THE PRIVACY SIDE EFFECT OF CHEAP INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reducing the numerical precision of model weights is common in deployment because it lowers computational cost. This reduction also selectively lowers a model's ability to reproduce the training data word-for-word, by several bits, before it substantially affects general capability. This difference is termed the selectivity gap. The gap is measured across a 29-fold range in parameter count and is replicated in a second model family, an error-correcting production quantizer, and text-to-image diffusion, where the gap is more pronounced. Although the gap narrows as models grow, it stays above 24.2 % points of retention across all sizes tested. The content destroyed is not predicted by data duplication, although a magnitude filter, such as quantization, should spare duplicated content. Instead, the sharpness of the loss surface around a memorized sequence predicts its sensitivity to numerical rounding. This extends prior observations linking memorization to sharp loss regions by showing that the same sharpness also predicts which memorized content is lost under quantization. In diffusion models, the selectivity gap persists, but weight-space sharpness is no longer predictive. Instead, the gap is associated with the stability of image generation to perturbations in the starting noise. Finally, decomposing a realistic deployment recipe shows that the 8-bit activation quantizer, rather than the 4-bit weight quantizer, is primarily responsible for the loss of training-image reproduction. Numerical precision is therefore both an efficiency-related deployment choice and a privacy-relevant configuration that can substantially affect a model's ability to reproduce its training data.

1 INTRODUCTION

Quantization is applied to almost every deployed model for reasons unrelated to privacy. It makes models cheaper to serve. This work examines how quantization affects training data memorized by these models and finds an asymmetric effect. Reducing weight precision destroys a model's ability to reproduce its training data verbatim several bits before substantially affecting general capability. There is a window of roughly 6 to 4 bits in which a model loses much of its verbatim recall while essentially retaining all of its measured capability, visible directly in Figure 1(a).

Larger models memorize more, so a separation that shrank with scale would matter less for models used in practice. The separation does shrink, but then stops. From 410M to 2.8B it narrows enough that a linear extrapolation would close it below 10B parameters. From 2.8B onward, however, it does not close. The separation appears at every size and under every capability anchor. It is also present under an error-correcting production quantizer and replicates in a second model family whose pretraining corpus was never deduplicated. The effect reappears more sharply in text-to-image diffusion, where 5-bit weights eliminate replication of memorized training images without measurable changes in prompt fidelity or distribution-level image quality.

The next question is why the separation occurs. Since quantization acts as a magnitude filter, one might expect it to preserve content that received stronger training, such as duplicated training data. After adjusting for competing factors, however, duplication does not predict which memorized sequences are destroyed by quantization, while the curvature of the loss surface around a sequence does. The association between memorized content and sharp minima is already established (Garg et al., 2024; Ravikumar et al., 2024; Merullo et al., 2025). The same sharpness also predicts which memorized content is disrupted by numerical rounding. This holds at four model sizes in the first family and in a second family, where duplication was counted exactly against the model's own

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.

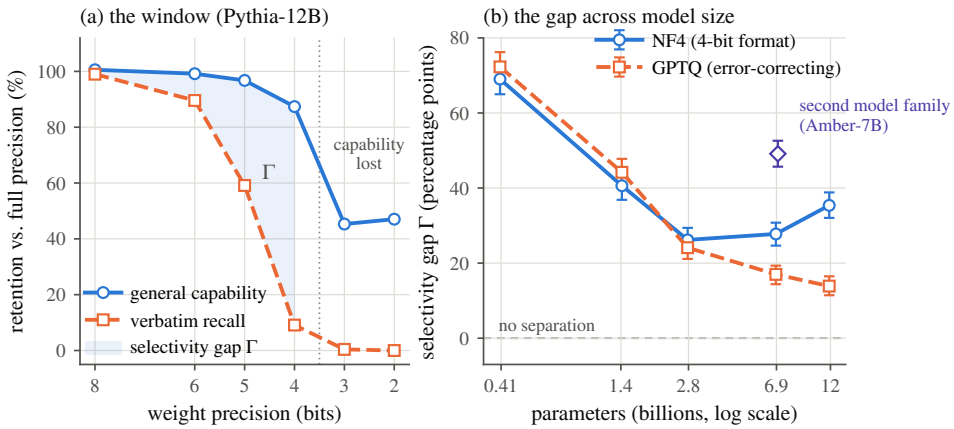


Figure 1: The precision window and how it changes with model scale. (a) On the largest audited model, verbatim reproduction of training data degrades several bits before general capability, leaving a band of roughly 6 to 4 bits in which the model loses much of its memorized content while retaining all of its measured capability essentially. The shaded region between the curves is the *selectivity gap* Γ, defined in Section 2. Below 4 bits, both curves collapse, so the separation no longer provides a useful window, as the model is no longer usable at that precision. (b) The same gap Γ is measured across all model sizes. It narrows from +69.0 to +26.2 points, then stops narrowing, remaining above zero across both quantizer families and in the second model family. Bars show 1,000-resample bootstrap 95% intervals.

training corpus rather than estimated from an index. This choice was fixed in advance to avoid defining the comparison based on its outcome.

The same effect appears in diffusion, but with a different geometry. Weight-space curvature does *not* transfer, and the magnitude effect observed in language models can be ruled out. Instead, the relevant predictor is a property of the sampling trajectory. Memorized images that survive quantization lie within wide attractor basins, which provide explanatory power beyond what memorization alone can offer. Thus, the same selectivity phenomenon appears across two modalities but arises through different mechanisms. This work makes the following four contributions.

1. **The separation does not close with scale.** The selectivity gap between memorization damage and capability damage stays above 24.2 % points of retention under NF4 and above 12.4 under GPTQ, across a 29-fold range in parameter count and three capability anchors. It also replicates in a second model family at $\Gamma = +49.1$ with no capability measure below 0.962 of full precision (Section 3).
2. **A mechanism, tested against a fairly measured competitor.** Loss-landscape curvature predicts which sequences are destroyed at every model size tested and in both families. Weight magnitude is secondary, while training-set duplication is uninformative after adjustment, including where duplication is counted exactly rather than proxied (Section 4).
3. **Transfer to diffusion, with a different mechanism.** Replication of memorized images is eliminated at 5-bit weights at no measurable quality cost. The language-model mechanism does not transfer, while basin stability of the sampling trajectory does (Section 5).
4. **An attribution result for the deployment recipe.** At deployment-grade quality, the 8-bit activation quantizer eliminates image replication, while the 4-bit weights alone do not (Section 5).

2 Setup

Metric. Quantization can damage memorization and general capability at the same time. The quantity of interest is therefore how much *more* it damages memorization than general capability. For a quantizer q, two retentions are measured relative to the same full-precision model in the same environment. Memorization retention $\rho_{\text{mem}}(q)$ is the ratio of verbatim extraction accuracy under q to verbatim extraction accuracy at full precision. Capability retention $\rho_{\text{cap}}(q)$ is defined in the same way using a general benchmark. The *selectivity gap* is their difference, in percentage points,

$$\Gamma(q) \;=\; 100 \cdot \big(\rho_{\mathrm{cap}}(q) - \rho_{\mathrm{mem}}(q)\big). \tag{1}$$

$\Gamma > 0$ means that quantization removes verbatim recall faster than general ability does. $\Gamma = 0$ means the two degrade together, and precision is not a useful control. Both retentions are ratios to a baseline measured in the same environment, so kernel-level non-determinism and hardware changes cancel rather than propagate.

The benchmark used for ρ_{cap} is a meaningful choice, termed the *capability anchor* here. Three anchors are reported throughout. They are the best retained of four benchmarks, a single benchmark held fixed across every model and quantizer, and the mean of all four. The first is permissive for a positive result, so the other two are reported alongside it in Appendix C. The choice of anchor changes the level by a few points but does not change any conclusion.

Language models. The Pythia suite (Biderman et al., 2023b) is audited at 410M, 1.4B, 2.8B, 6.9B and 12B parameters, covering a 29-fold range in parameter count. For each model, 1,000 64-token sequences that the model reproduces exactly are built, verified directly rather than taken from a released label. 1,000 paired low-accuracy controls from the same corpus (Gao et al., 2020) are also constructed, along with 200 shuffled windows that no model should reproduce, which serve as an instrument check. Capability is measured using LAMBADA, HellaSwag, ARC-Easy, and a fixed perplexity slice. Quantizers include a round-to-nearest implementation built for this work at $b \in \{8, 6, 5, 4, 3, 2\}$ bits, NF4 (Dettmers et al., 2023), INT8 (Dettmers et al., 2022), and GPTQ (Frantar et al., 2023), an error-correcting production quantizer. Items are paired across precisions, so every comparison is within-item. Intervals are 1,000 bootstrap resamples, whereas exact counts, revisions, quantizer configurations, and seeds are provided in Appendix A.

The second model family is Amber (Liu et al., 2024), a 7B LLaMA-architecture model whose pretraining corpus is *not* deduplicated and is fully released. This makes it suitable for replication because the number of times each memorized sequence appeared in the model's own training corpus can be counted exactly. The convenient alternative was an off-corpus n-gram search index, which measures how often a sequence occurs in another large corpus, thereby estimating exposure rather than counting training occurrences. Section 4 explains why this alternative was rejected in advance. 1,000 memorized text sequences and 1,000 code sequences are built using the same protocol, with source-matched controls.

Diffusion. Stable Diffusion v1.4 (Rombach et al., 2022) is audited using released captions known to reproduce their training images (Webster, 2023), together with a companion audit of Stable Diffusion 2.1-base. Replication is scored with a self-supervised copy detector (Pizzi et al., 2022) at $\tau = 0.75$. Quality is measured using CLIP-score (Radford et al., 2021) against the prompt, with FID (Heusel et al., 2017) against COCO as an independent distribution-level check.

The diffusion reproducing sets, R for the bit-width sweep and attribution, and S for the mechanism regression, are defined in Appendix A. Intervals over them are resampled at the caption level and read descriptively (Appendix G).

3 The selectivity gap across model scale

The primary result. Figure 1b and Table 1 show how the separation changes as models grow. Under NF4, the gap falls steeply at first, from +69.0 points at 410M to +26.2 at 2.8B. If this early decline continued at the same rate, the effect would close before 10B parameters. The decline does not continue. The step from 2.8B to 6.9B is +1.6 points and includes zero, so the gap essentially holds steady rather than continuing to shrink. The step from 6.9B to 12B is +7.6 and excludes zero, so the gap widens again rather than closing further. Across every model size and all three capability anchors tested, the gap never drops below 24.2 points, and its lower confidence bound never drops below 21.3.

The primary claim remains unchanged across the checks applied here. *The separation does not close over a 29-fold range in parameter count.* This result holds for all three capability anchors and whether adjacent differences are tested with paired or independent resampling. The 12B point is also not simply an artifact of switching the benchmark used as the anchor at that size (Appendix C).

Table 1: The selectivity gap Γ in percentage points, with 1,000-resample bootstrap 95% intervals. The gap is positive with a lower bound above zero in every cell. Bold marks the row used for the primary claim. At 12B, the largest model tested and the far end of the 29-fold range in parameter count covered by this study, the separation is larger than at 2.8B. All three capability anchors are tabulated in Appendix C. Exact parameter counts are given in Appendix A.

	NF4	GPTQ
Pythia-410M	+69.0 [+65.0, +73.4]	+72.3 [+68.2, +76.2]
Pythia-1.4B	+40.6 [+36.9, +44.3]	+44.2 [+40.7, +47.8]
Pythia-2.8B	+26.2 [+23.0, +29.4]	+24.1 [+21.1, +27.1]
Pythia-6.9B	+27.8 [+24.7, +30.8]	+17.0 [+14.4, +19.3]
Pythia-12B	**+35.4 [+32.0, +38.8]**	**+13.9 [+11.4, +16.5]**
Amber-7B (second family)	+49.1 [+45.7, +52.6]	+32.5 [+29.6, +35.4]

Under direct rounding, the gap is larger at 12B than at 6.9B. This increase does not reproduce under GPTQ, and more details can be found in (Appendix C).

Under an error-correcting production quantizer. Round-to-nearest is the weakest quantizer used in practice, so GPTQ, which corrects for rounding error rather than simply discarding it, is tested as well. Under GPTQ, the gap remains positive, with a lower confidence bound above zero for every model size and across all three capability anchors. The minimum lower bound is 10.4 points (Table 6). The phenomenon is therefore not an artifact of unsophisticated rounding.

