothQuant among the supported formats
Hugging Face Optimum concept guide (Hugging Face, 2026)	"typically by passing representative data through the model and recording the activation values."	about 200	–

Candidate mismatch. Table 21 changes only the attacker's copy of the candidate. Leakage degrades gracefully: half the calibration tokens keep AUC 0.991 and 80% of the detection at 10^{-3},

and 64 tokens still give AUC 0.90. Preprocessing mismatches (normalized whitespace, a missing prefix, a prepended header) keep AUC at 0.84–0.92 and lower TPR at 10^{-4} to 0.01–0.16.

Table 21: Attacker's copy of the candidate (large-null benchmark; three draws; null thresholds use the same copy).

Attacker's candidate	AUC	TPR@10^{-2}	TPR@10^{-3}	TPR@10^{-4} (FP/16,384)	TPR@0 obs. FP
first 64 tokens	0.903	0.102	0.028	0.012 (1)	0.011
first 256 tokens	0.991	0.888	0.675	0.249 (1)	0.350
exact (512 tokens)	0.995	0.919	0.843	0.679 (1)	0.578
header line prepended	0.916	0.453	0.281	0.156 (2)	0.096

Email calibration data. Calibrating Qwen2.5-1.5B on 512 Enron emails (AESLC) instead of Pile-10k, with attacks selected on Pile-10k exports, leaks at the same level: AUC 0.73–0.74 against 0.76–0.77, and TPR 0.12–0.14 against 0.11 at 1% FPR (Table 22). Certificates are rare in this bfloat16 export on both corpora.

Table 22: Enron emails versus Pile-10k as calibration data (Qwen2.5-1.5B; 2,000 same-corpus non-members; three draws). The Pile-10k range classifier was trained on those draws.

Calibration statistic	calibration corpus	AUC	TPR@1%	TPR@0.1%	TPR@0 obs. FP
stock signed range	Enron emails	0.741	0.136	0.064	0.043
	Pile-10k	0.768	0.107	0.046	0.014
absolute max	Enron emails	0.727	0.117	0.020	0.012
	Pile-10k	0.764	0.110	0.060	0.038