Error correction does change one important aspect. GPTQ *recovers* part of what direct rounding destroys, and it recovers more when the quantizer is calibrated specifically on memorized data. This means that an adversary who controls the calibration set can partially undo the privacy benefit that quantization otherwise provides (Appendix E).

A second model family. Pythia's corpus is deduplicated, and its memorized-sequence sets are released, so the results above inherit both properties. The replication test, therefore, uses a model that shares neither. Amber is a 7B LLaMA-architecture model trained on an undeduplicated corpus, and no list of its memorized sequences has been released. The ground truth for this model is therefore derived directly using the same protocol used for Pythia.

The separation transfers to this second family and is even larger there. On text, $\Gamma = +49.1[+45.7, +52.6]$ points, with a capability retention of 1.0053. In practical terms, 48.6% of verbatim extraction is destroyed, while no capability measure falls below 0.962 of full precision. The size-matched Pythia comparison is +27.8 [+24.7, +30.8], which is noticeably smaller. This difference is reported rather than formally tested because the two intervals are computed at different levels of clustering, so a direct comparison of their widths would not have a clear interpretation. The instrument behind this comparison is clean regardless. The shuffled sanity set, which no model should be able to reproduce, has an extraction accuracy of 0.0000 at all 10 precisions tested. Memorized code is markedly more robust than memorized prose (Appendix D).

4 Predictors of quantization damage

A simple explanation for the separation is that quantization acts as a magnitude filter. It should preserve content that was trained most strongly, and heavily duplicated text is trained most strongly. Under this explanation, duplication should predict survival.

The three candidates. This explanation is tested against two alternative measures of training strength and a geometric alternative. Two measures of quantization damage are regressed on three per-sequence predictors, (a) Curvature: SAM sharpness (Foret et al., 2021) of the loss surface at the sequence, at perturbation radius $\rho = 0.05$, computed end-to-end in fp32 because the perturbations are below half precision's resolution, (b) Magnitude: the full-precision continuation loss, which measures how strongly the sequence is fitted, and (c) Duplication: the number of times the sequence appears in the pretraining corpus. A predictor is called *dominant* when its partial correlation excludes zero and leads the best competitor by at least 0.10. Appendix A states the rules as they were fixed, and Appendix D gives the full output.

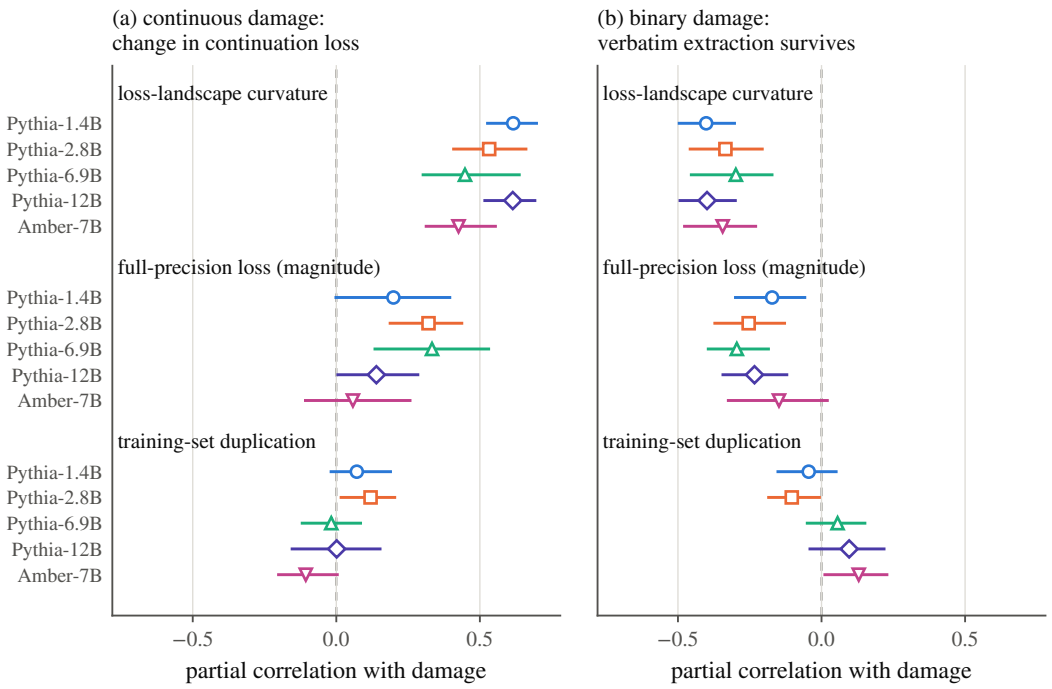


Figure 2: Curvature, rather than weight magnitude or training-set duplication, predicts which memorized sequences are destroyed by quantization. Each estimate is a partial correlation adjusted for the other two predictors, with a 1,000-resample bootstrap 95% interval. The two outcomes have opposite signs because one measures damage while the other measures survival. Curvature leads in every cell across four model sizes in both the first and second families. Duplication is the weakest predictor throughout. Its largest partial correlation in the figure is 0.13 in absolute value, including in the second family, where duplication is counted exactly against the model's own training corpus rather than estimated from a search index.

Curvature. Curvature is the strongest of the three predictors in almost all tests. For continuous damage, defined as the increase in sequence loss after quantization, sharpness leads at every model size, reaching partial $r = +0.62[+0.52, +0.70]$ at 1.4B ($n = 276$, $R^2 = 0.478$), +0.53 [+0.40, +0.67] at 2.8B ($n = 274$, $R^2 = 0.418$), +0.45 [+0.30, +0.64] at 6.9B ($n = 277$, $R^2 = 0.372$), and +0.61 [+0.51, +0.70] at 12B ($n = 289$, $R^2 = 0.520$). For the binary question of whether a memorized sequence survives extraction, curvature again leads at every size, reaching -0.40 [-0.50, -0.30], -0.33 [-0.46, -0.20], -0.30 [-0.46, -0.17], and -0.40 [-0.50, -0.30] (McFadden pseudo-R^2 0.229, 0.229, 0.222 and 0.294).

Across these eight tests, covering four sizes and two damage measures, curvature satisfies the dominance rule in six. In the remaining two, it is still the top-ranked predictor, but its margin over the closest competitor falls just short of the required threshold. These cases are therefore reported as leading rather than dominant (Appendix D).

Duplication. Duplication is substantially weaker than curvature. After adjusting for the other two predictors, its largest absolute partial correlation across the Pythia fits is 0.12. Where it differs marginally from zero, its sign runs *against* the prediction that heavily duplicated content should survive quantization better. This weakens rather than supports the original magnitude-based explanation.

Measuring duplication in the second family. A mechanism that outperforms duplication is only convincing if duplication is measured fairly. This is particularly important for the second family. An off-corpus search index, described in Section 2, was available and convenient, but it covers only about a quarter of a draw from Amber's corpus. Using it would have favored curvature by construction rather than evidence. It was therefore rejected *before measuring* in favor of counting exact 32-token occurrences directly against Amber's own corpus. A condition was fixed in advance that detection below 50% would be considered underpowered and treated as no comparative claim. Detection reached 62.0% on text and 65.7% on code, comfortably above that threshold, so the comparison uses a directly measured competitor.

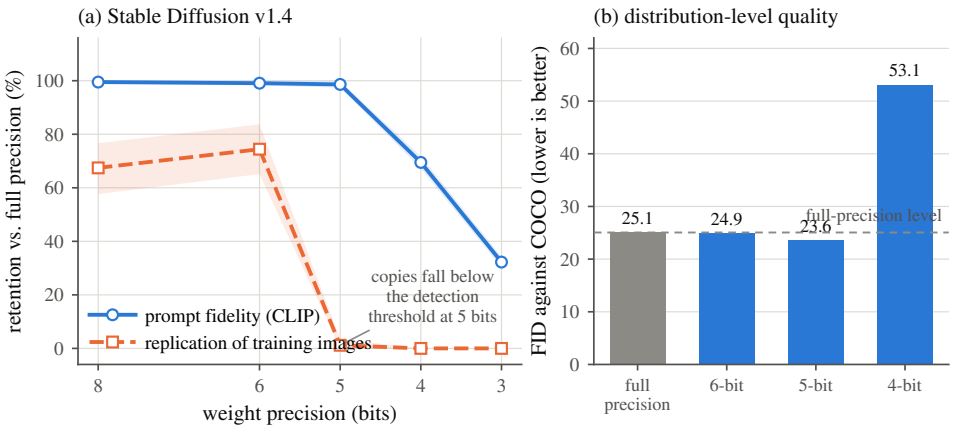


Figure 3: Replication of memorized training images collapses in a single bit-width step while both quality metrics remain flat. **(a)** Replication retention falls by 0.733 between 6-bit and 5-bit weights, reaching 0.012, while prompt fidelity remains at 0.986. **(b)** FID against COCO shows no measurable distribution-level quality loss at 5-bit weights, while 4-bit weights cause a substantial quality loss.

Replication in the second family. The same pattern holds outside Pythia. On Amber, sharpness reaches +0.43 [+0.31, +0.56] for continuous damage and -0.34 [-0.48, -0.22] for survival, is dominant on both, and has mechanistically consistent signs ($n = 300$, $R^2 = 0.211$, McFadden 0.155). Duplication, counted exactly rather than estimated, reaches -0.11 [-0.21, +0.01] for continuous damage and is similarly weak to its Pythia result. The two families differ in one respect. In Pythia, magnitude is a genuine secondary predictor, while on Amber, it is weak for both outcomes (+0.06 [-0.11, +0.26]). This difference is reported rather than averaged away.

5 TEXT-TO-IMAGE DIFFUSION

Replication across weight precision. In Stable Diffusion v1.4, the same phenomenon appears, but the transition differs from the language model case. Instead of declining gradually across several bit widths, replication collapses in a single step. Replication retention remains at 0.674 at 8-bit weights and 0.744 at 6-bit, with little change between them. It then falls sharply to 0.012 at 5-bit, a drop of 0.733 [0.642, 0.825]. Over the same step, prompt fidelity remains at 0.986 and FID changes only from 25.05 to 23.59 (Figure 3), a -1.5 FID change that is within sampling noise for a 5,000-image estimate. The claim is therefore that *no* distribution-level quality loss is detectable at 5 bits, not that quality improves. At 4-bit weights, replication falls further to 0.000, but this does not represent a further useful reduction. Prompt fidelity falls to 0.695 and FID rises to 53.11, indicating that the model itself is degraded at this precision. The absence of copies at 4 bits, therefore, does not show that anything useful has happened. Figure 4 shows this same collapse directly on six example captions, four memorized and two ordinary, making the one-bit offset between when memorized content fails and when ordinary content fails visible at the level of individual images.

At 5-bit weights, copies fall just below the detection threshold, so the result depends on the chosen threshold, unlike the 4-bit result (Appendix H). Stable Diffusion 2.1-base shows the transition at the same bit width, but the reduction is less complete (Appendix H).

Predictors of image survival. Weight-space curvature, which was a strong predictor for language models, does not carry over to diffusion. It does not predict which memorized images survive quantization with comparable strength, although an effect below $|r| \approx 0.28$ would be undetectable at the sample size used here (Appendix D.1). This failure points to a different location for memorization in diffusion models. A diffusion model produces an image by integrating a trajectory from an initial noise sample, *so memorization may depend on the sampling dynamics rather than the weights*. The quantity used to capture this is *basin stability*, which measures how much a generation changes when the initial noise is perturbed in a norm-preserving way. A memorized image in a wide attractor basin returns to the same picture from many different perturbed starting points, while one in a narrow basin does not. The attraction-basin account of diffusion memorization is established (Jain et al., 2025), as

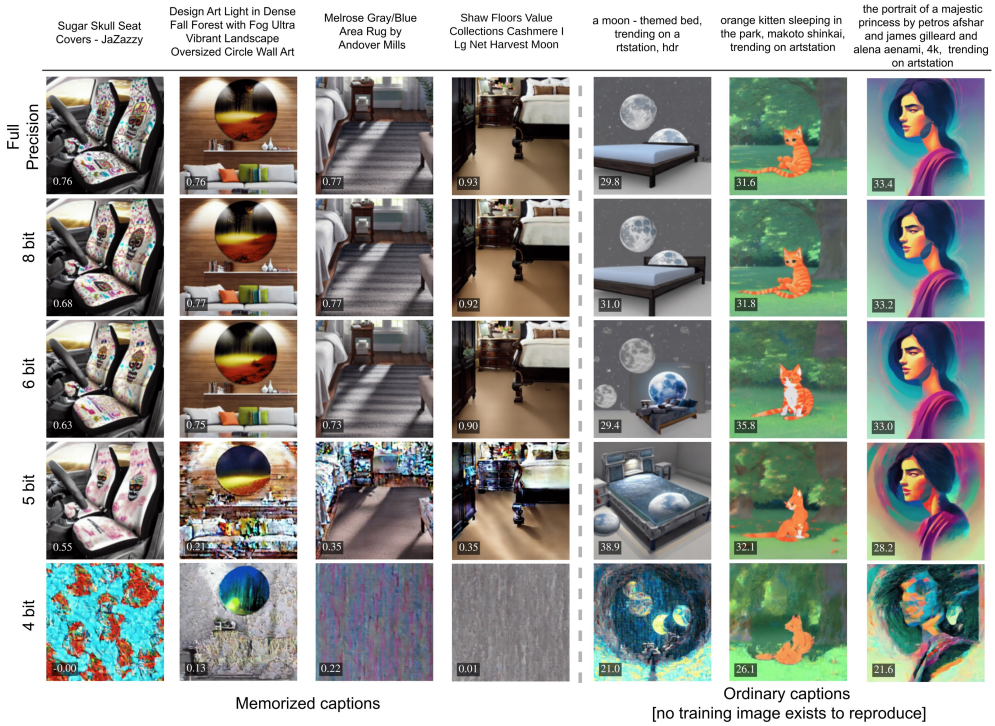


Figure 4: Numbers show copy-detector similarity to the caption's training image on the left ($\tau = 0.75$ counts as a copy) and CLIP fidelity to the prompt on the right. The two metrics use different scales, so comparisons should be made within each column rather than across the divider. Memorized captions break at 5-bit, while ordinary captions continue to match their prompts and fail only at 4-bit. This one-bit offset makes Γ (Eq. 1) visible. An unmemorized caption is not tied to a particular image, so the right-hand compositions change across precisions while their scores remain stable.

Table 2: Weight and activation quantization, decomposed. Three conditions share a byte-identical activation path or a byte-identical weight quantizer, making the comparison controlled by construction. The bold row shows the main finding. 8-bit activations with *full-precision weights* eliminate replication entirely while retaining the highest quality among the three conditions compared here.

	Replication retention	Prompt fidelity	Median similarity
4-bit weights only	0.1429 [0.0690, 0.2278]	0.9747	0.643
4-bit weights + 8-bit act.	0.0000	0.9694	0.177
8-bit activations only	**0.0000 [0.0000, 0.0000]**	**0.9949**	**0.322**

is the role of the initial noise in escaping that basin (Han et al., 2025). The question tested here is whether basin width predicts which copies are removed by quantization.

It does, and the key test is whether basin stability explains survival beyond baseline memorization strength. It reaches partial $r = +0.57[+0.36, +0.73]$ on survival and adds partial $r = +0.61[+0.40, +0.76]$ in the incremental test, raising McFadden pseudo-R^2 from 0.271 to 0.610 (Holm-adjusted $p = 0.0003$, $n = 103$). This signal is also nearly independent of the factors measured in the earlier analysis, with correlations of 0.085 with baseline memorization strength, 0.219 with static sharpness, and 0.076 with duplication. Basin stability, therefore, captures a distinct survival axis that static weight-space curvature does not capture. Memorized images that survive quantization sit in wide attractor basins of the sampling trajectory rather than in flat regions of the weight-space loss. Two competing explanations were tested and rejected (Appendix D.1).

Weight and activation quantization in a deployment recipe. A realistic deployment recipe quantizes weights and activations together, so it does not reproduce the clean result reported above. With 4-bit group-wise weights alone, at 0.9747 prompt fidelity, 0.1429 of replications return. Thus, the weight quantizer by itself only partially removes copying. A 2×2 decomposition identifies which component is responsible (Table 2). With weights kept at full precision and only an 8-bit activation quantizer applied, replication retention falls to 0.0000 at prompt fidelity 0.9949. Across all 84 reproductions, *not one is eliminated by the 4-bit weights that the activation quantizer does not*

also eliminate, at every detection threshold from 0.60 to 0.80. Those reproductions span only three independent reference images, so this is stated as a structural claim rather than a statistical claim based on a complete observed set. Full-precision weights combined with 8-bit activations provide a useful mechanistic condition for isolating this effect, not a deployment recommendation in itself. Appendix I gives the scope and the secondary-set replication.

6 Related work

Memorization and extraction. Language models can reproduce training data verbatim, with extraction rates that increase with model size, corpus duplication, and prompt length (Carlini et al., 2021; 2023b; Nasr et al., 2025). Duplication is one of the best-established drivers, and corpus deduplication measurably reduces extraction (Lee et al., 2022; Kandpal et al., 2022). The Pythia suite was developed to support this type of analysis and provides verified memorized-sequence sets (Biderman et al., 2023b;a). The mechanism result reported later in this paper differs from the duplication account on one specific question, namely, which memorized sequences survive an intervention. Duplication still predicts what a model memorizes in the first place. After adjusting for curvature, however, it does not predict which of those memorized sequences quantization destroys.

Curvature and memorization. Memorized examples are known to lie in sharp regions of the loss surface. Curvature has been used as a per-example measure for identifying which points a network has memorized (Garg et al., 2024; Ravikumar et al., 2024) and more recently for editing memorized content from a transformer (Merullo et al., 2025). These works use curvature to identify memorized content. The proposed work uses it for a different purpose, namely to predict which memorized content survives an intervention that curvature was not designed to detect. The two uses are not equivalent. A statistic that separates memorized examples from non-memorized examples does not necessarily rank memorized examples by how much damage numerical rounding will cause.

Post-training quantization. Modern methods reduce weights to 4 bits or fewer with small quality costs. They do so through error-correcting round-off using a calibration set (Frantar et al., 2023), scaling of salient channels (Lin et al., 2024), or information-theoretically motivated number formats (Dettmers et al., 2022; 2023). Compression can affect examples unevenly, with a small subset experiencing disproportionate changes while headline accuracy remains largely unchanged (Hooker et al., 2019). This literature focuses on perplexity, benchmark accuracy, and throughput. Memorized content is not among the reported quantities, while the memorization literature does not manipulate numerical precision. This paper examines the gap between these two lines of work.

Quantization and unlearning. Zhang et al. (2025b) report a result that appears to reverse the main result in this paper. In their setting, models that have been unlearned retain about 21% of the forgotten knowledge at full precision, but this increases to 83% after 4-bit quantization. Quantization, therefore, restores knowledge that unlearning had removed rather than further removing it. The two results are compatible once the measured objects are separated. Zhang et al. (2025b) evaluate question-answering knowledge in a model whose unlearning update is a small weight-space delta. Rounding erases a delta of that size and moves the model toward its pre-unlearning state. This paper instead evaluates verbatim extraction from a pretrained model without such a delta, in which rounding perturbs the sharp minima that directly encode memorized sequences. Both findings show that low-precision weights do not preserve fine-grained weight-space structure. The effect of losing that structure differs by setting. It is protective in the unlearning case and destructive in this one. Neither result supports treating quantization as a form of unlearning explicitly stated in Section 7.

Diffusion memorization. Text-to-image diffusion models can replicate training images for a small fraction of prompts (Carlini et al., 2023a; Somepalli et al., 2023a). The phenomenon is concentrated on duplicated images and template-like captions (Somepalli et al., 2023b). Reproducible extraction sets based on these findings (Webster, 2023) are the standard instrument and are the sets whose independence structure is characterized in Appendix G. Existing mitigation methods act on the caption, the sampling trajectory, or the fine-tuning update (Somepalli et al., 2023b; Wen et al., 2024). Numerical precision has not been considered among these interventions.

Mitigation by intervention on the model. Machine unlearning aims to remove specific training data after training, with formal guarantees in restricted settings and heuristic approximations otherwise.

Quantization is not unlearning. It provides no guarantees and is only partially reversible (Section 7). Reference-free membership and contamination detectors (Shi et al., 2024; Zhang et al., 2025a) can establish whether a sequence was used for training, but they cannot remove it. Quantization does the opposite. It is a training-free intervention already deployed at scale that can remove a model's ability to emit a sequence without detecting that sequence.

7 LIMITATIONS

Model coverage at larger scales. The central scaling result spans five model sizes over a 29-fold range in parameter count and is supported by an independent model family. At every size, however, the analysis uses one checkpoint, and above 7B, there is only one checkpoint in total. The reported confidence intervals are computed over sequences rather than independent training runs, so they measure consistency across items within each model. The rise observed at 12B therefore reflects that checkpoint and is reported without being used as a basis for later claims.

Mechanism results are predictive. The mechanism results in Sections 4 and 5 are correlational. Curvature and basin stability are measured as properties of the sequences and images rather than directly manipulated. The results, therefore, establish that these quantities predict survival under quantization, rather than that changing either quantity would produce a corresponding change in survival.

Independence of the diffusion units. The primary reproducing set contains 70 captions but only 5 independent reference images because multiple captions can share a reference image (Appendix G). Caption-level intervals are therefore narrower than the independent information in the data supports and are reported descriptively rather than as formal coverage statements. The diffusion mechanism result uses the same caption-level structure, while its point estimate is unchanged.

Verbatim reproduction only. The measurements in this paper target exact reproduction. A model that no longer emits training data word-for-word may still retain the same content in another form. The results should therefore be interpreted specifically as a reduction in verbatim recall.

Not an unlearning guarantee. The results concern what a specific extraction procedure recovers from a specific quantized model and do not provide a formal unlearning guarantee. A stronger search-and-sampling attacker is evaluated separately in Appendix F to assess the persistence of the effect under additional extraction effort. An attacker capable of restoring full precision is outside the threat model considered here. Within the considered setting, Section 3 further shows that an attacker who controls the quantizer's calibration set can partially recover the effect.

8 CONCLUSION

This paper asks what happens to a model's memorized training data when its numerical precision is reduced, a change already made in nearly every deployment for reasons of cost rather than privacy. We show that the effect is asymmetric. Verbatim memorization is damaged before general capability is, and this separation does not close as models grow. It holds across a wide range of parameter counts, under an error correcting production quantizer, and in a second, independently trained model family. We then ask why this happens. The natural expectation is that quantization should spare whatever was trained hardest, meaning heavily duplicated content, but we show that duplication does not predict which memorized sequences survive. What predicts survival instead is the sharpness of the loss surface around a memorized sequence. We find that the same selectivity also holds in text-to-image diffusion, but through a different mechanism. Weight-space sharpness does not carry over to this setting, and the property that predicts survival is the stability of the sampling trajectory. Finally, by decomposing a realistic deployment recipe into its parts, we show that this benefit in diffusion models comes primarily from the activation quantizer rather than from the weight quantizer, which is usually credited with it. These results are neither an unlearning method nor a privacy guarantee of any kind. The effect is partial, and an adversary who controls the quantizer's calibration set can reverse part of it. What our results do show is that numerical precision, a setting already present in every deployed model, is a privacy-relevant configuration choice, one that is currently made without any consideration of privacy.

REPRODUCIBILITY STATEMENT

Every threshold, decision rule and outcome category in this paper was committed to a version-controlled repository before the corresponding measurement existed, and each is applied by code rather than by inspection; Appendix B describes the apparatus and Appendix A lists the rules verbatim. Raw measurement directories are write-once, and every table and figure in this paper is regenerated from them by a script that emits the numbers directly into the manuscript, so no value here is transcribed by hand. Appendix A gives model identifiers and revisions, data-construction procedures, quantizer configurations, evaluation protocols, seeds, and the bootstrap procedure used for every interval. One diffusion checkpoint we audit is no longer publicly retrievable; results depending on it are pinned by a content hash over the weights, verified at load, and an archival deposit with checksums and license texts is prepared. Anonymized source and the analysis outputs will be released; the archival deposit identifier will be added on acceptance, since minting it now would identify the authors.

ETHICS STATEMENT

This work measures the extraction of memorized pretraining text and the reproduction of copyrighted training images, using publicly released models and publicly released research artifacts built for exactly this purpose. We construct no new corpus of personal data. Training images were retrieved only to score similarity and are not redistributed. The only extracted content shown is Figure 4: model reproductions of four commercial product images, none depicting a person, already documented as memorized in the public reproducing-prompt set. Otherwise, what we release are aggregate rates, per-sequence scores keyed to public identifiers, and the code that produces them. The extraction procedures we use are already published, and the reproducing-prompt sets we analyze are already public; we judge that the marginal uplift to a bad actor is small relative to the value of quantifying a mitigation. The intended use of this work is defensive, and we have tried to state its limits sharply enough that it cannot be cited as a privacy guarantee: Section 7 records that the effect is partly reversible, that it addresses verbatim reproduction rather than retained knowledge, and that it provides no formal guarantee of any kind. Our archival deposit of model weights carries the original licenses and their use-based restrictions unchanged.

USE OF AI

Large language models were used for the following, in all cases with the authors verifying the output. *Used:* drafting and editing prose in this manuscript; generating and refactoring analysis, figure and build code; summarizing literature during scoping. *Not used:* generating any synthetic data used in an experiment; producing, selecting or altering any experimental result; formulating the hypotheses, thresholds or decision rules, all of which were specified by the authors before measurement.

Verification of AI-assisted work was structural rather than by inspection alone. Every numerical value in this manuscript is emitted into the text by a generator that reads the recorded result files, so a mis-transcribed number is not expressible; analysis code was checked against previously recorded outputs, and the errors this caught are reported in Appendix B rather than silently corrected. The authors take responsibility for the entire contents of this submission.

REFERENCES

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems*, 2023a.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023b.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, 2023a.

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Isha Garg, Deepak Ravikumar, and Kaushik Roy. Memorization through the lens of curvature of loss function around samples. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15083–15101, 2024. arXiv:2307.05831.

Hyeonggeun Han, Sehwan Kim, Hyungjun Joo, Sangwoo Hong, and Jungwoo Lee. Adjusting initial noise to mitigate memorization in text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Chunsan Hong, Tae-Hyun Oh, and Minhyuk Sung. MemBench: Memorized image trigger prompt dataset for diffusion models. *Transactions on Machine Learning Research*, 2025.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, 2022.

Antoni Kowalczuk, Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding DoRI: Discovery of retained images in diffusion models. In *International Conference on Machine Learning*, 2026.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

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*, 2024.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. LLM360: Towards fully transparent open-source LLMs. In *Conference on Language Modeling*, 2024. arXiv:2312.06550.

Jack Merullo, Srihita Vatsavaya, Lucius Bushnaq, and Owen Lewis. From memorization to reasoning in the spectrum of loss curvature. *arXiv preprint arXiv:2510.24256*, 2025.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramer, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *International Conference on Learning Representations*, 2025. arXiv:2311.17035.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Deepak Ravikumar, Efstathia Soufleri, Abolfazl Hashemi, and Kaushik Roy. Unveiling privacy, memorization, and input curvature links. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42192–42212, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, 2024.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems*, 2023b.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *International Conference on Learning Representations*, 2024.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-K%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, 2025a.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *International Conference on Learning Representations*, 2025b. arXiv:2410.16454.

A Protocol and pre-registered rules

A.1 Models and revisions

Language models are the Pythia deduplicated suite at 405,334,016, 1,414,647,808, 2,775,208,960, 6,857,302,016 and 11,846,072,320 parameters (Biderman et al., 2023b), and Amber, a 7B LLaMA-architecture model with a fully released, non-deduplicated pretraining corpus (Liu et al., 2024). Diffusion models are Stable Diffusion v1.4 (Rombach et al., 2022) and Stable Diffusion 2.1-base.

One checkpoint was withdrawn from public hosting during this work. Its mirror does not include a license file, so the archival deposit attaches the license text with the manifest and checksums. The deposit is prepared, but publishing it would identify the authors, so no archive identifier appears in this anonymous version.

A.2 Constructing the memorized sets

For each language model, candidate sequences are sampled, and memorization is re-verified directly rather than taken from a released label. A sequence is retained only if greedy decoding of a 32-token continuation from its 32-token prefix reproduces the continuation exactly. The first 1,000 verified sequences are retained. Alongside them, 1,000 paired low-accuracy control sequences from the same corpus and 200 shuffled windows are constructed. The shuffled set serves as an instrument check and reads zero everywhere it is reported.

For the second family, no memorized-sequence list has been released, so the ground truth is derived using the same protocol. The base rate was measured independently of the released evaluation subsets and of a stratified, representative draw over the corpus. The two agree, validating the port. The representative draw is used because the released subsets are domain-skewed and would make the memorized set mostly code.

For diffusion, the reproducing set is built from released captions that reproduce training images, using seeds verified by the release rather than arbitrary ones. At arbitrary seeds, the observed reproduction rate is only a few percent, while at the verified seeds, it reaches roughly one in five. Two sets of reproductions support the diffusion experiments. The primary set is used for the main sweep and attribution result, while the secondary set is used for mechanism analysis. From 1,264 generated prompt-seed pairs over 316 captions, the primary set R contains 86 pairs spanning 70 captions that reproduce a training image at full precision. The secondary set S contains 103 reproducing pairs, one per caption, using captions from R together with 33 additional captions found through a wider seed sweep. Their union contains 119 pairs. R is used for the bit-width sweep and attribution result, while S is used for the mechanism regression and attribution replication. Captions in these releases are scored against retrieved reference images, and template-style captions frequently share one reference. Therefore, intervals in Section 5 are resampled at the caption level and should be interpreted as descriptive. Appendix G measures this structure. The control replication rate is 0.000 at every precision in every diffusion experiment reported here.

A.3 Quantizers

Round-to-nearest is an implementation built for this work. It is per-output-channel symmetric, applied to transformer-block linear layers, and excludes the language-model head. Quantizing the embedding matrices as well, at 4 bits, removes a further 0.5 to 6.0 points of extraction across the Pythia sizes and 13.0 points of text extraction retention in the second family (Table 9). This is small relative to the 73 and 74 points already removed by the transformer blocks alone. The effect is therefore mainly located in the transformer blocks rather than the embeddings. NF4 and INT8 use the standard implementations (Dettmers et al., 2023; 2022). GPTQ is a production error-correcting quantizer (Frantar et al., 2023) run in an isolated dependency tree, with kernels set to a reference backend rather than optimized ones, which affects speed but not numerics.

For diffusion, weight-only round-to-nearest is applied per output channel to the denoiser's linear and convolutional layers. The deployment conditions use 4-bit group-wise asymmetric weight quantization at group size 128 and an 8-bit activation path with static per-input-channel scales derived from a recorded calibration bank. The three conditions compared in Section 5 share either a byte-

identical activation path or a byte-identical weight quantizer, making the comparison controlled by construction. The activation configuration is hashed over its mode, bit width, hook class, target layer types, calibration bank and checksum, and scale rule. That hash is asserted equal before any image is generated. The realized per-layer scale tensors are also verified bit-identical across all 282 layers.

A.4 Evaluation

Capability is LAMBADA, HellaSwag, ARC-Easy and perplexity on a fixed 500,000-token slice. Extraction is exact-match on the 32-token continuation, with a multi-prefix variant at offsets 0, 8 and 16 recorded alongside. Every item and prompt is paired across precisions, so each comparison is within-item and quantization is the only changing factor.

For diffusion, replication is scored with a self-supervised copy detector (Pizzi et al., 2022) at $\tau = 0.75$, taking the maximum over each caption's reference image set. Quality is CLIP-score (Radford et al., 2021) against the prompt on a frozen 2,000-caption sample, with FID (Heusel et al., 2017) against a frozen 5,000-image COCO reference whose file list is fixed by checksum. Latents are paired across conditions, and the same generation identifier keeps the seed fixed at every bit width.

Intervals are 1,000 bootstrap resamples at seed 1234 throughout. Where the resampling unit is not the item, as for every diffusion result (Appendix G), the unit is stated with the interval.

A.5 The pre-registered rules

Mechanism dominance. A predictor is *dominant* if its partial correlation interval excludes zero and it leads the best competitor by at least 0.10. If no predictor qualifies or the fit falls below $R^2 = 0.05$, the outcome is recorded as no finding. This rule is applied by a function over the fitted output rather than by inspection.

Duplication fairness. In the second family, a comparison against duplication is allowed only when duplication is measured well enough to compete. If exact-count detection had fallen below 50%, it would have been declared under-powered and treated as no comparative claim. An off-corpus proxy index was available but rejected in advance because it covers roughly a quarter of a draw from that corpus and would have determined the comparison by construction.

Diffusion candidates. Three replacement candidates were fixed in order. Each candidate had to add explanatory power beyond baseline memorization strength rather than merely correlate with survival, with Holm correction across the three.

Deployment quality condition. A replication verdict is assigned only when quality retention is at or above 0.90. When the first deployment recipe considered failed this condition, no verdict was assigned, which motivated the later diagnosis and correction.

Detection threshold. The threshold was fixed at $\tau = 0.75$ before any measurement and was never changed. Sensitivity at 0.60, 0.70 and 0.80 is reported beside every verdict rather than used to select one.

B The measurement apparatus

The measurement apparatus is designed to make the decisions and the implementation checkable, so that the reported results can be reproduced rather than taken on faith.

Thresholds were fixed in code before the data existed. Every decision rule in this paper, including the dominance rule of Section 4, the detection threshold of Section 5, and the quality conditions used for the deployment comparison, was committed to the repository before the corresponding measurement. Each rule is applied by a function rather than by inspection. The commitment can be verified by comparing the recorded revision with the final analysis. For the mechanism analysis, the pre-registered rules were unchanged between commitment and the reported results.

Results are write-once and every table regenerates from them. Raw measurement directories are never overwritten. Regenerating every table and figure on a different machine, with a different

Table 3: The selectivity gap Γ under all three capability anchors, NF4, with 1,000-resample bootstrap 95% intervals. Every cell is positive with a lower bound above zero. The anchor changes the level by a few points and changes no conclusion.

model	best retained (primary)	HellaSwag, fixed	mean of four
Pythia-410M	+69.0 [+65.0, +73.4]	+69.0 [+65.0, +73.4]	+44.8 [+41.8, +47.8]
Pythia-1.4B	+40.6 [+36.9, +44.3]	+40.6 [+36.9, +44.3]	+28.5 [+25.4, +31.7]
Pythia-2.8B	+26.2 [+23.0, +29.4]	+26.2 [+23.0, +29.4]	+24.2 [+21.3, +27.0]
Pythia-6.9B	+27.8 [+24.7, +30.8]	+27.0 [+23.8, +30.2]	+25.7 [+22.9, +28.5]
Pythia-12B	+35.4 [+32.0, +38.8]	+34.7 [+31.1, +38.1]	+33.3 [+30.3, +36.4]

Table 4: Adjacent differences in Γ between neighbouring model sizes, paired bootstrap. The two smallest steps are strongly negative; the 2.8B→6.9B step includes zero under every anchor; the largest step is positive and excludes zero under every anchor.

step	best retained (primary)	HellaSwag, fixed	mean of four
410M→1.4B	-28.4 [-33.8, -23.2]†	-28.4 [-33.8, -23.2]†	-16.3 [-20.6, -11.7]†
1.4B→2.8B	-14.4 [-19.2, -9.5]†	-14.4 [-19.2, -9.5]†	-4.3 [-8.4, -0.2]†
2.8B→6.9B	+1.6 [-3.0, +6.2]	+0.9 [-3.9, +5.4]	+1.5 [-2.5, +5.8]
6.9B→12B	+7.6 [+2.9, +12.3]†	+7.7 [+2.9, +12.2]†	+7.5 [+3.2, +11.8]†

† interval excludes zero.

Table 5: Memorization retention / capability retention under round-to-nearest at each weight precision. Memorization falls first at every model size; below 4 bits capability follows.

model	8-bit	6-bit	5-bit	4-bit	3-bit	2-bit
410M	0.985 / 0.997	0.814 / 1.009	0.421 / 0.988	0.065 / 0.927	0.000 / 0.667	0.000 / 0.656
1.4B	0.993 / 1.004	0.912 / 0.981	0.654 / 0.991	0.119 / 0.914	0.000 / 0.578	0.000 / 0.542
2.8B	0.991 / 1.002	0.919 / 1.012	0.725 / 1.004	0.270 / 0.952	0.002 / 0.565	0.000 / 0.525
6.9B	0.986 / 1.003	0.910 / 1.007	0.662 / 0.978	0.194 / 0.920	0.001 / 0.495	0.000 / 0.472
12B	0.990 / 1.006	0.896 / 0.996	0.591 / 0.968	0.091 / 0.874	0.004 / 0.453	0.000 / 0.471

accelerator and an older framework version, reproduced the recorded values byte identically for the current-era results and to machine epsilon for the oldest ones.

One model audited here is no longer publicly retrievable. The second diffusion checkpoint was withdrawn from public hosting during this work. Every result that depends on it is pinned by a content hash over the weights and verified at load. An archival deposit is prepared with the manifest, checksums, and both license texts, since the mirror used to retrieve the checkpoint does not provide its own license file. The archive preserves the provenance needed to reproduce the diffusion results.

This apparatus also provides direct checks on the analysis pipeline. Applying the frozen decision statistic rather than an earlier implementation changed one verdict, demonstrating why the fixed rule is applied independently of the result. A pre-registered analysis pass was found to be skipped by a conditional with no failure branch, a case that was detectable through the write-once analysis checks. A per-sequence predictor table was also found to be keyed by position within a subset rather than by item identity, and the regeneration check exposed the mismatch. In the diffusion pipeline, a preprocessing asymmetry affecting one condition was identified and disclosed after the same checks surfaced it. These checks ensure that such implementation issues are identified during validation rather than entering the reported results.

C Capability anchors and the full scaling sweep

Table 3 provides the robustness check for the claim that the primary result is invariant to the capability anchor. The best-retained anchor is the pre-registered definition used in the main text. Holding one benchmark fixed across all models and quantizers avoids per-cell maximization, while averaging four

Table 6: The selectivity gap under GPTQ, an error-correcting production quantizer, at all three anchors. Positive with a lower bound above zero everywhere. The last column is the reason the 12B rise does not reproduce here: unlike direct rounding, error correction recovers more memorization as models grow.

model	best retained (primary)	HellaSwag, fixed	mean of four	extraction ret.
Pythia-1.4B	+44.2 [+40.7, +47.8]	+43.3 [+39.5, +46.9]	+34.0 [+30.8, +37.4]	0.541
Pythia-2.8B	+24.1 [+21.1, +27.1]	+24.1 [+21.1, +27.1]	+21.9 [+19.2, +24.5]	0.758
Pythia-6.9B	+17.0 [+14.4, +19.3]	+15.2 [+12.5, +17.8]	+15.9 [+13.5, +18.2]	0.830
Pythia-12B	+13.9 [+11.4, +16.5]	+13.9 [+11.4, +16.5]	+12.4 [+10.4, +14.8]	0.864

benchmarks provides the most conservative of the three anchors. The gap is positive with a lower bound above zero in all fifteen cells. Switching from the fixed anchor to the maximizing anchor increases the 6.9B value by less than one point, which bounds the effect of anchor selection.

Table 4 shows how the gap changes across model sizes. The two smallest steps are strongly negative under every anchor. The 2.8B→6.9B step includes zero under every anchor, indicating that the gap stops narrowing. The 6.9B→12B step is positive and excludes zero under every anchor. This provides the secondary scaling result, with the qualification that it rests on one checkpoint at that size and does not reproduce under an error-correcting quantizer.

Table 6 provides the corresponding GPTQ results. Extraction retention under GPTQ rises monotonically with model size, from 0.541 at 1.4B to 0.864 at 12B. Thus, error correction recovers more of what direct rounding removes as models grow. Together with the cross-quantizer comparison, this indicates that the 12B rise is specific to direct rounding rather than quantization in general.

The 12B rise. Under direct rounding, extraction retention is non-monotone across model sizes (0.287 / 0.580 / 0.727 / 0.709 / 0.627 from 410M to 12B), and this pattern is unchanged across the capability anchors. The rise is reported as observed without assigning a cause. The tested explanation, that increasing curvature accounts for the rise, is not supported. Median sharpness at 12B exceeds that at 6.9B by only 0.10 [-0.06, 0.32] ($p = 0.289$, $n = 300$ per size), which is too small to account for the size of the observed rise. Under GPTQ, the same step is -3.1 [-6.7, +0.2] points, further showing that the rise is specific to direct rounding rather than quantization in general.

C.1 The intervention, item by item

Figure 5 shows the intervention item by item. Each row represents one memorized sequence, one benchmark item, or one memorized image, with weight precisions aligned across all three panels.

C.2 Which exact continuations quantization removes

The selectivity gap compares exact reproduction of a 32-token continuation with benchmark capability. By itself, this does not establish whether the effect is specific to memorized text or also applies to exact recall more generally. Table 7 addresses this directly. Every item is a continuation that the full-precision model reproduces exactly. Each non-training item is paired with a memorized sequence having the same full-precision confidence, matched on mean loss, mean top-1 margin, and minimum top-1 margin over the 32 positions, with every standardized mean difference at most 0.08 after matching. Two types of text are compared with memorized text. *Held-out* items are 400 windows per size from the Pile validation and test splits whose full 64-token window does not occur in the training split. *Copied from context* items (363 to 400 per size) are held-out 64-token windows shown to the model once, followed by a request to generate the second half when the first half recurs.

The two comparisons show different patterns. Held-out text is lost at least as often as memorized text, with differences ranging from -6.2 to +2.5 percentage points under NF4 and from -7.7 to +3.5 under GPTQ. Text copied from context, in contrast, is retained more often. From 1.4B to 12B, it is retained +22.8 to +47.2 points more often than matched memorized text under NF4 and +14.0 to +33.8 points more under GPTQ. Together, these results show that quantization removes exact recall supported by the weights, whether the recalled text is memorized or held out, while largely preserving recall supplied by the immediate context.

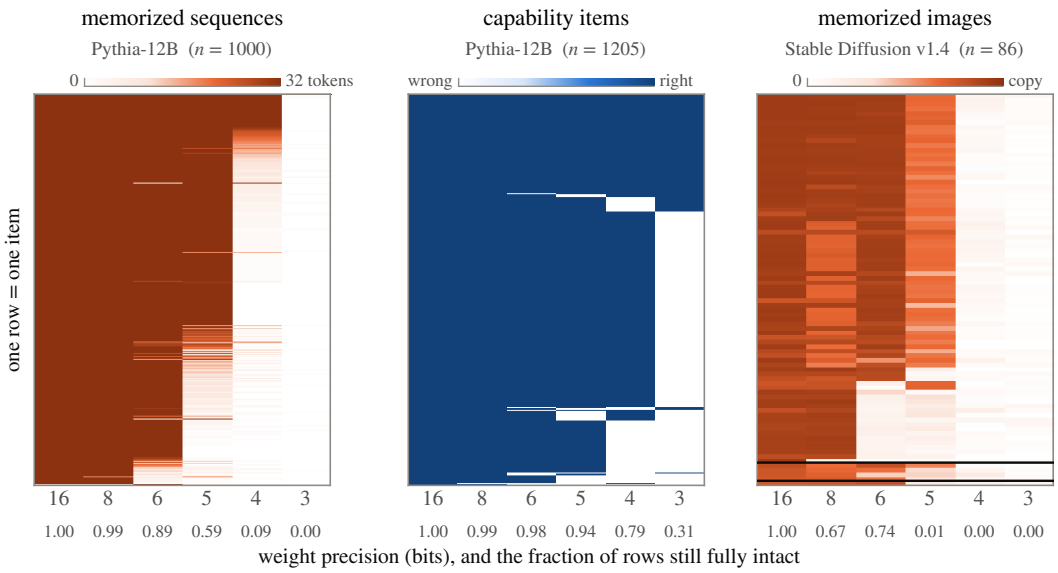


Figure 5: The same intervention, item by item, on a shared precision axis. Memorized content (verbatim continuations on the left, reproduced training images on the right) dissolves while the capability block holds through the same columns. The three color scales measure different quantities and are not comparable across panels, and the per-column figures are item-level rates, not the retention ratios of Eq. 1. The frontier is ragged in both memorized panels; which items fall on which side of it is the question of Section 4. Rules separate independent reference images, whose structure Appendix G measures.

Two checks further support this comparison. The minimum margin identifies the position where a continuation is most likely to break. Matching on this quantity was specified before it was computed, and adding it changes no NF4 difference by more than 2.7 points. Repeating the comparison on a second accelerator, with matching on mean loss and mean margin instead, gives +24.3 to +50.0 points for text copied from context under NF4 over the same sizes. Two scope conditions apply. The comparison with held-out text was specified before measurement, while the comparison with text copied from context was added afterward. Copied windows are also scored after a 96-token context rather than a 32-token prefix.

Table 7: Retention of exact 32-token continuations under quantization for two kinds of text the model was not trained on, each minus the retention of memorized sequences matched to it on full-precision mean loss, mean top-1 margin and minimum top-1 margin, in percentage points with 1,000-resample 95% intervals over source documents. Held-out: 64-token windows that do not occur in the training split. Copied from context: a held-out passage the model reproduces after seeing it once in its context. Negative values mean the held-out text is lost more often than memorized text.

	NF4		GPTQ	
model	held-out	copied from context	held-out	copied from context
Pythia-410M	-6.0 [-11.6, -0.8]	+7.2 [+0.8, +13.4]	-2.5 [-7.8, +2.6]	+7.2 [+0.8, +13.8]
Pythia-1.4B	-4.2 [-10.7, +2.0]	+22.8 [+16.1, +29.1]	-7.7 [-14.3, -0.7]	+33.8 [+27.5, +39.9]
Pythia-2.8B	-6.2 [-12.7, -0.2]	+32.5 [+27.7, +37.8]	-5.5 [-11.3, +0.7]	+22.8 [+18.1, +27.5]
Pythia-6.9B	-1.0 [-7.3, +5.0]	+32.8 [+27.5, +37.8]	+1.2 [-4.3, +7.0]	+23.2 [+18.6, +27.8]
Pythia-12B	+2.5 [-4.3, +8.6]	+47.2 [+42.3, +52.7]	+3.5 [-1.5, +8.9]	+14.0 [+10.2, +17.9]

D MECHANISM: FULL REGRESSION OUTPUT

Table 8 gives the complete output behind Figure 2, covering four sizes in the first family and both strata of the second family. Three points are especially relevant to the claims in the main text.

First, curvature does not dominate every cell, and the two cases where it only leads are important to state clearly. On the binary outcome at 2.8B and 6.9B, curvature leads the full-precision loss by 0.081

Table 8: Complete partial-correlation output for the mechanism analysis: every predictor adjusted for the other two, on both damage outcomes, at four model sizes in the first family and in both strata of the second. Signs on the two outcomes are opposite because one measures damage and the other survival. Curvature leads every cell. The code stratum is the one place where the two outcomes disagree about which predictor leads, and it is reported as such rather than averaged with the text stratum.

model	predictor	continuous damage	extraction survives
Pythia-1.4B	curvature (SAM sharpness)	+0.62 [+0.52, +0.70]	-0.40 [-0.50, -0.30]
	full-precision loss (magnitude)	+0.20 [-0.01, +0.40]	-0.17 [-0.30, -0.05]
	training-set duplication	+0.07 [-0.02, +0.19]	-0.04 [-0.16, +0.06]
$n = 276$, $R^2 = 0.478$, McFadden pseudo-$R^2 = 0.229$			
Pythia-2.8B	curvature (SAM sharpness)	+0.53 [+0.40, +0.67]	-0.33 [-0.46, -0.20]
	full-precision loss (magnitude)	+0.32 [+0.18, +0.44]	-0.25 [-0.38, -0.12]
	training-set duplication	+0.12 [+0.01, +0.21]	-0.10 [-0.19, -0.00]
$n = 274$, $R^2 = 0.418$, McFadden pseudo-$R^2 = 0.229$			
Pythia-6.9B	curvature (SAM sharpness)	+0.45 [+0.30, +0.64]	-0.30 [-0.46, -0.17]
	full-precision loss (magnitude)	+0.33 [+0.13, +0.54]	-0.29 [-0.40, -0.18]
	training-set duplication	-0.02 [-0.12, +0.09]	+0.06 [-0.05, +0.16]
$n = 277$, $R^2 = 0.372$, McFadden pseudo-$R^2 = 0.222$			
Pythia-12B	curvature (SAM sharpness)	+0.61 [+0.51, +0.70]	-0.40 [-0.50, -0.30]
	full-precision loss (magnitude)	+0.14 [+0.00, +0.29]	-0.23 [-0.35, -0.12]
	training-set duplication	+0.00 [-0.16, +0.16]	+0.10 [-0.05, +0.22]
$n = 289$, $R^2 = 0.520$, McFadden pseudo-$R^2 = 0.294$			
Amber-7B (text)	curvature (SAM sharpness)	+0.43 [+0.31, +0.56]	-0.34 [-0.48, -0.22]
	full-precision loss (magnitude)	+0.06 [-0.11, +0.26]	-0.15 [-0.33, +0.03]
	training-set duplication (exact count)	-0.11 [-0.21, +0.01]	+0.13 [+0.01, +0.23]
$n = 300$, $R^2 = 0.211$, McFadden pseudo-$R^2 = 0.155$			
Amber-7B (code)	curvature (SAM sharpness)	+0.33 [+0.20, +0.52]	-0.32 [-0.44, -0.21]
	full-precision loss (magnitude)	+0.09 [-0.03, +0.27]	-0.32 [-0.43, -0.19]
	training-set duplication (exact count)	-0.00 [-0.07, +0.07]	+0.05 [-0.06, +0.14]
$n = 300$, $R^2 = 0.143$, McFadden pseudo-$R^2 = 0.242$			

Table 9: The second family across every precision, split by domain. Memorized code is markedly more quantization-robust than memorized prose at every width. The gap peaks at 4-bit and then falls at 3 and 2 bits because capability collapses; memorization does not return.

configuration	capability ret.	text		code	
		extraction ret.	Γ	extraction ret.	Γ
full precision	1.000	1.000	+0.0 [+0.0, +0.0]	1.000	+0.0
8-bit	1.001	0.983	+1.8 [+0.8, +3.0]	0.988	+1.3
6-bit	1.007	0.914	+9.3 [+7.4, +11.5]	0.952	+5.5
5-bit	1.002	0.780	+22.2 [+19.6, +25.1]	0.889	+11.3
4-bit	0.964	0.261	+70.3 [+67.2, +73.8]	0.665	+29.9
3-bit	0.598	0.036	+56.2 [+53.0, +59.7]	0.084	+51.4
2-bit	0.437	0.000	+43.7 [+40.5, +47.0]	0.000	+43.7
4-bit incl. embeddings	0.886	0.131	+75.5 [+72.6, +78.6]	0.355	+53.1
NF4	1.005	0.514	+49.1 [+45.7, +52.6]	0.792	+21.3
INT8	1.001	0.942	+5.9 [+4.4, +7.5]	0.967	+3.4
GPTQ	0.998	0.672	+32.5 [+29.6, +35.4]	0.843	+15.4

and 0.003, respectively, against the required margin of 0.10. These cells are therefore reported as curvature leading rather than dominating. At 6.9B, the margin is small enough that the two predictors are effectively indistinguishable on this outcome at that size.

Second, the two model families differ in their second-ranked predictor. In Pythia, the full-precision loss is a genuine secondary predictor, with an interval excluding zero in most cells. In the second family, it is weak on both outcomes. This difference is reported directly rather than averaged away. The result that replicates across the two families is the leading role of curvature, not the complete ordering of predictors.

Table 10: The three pre-registered candidates for the diffusion mechanism, in the order they were fixed. A candidate had to add explanatory power over baseline memorization strength, not merely correlate with survival. The lead candidate fires and then adds nothing; the second adds a great deal; the third does not clear the pre-set floor.

candidate	marginal partial r	incremental partial r	pseudo-R^2 before	after
text-conditioning attention concentration	+0.385 [+0.152, +0.588]	+0.023 [-0.214, +0.282]	0.271	0.271
basin stability	+0.567 [+0.361, +0.734]	+0.611 [+0.398, +0.759]	0.271	0.610
cross-attention localization	+0.070 [-0.011, +0.153]	—	—	—

Holm-adjusted p: attention 0.0020, basin stability 0.0003, localization 0.0894.

Table 11: Quantizing one block group at a time, everything else left at full precision. A random size-matched non-attention subset destroys more replication than cross-attention does, and the convolutional blocks at intermediate resolutions eliminate every copy on their own, which localization to text conditioning would not produce.

block group quantized alone	mean similarity drop	replication survives
cross-attention	0.249	61%
self-attention	0.179	43%
feed-forward	0.106	76%
projections	0.155	68%
all ResNet blocks	0.556	0%
ResNet, 8×8	0.050	82%
ResNet, 16×16	0.509	0%
ResNet, 32×32	0.566	0%
ResNet, 64×64	0.218	74%
random size-matched control	0.339	38%

Third, the code stratum of the second family is the only setting where the two damage outcomes differ in their leading predictor. For continuous damage, curvature is dominant. For the binary outcome, the full-precision loss ties it. Memorized code is more robust to quantization than memorized prose throughout this analysis, as in the bit-width sweep, and the text stratum is the one directly comparable to Pythia.

Code and prose in the second family. Memorized code is more robust to quantization than memorized prose. When the second family's memorized set is split by domain, the gap on text exceeds the gap on code by +27.8 [+24.1, +31.7] points, with extraction retention of 0.792 for code and 0.514 for text. The effect therefore varies by domain rather than being uniform. The text stratum is reported in the main body and is the one directly comparable to Pythia.

The 6.9B and 12B fits. At these two sizes, sharpness is computed with gradients held in host memory. This reproduces the standard computation bit-identically where both methods apply. Recomputing the 2.8B values in the environment used for these larger sizes reproduces the recorded values to a median absolute difference of 1.7×10^{-4} with a correlation of 1.000. The predictor tables at these sizes have duplication coverage of 91.3% and 93.0%, respectively, comparable to the smaller sizes, so the same measurement quality is maintained across the full range.

Duplication, and how it was counted. In the second family, duplication is the number of exact 32-token occurrences of a sequence in the model's own pretraining corpus. It is estimated from a uniform sample of that corpus at a fixed sampling fraction, with no model involved in the count. Detection reached 62.0% on text and 65.7% on code, comfortably above the 50calibration diagnostic against the rejected proxy index gives a Spearman correlation of around 0.81 on the overlapping subset. This supports the use of the direct corpus count while providing a check against the proxy measure.

D.1 The diffusion mechanism and alternative predictors

Weight-space curvature. The predictor that is strong for language models does not show the same relationship for diffusion. Fitting the same three language-model predictors to image survival gives a fit too weak to support a finding, and static weight-space sharpness has a partial correlation well

Table 12: Calibrating an error-correcting quantizer on memorized rather than generic data partially restores what quantization removed. The control moves in the opposite direction, so this is not generic calibration quality improving everything at once. Probe-grade: two model sizes, direction established, magnitude not.

model	quantity	generic calib.	memorized calib.	difference
1.4b	verbatim extraction	0.541	0.606	+0.065 [+0.036, +0.097]†
1.4b	control perplexity	0.907	0.878	-0.029 [-0.034, -0.023]†
1.4b	multi-prefix extraction	0.679	0.698	+0.020 [-0.025, +0.066]
2.8b	verbatim extraction	0.758	0.810	+0.052 [+0.025, +0.076]†
2.8b	control perplexity	0.974	0.963	-0.011 [-0.015, -0.008]†
2.8b	multi-prefix extraction	0.808	0.855	+0.047 [+0.007, +0.087]†

† interval excludes zero.

below the effect size measured for language models. Two points qualify this result. Weight-space curvature does not predict image survival with comparable strength to its relationship with sequence survival. The sample used here can detect correlations only above about 0.28 in absolute value, so a genuinely weak effect would not be detectable at this scale.

Three candidate predictors of image survival were tested, with one retained. The other two are reported because they represent alternative explanations for the same underlying data. They are text-conditioning attention concentration and localization of memorization to the cross-attention blocks.

Table 10 compares all three candidates. The key test is the incremental one, which asks whether a candidate explains variation in survival beyond baseline memorization strength. *Text-conditioning attention concentration* does not add explanatory power. Its marginal correlation is present, but its incremental correlation is indistinguishable from zero because it is already 0.694-correlated with baseline memorization strength. It therefore largely measures the same quantity through the attention map. *Basin stability*, in contrast, adds +0.61 incrementally and raises pseudo-R^2 from 0.271 to 0.610. The result is also robust to the choice of radius. Across three perturbation radii, the incremental correlation remains monotone in radius and every interval excludes zero.

Cross-attention localization is evaluated with a direct block-level test. Table 11 quantizes one block group at a time while keeping everything else at full precision to identify where replication is affected. A random size-matched non-attention subset destroys more replication than cross-attention (0.339 against 0.249), and the localization contrast is +0.070 [-0.011, +0.153] ($p = 0.089$). Replication instead depends on the convolutional backbone at intermediate resolutions. Quantizing those blocks alone leaves only 0% of copies standing, effectively eliminating them. This result is consistent with separate evidence that memorization in these models is not local (Kowalczuk et al., 2026), and it further supports the distinction between baseline memorization strength and the predictor of which images survive quantization.

E CALIBRATION-SET LEAKAGE

An error-correcting quantizer chooses rounding to minimize error on its calibration set. If the calibration set contains memorized content, the quantizer can therefore preserve that content more strongly because doing so reduces calibration error. The results support this possibility. Extraction retention increases by +0.065 at 1.4B and +0.052 at 2.8B when the quantizer is calibrated on memorized data rather than generic data, with both intervals excluding zero. Control sequences move in the opposite direction under the same change. This indicates that the increase is not simply a general improvement in retention from using a different calibration set, but is specific to the memorized sequences. The multi-prefix variant shows the same direction.

This result defines an important condition for the effect studied in this paper. An adversary who controls the calibration set can recover part of what quantization would otherwise remove without changing the bit width. The result is reported as probe-grade evidence rather than a fully established finding. Two model sizes establish the direction of the effect, but not its magnitude, and do not determine whether the effect increases or decreases at larger model scales.

F EXTRACTION AGAINST A SEARCH-AND-SAMPLING ATTACKER

Table 13: Verbatim extraction removed by quantization, under greedy decoding and under an attacker that pools greedy decoding, beam search with widths 5, 10 and 20 (all beams kept) and 100 samples at temperature 1, and submits the candidate the quantized model itself scores most likely. Reduction is the share of the full-precision model's extraction that the quantized model loses against the same attacker. The pooled-candidate column is an upper bound: it counts a sequence as extracted if the true continuation is anywhere in the pool. The gap uses the recorded capability retention of each configuration. Intervals are 1,000-resample bootstrap 95% intervals.

configuration	destroyed (greedy)	extraction removed (%) greedy	attacker, top-1	pooled candidates	Γ greedy / attacker
Pythia-12B, RTN 4-bit	907	90.9	88.0 [86.6, 89.3]	80.6	+78.3 / +75.3 [+74.0, +76.7]
Pythia-12B, NF4	372	37.3	23.7 [21.1, 26.3]	11.1	+35.4 / +21.8 [+19.2, +24.4]
Pythia-2.8B, NF4	270	27.0	19.9 [16.5, 23.4]	6.4	+25.9 / +18.8 [+15.4, +22.3]

Table 14: Where the attacker's recoveries come from: the share of sequences that greedy decoding no longer extracts after quantization whose exact continuation each component produces, and the share each ranking of the pooled candidates places first.

component (% of destroyed sequences)	12B, RTN 4-bit	12B, NF4	2.8B, NF4
beam search, width 5	7.4	53.0	63.0
beam search, width 20	11.0	69.9	76.7
100 samples, temperature 1	1.4	44.1	54.4
100 samples, temperature 0.7	5.7	54.0	63.0
any candidate in the pool	11.2	70.2	76.7
top-1 by model likelihood	4.0	34.4	33.0
top-10 by model likelihood	10.6	67.5	74.1
top-1 by compression ratio	5.4	36.0	33.0
top-1 by sample frequency	0.9	23.9	25.2

Extraction elsewhere in this paper uses greedy decoding. A sequence counts as extracted when the most likely continuation of its 32-token prefix matches all 32 following tokens. A real attacker is not limited to this decoding strategy, so the question here is whether the selectivity gap persists against a stronger search-and-sampling attack. The attacker sees only the prefix and the quantized model. It pools the greedy continuation, every beam from beam search at widths 5, 10 and 20, and 100 samples at temperature 1, then selects the candidate that the quantized model scores most likely. A further 100 samples at temperature 0.7 are reported separately and do not change the conclusion. The attacker is run on every sequence that is extracted by greedy decoding at full precision but not after quantization, on 100 randomly chosen sequences that remain extractable, and on 200 prefixes from the control set. The same attacker is also run at full precision on 200 randomly chosen memorized sequences, where it extracts a fraction 0.980 of them at 12B. The reduction reported throughout is the share of this full-precision extraction lost under quantization, with each reduction measured on a single accelerator.

The question is whether this reduction remains above a meaningful threshold against the stronger attacker. The pre-specified criterion required a reduction of at least 50% at 12B, with 20% to 50% counted as a partial effect. Under 4-bit round-to-nearest, the observed reduction is 88.0% [86.6, 89.3], compared with 90.9% under greedy decoding, so the criterion is met in full. Under NF4, it is 23.7% [21.1, 26.3], compared with 37.3% under greedy decoding, which falls in the partial range. The selectivity gap nevertheless remains positive with its interval above zero in every configuration measured, reaching +21.8 [+19.2, +24.4] for NF4 at 12B (+35.4 under greedy decoding), +75.3 [+74.0, +76.7] for round-to-nearest, and +18.8 [+15.4, +22.3] for NF4 at 2.8B. Ranking the pool by compression ratio or sample frequency instead of likelihood gives reductions of 22.8% and 28.1% for NF4 at 12B. This shows that the result does not depend on the particular ranking rule. No control prefix was ever completed with its true continuation, giving a false-recovery rate of 0.000 in every configuration tested.

The next question is where the attacker's recovered sequences come from. This helps identify which part of the attack contributes most to the recovery. Beam search is the strongest single component (Table 14). For NF4 at 12B, width-20 beams alone contain the true continuation for 69.9% of

destroyed sequences, compared with 44.1% for 100 samples at temperature 1. Ranking is the main bottleneck rather than candidate recall. The pool contains the correct continuation for 70.2% of destroyed sequences, but the attacker's first choice is correct for only 34.4% of them. The pooled-candidate column of Table 13 therefore provides an upper bound that none of the ranking rules tested here reaches. Under this upper bound, NF4 at 12B would remove 11.1% of extraction and round-to-nearest would remove 80.6%. Likelihood ranking also sometimes displaces a continuation that greedy decoding alone still produces correctly. 99.0% of such sequences are ranked first under NF4 at 12B, and 90.1% under round-to-nearest.

As an independent check of the sampling component, a closed-form estimate for the temperature-1 sampler follows directly from the teacher-forced loss. A single draw emits the exact 32-token continuation with probability $e^{-32,\bar{\ell}}$, where $\bar{\ell}$ is the mean per-token loss of that continuation. Thus, K draws emit it at least once with probability $1-(1-e^{-32\bar{\ell}})^K$. This estimate closely tracks the sampler that was actually run. It predicts 42.5% compared with 44.1% observed for NF4 at 12B with 100 samples, 59.1% compared with 60.5% with 1,000 samples on a stratified subsample of 200 destroyed sequences, and 1.6% compared with 1.4% under round-to-nearest. As an estimate of the full attacker rather than sampling alone, the closed form runs low because beam search recovers more than sampling does on its own.

GPTQ and the 6.9B model are excluded from this analysis because of a data integrity safeguard. Each configuration was first required to reproduce its recorded greedy labels on at least 99% of sequences before being included. The NF4 checkpoint at 12B reproduced them on 100.0%, clearing this threshold. Re-quantizing with GPTQ instead gives a checkpoint whose labels agree on 98.0% at 12B, below the threshold. At 6.9B, the recorded labels come from a different accelerator, whose argmax near-ties differ from the recorded ones by as much as two runs of the same configuration on two different accelerators differ from each other, with 98.7% agreement. Both configurations were therefore excluded rather than reported with unreliable comparisons.

G Diffusion: benchmark structure and the independence of its units

Table 15: Benchmark structure in all four models these released caption sets cover. "shared" is the percentage of captions that share a byte-identical reference file with another caption; "top file" is how many captions the single most reused photograph is scored against. Realistic Vision is a fine-tune of v1.4 and is not independent evidence; DeepFloyd IF is the genuinely independent point, with a different architecture and text encoder, and it shows the same structure.

model	captions	GT files	distinct	shared	top file	groups	largest	singletons
Stable Diffusion v1.4	403	1,104	401	72%	155	124	155	114
Stable Diffusion 2.1-base	196	717	211	95%	158	13	158	9
DeepFloyd IF	144	475	159	79%	90	35	90	30
Realistic Vision	204	767	223	96%	145	14	145	9

This appendix records a property of the instrument that was not known when the diffusion experiments were designed. It affects how the intervals in Section 5 should be interpreted and applies to the measurements in this paper as well as to those from other studies.

Diffusion memorization is studied using released sets of captions known to reproduce their training images. Each caption is scored by comparing the generated image against a set of retrieved reference images for that caption. For template-style captions, such as product listings that differ only in a model number, these reference sets can resolve to *the same photograph*. Captions scored against a byte-identical reference file therefore cannot be treated as independent outcomes.

The effect is large and does not depend on a detection threshold. It is visible by hashing the files, without using a model or similarity measure. Across the four models covered by these benchmarks, 72-96% of memorized captions share a reference file with another caption. One file is shared by as many as 158 captions, and a single group accounts for 38-81% of all captions with ground truth. The sharing is therefore not a rare edge case but a common property of the benchmark. The image banks themselves are not degenerate, since they contain hundreds of genuinely distinct images. The

dependence arises specifically from *which images each caption can be scored against*. This property is determined by how the benchmarks are constructed, rather than by any one model. It holds for a pixel-space model with a different text encoder as strongly as for the latent models, and for that model the file-hash grouping exactly matches the similarity-based grouping.

Effect on these reproducing sets. The primary reproducing set R has 70 captions but only 5 independent reference images. The second family has 159 captions and 4 independent reference images, one of which covers 155 of them. Caption-level intervals in Section 5 are therefore narrower than the independent information supports, by roughly an order of magnitude in nominal n. They should be read as descriptive rather than as formal coverage statements. No point estimate changes as a result, and no result is removed. The structural claims about which conditions eliminate which reproductions are stated over complete observed sets and do not depend on a significance test, so they are unaffected by this issue.

Table 15 provides the evidence behind these numbers. Two columns are central to the argument. 'Shared" is the fraction of memorized captions that are scored against a reference file also used for another caption, and it ranges from 72% to 96%. 'Top file" is the number of captions scored against the single most reused photograph, which reaches 158.

Two checks are important for interpreting this table. First, the grouping is based on byte-level identity. Taking the transitive closure of "these two captions share an identical reference file" agrees with similarity-based clustering to within a few groups in every model, and exactly for the pixel-space model. This rules out the grouping being an artifact of a particular similarity threshold. Second, the image banks are not themselves degenerate. Clustering the reference images against each other merges at most one image beyond byte identity in any model, leaving between 159 and 401 genuinely distinct images. The dependence is therefore specifically in which images each caption can be scored against, not in the images themselves.

Two caveats apply to the four-model claim. Realistic Vision is a fine-tune of v1.4, so its near-identical memorized set is expected and does not provide independent evidence. DeepFloyd IF is the genuinely independent point because it has a different architecture, operates in pixel space, and uses a different text encoder. It is also the model for which byte identity exactly reproduces the clustering. Its memorized caption set is a subset of v1.4's, and its reference images are already in the existing bank. Adding more models therefore does not provide more independent items.

G.1 Independent groups outside the largest one

Table 16: Replication retention in every independent reproducing group of reference images, at every weight bit width and activation condition, at the pre-specified detection threshold. A group is the set of captions scored against the same reference image; the largest group of each architecture is listed separately, over all of its recorded reproducing pairs, and is not part of any count in the text. The activation conditions quantize the activations of every linear and convolutional layer statically to 8 bits, with full-precision or 4-bit group-wise weights; – marks a condition not generated for that architecture.

group	pairs	8-bit	6-bit	5-bit	4-bit	8-bit act.	4-bit g128	4-bit g128 + 8-bit act.
v1.4, MemBench group 1	27	0.89	0.19	0.00	0.00	0.00	0.00	0.00
v1.4, MemBench group 2	20	1.00	0.35	0.00	0.00	0.00	0.00	0.00
v1.4, MemBench group 3	6	0.83	0.67	0.00	0.00	0.00	0.00	0.00
v1.4, MemBench group 4	3	0.67	0.00	0.00	0.00	0.00	0.00	0.00
v1.4, MemBench group 5	2	0.50	0.00	0.00	0.00	0.00	0.00	0.00
v1.4, MemBench group 6	1	1.00	1.00	0.00	0.00	0.00	0.00	0.00
v1.4, MemBench group 7	1	1.00	1.00	0.00	0.00	0.00	0.00	0.00
v1.4, Webster group 1	6	0.17	0.17	0.00	0.00	0.00	0.00	0.00
v1.4, Webster group 2	1	1.00	1.00	0.00	0.00	0.00	0.00	0.00
v1.4, Webster group 3	1	1.00	0.00	0.00	0.00	0.00	0.00	0.00
v1.4, Webster group 4	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
v2.1-base, Webster group 1	2	1.00	0.00	0.00	0.00	0.00	0.00	0.00
v2.1-base, Webster group 2	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
v2.1-base, Webster group 3	1	1.00	1.00	0.00	0.00	0.00	0.00	0.00
v1.4, largest group	110	0.77	0.81	0.06	0.00	0.00	0.11	0.00
v2.1-base, largest group	155	0.99	0.96	0.46	0.00	–	–	–

Because nearly every reproduction in the primary set traces back to one dominant reference image, a natural concern is that the main diffusion findings, namely that the activation quantizer eliminates replication and that replication drops sharply between 6 and 5 bits, might reflect that single dominant cluster rather than a broader pattern. To test this directly, the bit-width sweep and activation conditions were repeated on every other group of reproducing captions available, 14 groups in total outside the largest one of each architecture. For v1.4, these are 7 groups formed by captions from MemBench (Hong et al., 2025) that attach to no existing group, and 4 groups from the captions of Webster (2023). For v2.1-base, they are 3 groups from the same source. Together, these groups contain 73 reproducing prompt-seed pairs, each generated with its recorded seed under every condition of Table 16. The unit of inference throughout is the group, not the pair. Before generating any new condition, 20 recorded pairs from the primary set were regenerated at full precision and with 8-bit activations as a pipeline check. The copy decision agreed with the recorded result for every pair under both conditions, with 20 and 20 agreement.

Two criteria were fixed in advance, one for each of the paper's two main diffusion findings. The first asked whether the activation quantizer result generalizes beyond the largest cluster, requiring replication retention of at most 0.10 with full-precision weights and 8-bit activations. This criterion was met in 14 of 14 groups (Clopper-Pearson 95% interval [0.77, 1.00]), and retention fell under the activation quantizer in every group tested (exact sign test, $p = 0.0001$). The second asked whether the single-step 6 to 5 bit collapse generalizes in the same way, requiring retention of at least 0.50 at 6 bits and at most 0.10 at 5 bits. This criterion was met in only 5 of 14 groups ([0.13, 0.65]). Retention was lower at 5 bits than at 6 in all 8 groups without a tie ($p = 0.008$). The remaining 6 groups tied because no replication remained at either width. In 9 groups, retention was already below 0.50 at 6 bits. Thus, outside the largest group, the transition is still complete by 5 bits, but it is not generally a single step from 6 to 5 as it is inside the largest group.

Whether these pass rates are meaningful depends on how many groups would need to pass by chance alone. With only 14 groups available, a one-sided test against a baseline rate of one half requires at least 11 successes to be considered statistically meaningful rather than incidental. The activation criterion clears this bound, passing in every group tested, while the single-step criterion does not, passing in only 5 of 14 groups. In short, the activation result generalizes cleanly beyond the dominant cluster, while the exact shape of the bit-width transition is less consistent, although it still moves in the same direction across all groups.

As a final check across all smaller groups pooled together, with groups resampled and then pairs resampled within groups, replication retention is 0.822 [0.513, 0.947] at 8 bits, 0.288 [0.155, 0.481] at 6, 0.000 at 5, 0.000 at 4, and 0.000 with 8-bit activations. This pooled analysis confirms the same qualitative pattern seen in the largest cluster, with near elimination by 5 to 6 bits and complete elimination under the activation quantizer. The small number of available groups means that no group-level regression is attempted here.

H DIFFUSION: FRONTIER AND THRESHOLD SENSITIVITY

Table 17: Replication retention at every weight precision, as the copy-detection threshold varies. The pre-specified threshold governs the claims in the body; the sensitivity is reported beside it rather than used to relabel the result. The elimination width is 5 bits at three of the four thresholds and moves to 4 bits at the loosest, which is what "5-bit suppresses, 4-bit destroys" means operationally.

detection threshold	n	8-bit	6-bit	5-bit	4-bit	3-bit	elimination width
0.60	137	0.949	0.745	0.577	0.000	0.000	4-bit
0.70	105	0.762	0.705	0.029	0.000	0.000	5-bit
0.75*	86	0.674	0.744	0.012	0.000	0.000	5-bit
0.80	79	0.658	0.759	0.000	0.000	0.000	5-bit

* pre-specified.

The distinction in the main text between suppression and elimination depends on where residual similarity falls relative to the detection threshold, not simply on whether a copy crosses that threshold. The continuous similarity values behind this distinction are therefore reported here. Median similarity

Table 18: The second diffusion architecture, with both interval conventions shown side by side. The cluster-level interval is the appropriate one and it is wide, because 159 captions resolve to 4 independent reference images. The caption-level interval is what the nominal sample size would have produced, and is included to show the size of the correction rather than to be quoted.

configuration	replication ret.	95% CI (independent)	95% CI (caption)	prompt fidelity	median sim.
full precision	1.000	[1.000, 1.000]	[1.000, 1.000]	1.000	0.873
8-bit	0.981	[0.400, 1.000]	[0.956, 1.000]	0.998	0.872
7-bit	0.969	[0.400, 1.000]	[0.937, 0.994]	0.996	0.871
6-bit	0.943	[0.000, 0.962]	[0.906, 0.975]	0.990	0.861
5-bit	0.447	[0.000, 0.457]	[0.365, 0.522]	0.987	0.740
4-bit	0.000	[0.000, 0.000]	[0.000, 0.000]	0.839	0.058

on the reproducing set is 0.901 at 6 bits, 0.638 at 5, and 0.068 at 4. Five-bit quantization pushes copies just below the detection threshold, while four-bit quantization destroys them outright. At a looser threshold, 5-bit replication rises to 0.577. This confirms that the 5-bit result is suppression near the detection boundary rather than clean elimination.

One ordering in these results initially appears unexpected and therefore requires a direct check. Retention at 8 bits (0.674) is lower than at 6 bits (0.744), which would be unexpected if it reflected the model rather than an artifact of image generation. The 8-bit images for the first seed of each caption had been reused from an earlier run of the same configuration, which raised this concern. Regenerating all 316 of them changes median similarity by only 0.002 and flips 2 copy decisions, while leaving 8-bit retention on the primary set unchanged. The ordering therefore reflects the generated images rather than the earlier reuse.

Table 17 gives the sensitivity analysis behind the distinction between suppression and elimination. At the pre-specified threshold and at both stricter thresholds, the elimination width is 5 bits. At the loosest threshold, it moves to 4 bits because 5-bit quantization leaves residual similarity just below the pre-specified threshold rather than well below it. Thus, 4-bit elimination holds across the tested thresholds, while 5-bit elimination depends on the threshold. The number of reproducing pairs also decreases as the threshold becomes stricter, which follows from the stricter definition of what counts as a copy.

Stable Diffusion 2.1-base provides an independent check on this pattern because it uses a different text encoder and was trained on a deduplicated corpus. Replication falls from 0.943 at 6 bits to 0.447 at 5, with quality at 0.987. The two architectures agree on where the transition starts and where quality collapses, but differ in how complete the transition is. The fact that this second model also shows substantial memorization despite its deduplicated corpus provides a second line of evidence against the duplication account in Section 4.

Table 18 shows this second architecture under two interval conventions side by side. The purpose is to make the correction visible rather than to present two competing estimates. The cluster-level interval is the statistically appropriate one, but it is nearly uninformative on its own because 159 captions resolve to only 4 independent reference images, with one image alone covering 155 of them. The caption-level interval is shown alongside it not because it should be treated as reliable, but because the difference between the two shows how much narrower the naive interval would appear. The point estimates are unchanged by the choice of unit. Only the intervals around them change.

At the pre-specified threshold, this second architecture does not reach the same separation criterion as the first because 5-bit replication remains at 0.447 rather than falling to zero. At a stricter threshold, the same underlying data reaches the criterion at the same bit width as the first architecture. The reported verdict follows the threshold fixed in advance, with this sensitivity reported alongside it rather than used to select a more favorable threshold afterward. What replicates across the two architectures is where the transition starts and where quality collapses. What does not replicate to the same degree is how complete the transition becomes.

I DIFFUSION: WEIGHT AND ACTIVATION ATTRIBUTION

The main text shows that 4-bit weights alone only partially remove replication, while the activation quantizer removes it entirely. A natural question is whether this difference depends on the detection

Table 19: Weight and activation quantization decomposed, at every detection threshold. The weight-only condition traverses the whole range as the threshold moves; the two conditions containing the activation quantizer do not move at all. Prompt fidelity and median similarity are at the pre-specified threshold. Retention is relative to the pairs that reproduce at full precision in this run at each threshold (86, 86, 84, 78 pairs).

condition	replication retention at threshold				fidelity	median sim.
	0.60	0.70	0.75	0.80		
4-bit weights only	0.651	0.221	0.143	0.026	0.9747	0.643
4-bit weights + 8-bit activations	0.000	0.000	0.000	0.000	0.9694	0.177
8-bit activations only, weights untouched	0.000	0.000	0.000	0.000	0.9949	0.322

Contingency over the complete set of 84 reproductions: 72 eliminated by both, 0 by the 4-bit weights only, 12 by the activation quantizer only, 0 by neither (Jaccard 0.857).

threshold. Table 19 addresses this by sweeping the threshold. The weight-only condition changes across nearly the full range as the threshold moves, from 0.651 to 0.026. Both conditions that include the activation quantizer remain at 0.000 at every threshold tested. The weight-only result therefore depends on the detection threshold, while the activation quantizer result does not.

The contingency reported in the table note provides a stronger check. Across the complete set of 84 reproductions, no pair is eliminated by 4-bit weights while remaining under the activation-only condition. No significance test is attached to this claim because the relevant independent unit is the reference image, and these reproductions span only three such units, with one discordant unit. A paired test at this scale would have little statistical power. The claim is instead a set-containment result over the complete observed set, and it holds across detection thresholds.

Two further checks address potential alternative explanations. The activation-only condition leaves the weights verifiably unchanged and still gives the highest quality among the quantized conditions tested. The replication reduction is therefore not accompanied by a quality loss in this comparison. The weights are also not completely inert in continuous terms. Median similarity is lower when 4-bit weights are added to the activation quantizer than when the activation quantizer is used alone, but this additional change does not eliminate any further pair.

Context and scope. Copy elimination at 4 bits in the weight-only sweep depends on quality collapse rather than on a separate privacy effect. After removing the quality-destroying component of the first deployment recipe considered, 4-bit group-wise weights at 0.9747 prompt fidelity allow 0.1429 of replications to return. Across the 84 primary reproductions, only 0 pairs are eliminated by 4-bit weights alone, compared with 12 eliminated by the activation quantizer alone (Jaccard 0.857). The secondary set shows the same attribution (0.0000 against 0.0891 over 101 pairs), so the result is not specific to the primary set. Full-precision weights with 8-bit activations do not save memory on their own, so the deployable recipe remains 4-bit weights combined with 8-bit activations. The attribution changes only in terms of which component provides the privacy-relevant effect. The activation scales used here are per-input-channel, which a standard integer matrix-multiply kernel cannot realize directly. Appendix I.1 tests a variant that such kernels can realize.

I.1 Clipping, rounding, and a kernel-realizable activation quantizer

Table 20: The 8-bit activation quantizer decomposed into its clipping and rounding components, with full-precision weights and with 4-bit group-wise weights. Replication retention is reported on the primary reproducing set and on the independent groups outside the largest one; median similarity is on the primary reproducing set; fidelity is the prompt-fidelity ratio on the quality set. All values at the pre-specified detection threshold.

	full-precision weights				4-bit group-wise weights			
activation condition	retention	retention, indep.	median sim.	fidelity	retention	retention, indep.	median sim.	fidelity
no activation quantization	0.977	1.000	0.92	1.000	0.140	0.000	0.64	0.975
static 8-bit, clipping and rounding (as recorded)	0.000	0.000	0.32	0.995	0.000	0.000	0.18	0.969
clipping only	0.000	0.000	0.33	1.001	0.000	0.000	0.21	0.975
rounding only, same step	0.000	0.000	0.49	0.996	0.000	0.000	0.53	0.971
dynamic per-token 8-bit	0.000	0.000	0.48	0.990	0.000	0.000	0.23	0.966
static 8-bit, 512-prompt calibration	0.000	0.000	0.30	0.988	0.000	0.000	0.61	0.963

The activation quantizer used throughout performs two separate operations on the input of each linear and convolutional layer. It clamps each channel to the range set by its calibration maximum and then rounds the result to an 8-bit grid over that range. This subsection separates these two operations to determine which one is responsible for eliminating replication. Clipping only applies the same clamp without rounding. Rounding only uses the same step size while widening the integer range so that no value is clamped. A dynamic variant recomputes a symmetric 8-bit scale for every token, or for every spatial position in convolutions, from its own maximum. This is a scheme supported by standard integer kernels and does not clamp. A final variant keeps the static quantizer but estimates its calibration maximum from 512 non-memorized prompts instead of 32. Before testing these variants, the recorded quantizer output was checked against clipping followed by rounding on real activations. The two matched exactly at all 14,382 layer calls, with only 0.035% of activation values falling outside the calibrated range.

The decision rule fixed in advance distinguished four possible outcomes. Clipping alone could be sufficient, rounding alone could be sufficient, both could be required together, or either component alone could be sufficient. The observed result is that either component alone is sufficient, under both weight settings and on both the primary set and the independent groups (Table 20). With full-precision weights, clipping only gives replication retention 0.000 at fidelity 1.001, rounding only gives 0.000 at 0.996, and the dynamic per-token quantizer gives 0.000 at 0.990. The components differ in how strongly they suppress similarity rather than in whether they suppress it. Median similarity on the primary set is 0.33 with clipping only, close to the recorded quantizer's 0.32, compared with 0.49 for rounding only and 0.48 for the dynamic quantizer. At the loosest detection threshold of 0.60, clipping only leaves 0.012 of primary pairs above the threshold, rounding only leaves 0.035, and the dynamic quantizer leaves 0.209. From 0.70 upward, every variant under either weight setting and on either set remains at or below 0.058. The larger calibration set tested here also does not restore replication at the pre-specified threshold.

One further check addresses whether heavy clipping could itself indicate memorized content. At full precision, across every layer call in the sampling trajectory, 0.069% of activation values exceed the calibrated maximum for memorized prompts, compared with 0.063% for web captions from the same source that do not reproduce and 0.015% for the quality set. Memorized and non-memorized web captions, therefore, overshoot the calibrated range at essentially the same rate, so clipping activity alone does not distinguish memorized from non-memorized content.