Beyond Forget Scores: Auditing Quantization-Robust LLM Unlearning with Q-ROU

Anonymous ACL submission

Abstract

LLM unlearning is often judged by whether a target answer disappears, but deployed models fail in richer ways: the same edit can collapse semantic neighbors, induce degenerate generation, reintroduce forgotten content after INT4/NF4 compression, or recover it after benign post-edit training. We therefore argue that post-deployment unlearning should be evaluated as an audit problem, and introduce Quantization-Robust Orthogonal Unlearning (Q-ROU), a compact operating point combining active semantic-neighbor retention with bounded KL-to-uniform forgetting plus SLUG/QuantNoise deployment stabilizers. On a 28-probe 3B multi-entity stress test, Q-ROU reaches 27/28 in FP16 and 28/28 in INT4, while non-AR baselines collapse neighbor retention. On a 65-probe expanded suite, Q-ROU achieves 25/25 target suppression with 17/20 neighbor retention in FP16; aggressive baselines either collapse neighbors or reach stronger token suppression only in degenerate generation regimes. Real NF4 checks preserve the main Q-ROU operating point, and LoRA-concentrated baselines show the key distinction: low-bit survival alone does not guarantee semantic-neighbor preservation. Standardized TOFU runs support the same failure-mode analysis, while MUSE news, same-work neighbors, and post-edit recurrence expose boundaries that target-only metrics hide. The resulting claim is limited to INT4/NF4 deployment audits: Q-ROU delivers selective, low-bit-stable suppression at a substantially stronger operating point than target-only metrics suggest.

1 Introduction

A high forget score can be the least informative part of a post-deployment unlearning result. Large language models (LLMs) memorize vast amounts of training data, including information that may need removal after deployment—for privacy regulations (*e.g.*, GDPR ‘‘right to be forgotten’’ (European Parliament and Council of the European Union, 2016)), copyright concerns, or the emergence of harmful content. The natural first test is whether the target answer disappears, but that test is easy to pass for the wrong reasons. An edit can damage adjacent facts, make generation incoherent, fail after 4-bit deployment, or become easy to reverse through a small benign update. We use *deployment audit* to name the stronger requirement: post-hoc unlearning should be judged by the failure modes it rules out, not by a single target-side score.

Existing gradient-based methods (GA, GradDiff, RepBend) share at least three deployment-critical limitations. First, they can incur substantial collateral damage to semantically adjacent knowledge: recent mechanistic analyses suggest that concept representations are distributed and can share nearby feature structure (Li et al., 2025), so updates aimed at a forget set can overwrite nearby capabilities. In our multi-entity stress test, the non-AR baselines GA, GradDiff, and RepBend collapse neighbor retention to 0/8. Second, they exhibit a token–generation tradeoff: aggressive objectives can suppress target tokens only by entering regimes of severe generation degeneration (Appendix D.4). Third, these methods can be brittle under deployment compression (Zhang et al., 2025; Abitante et al., 2026). Post-training INT4 quantization perturbs local decision margins and can cause partially forgotten information to resurface. Recent work (Abitante et al., 2026) demonstrates that standard full-parameter unlearning produces weight updates smaller than the quantization grid size, causing forgetting effects to be ‘‘masked’’ or ‘‘erased’’ during quantization—a phenomenon we term *forgetting regression*. To address these limitations, we propose **Quantization-Robust Orthogonal Unlearning (Q-ROU)**, a deployment-oriented operating point that combines bounded forgetting, explicit neighbor anchoring, restricted updates, and quantization-aware stabilization. *Active Retention*

(AR) uses a KL-divergence constraint to explicitly preserve the model's output distributions on neighbor-domain texts, directly targeting the semantic knowledge boundary rather than relying on generic retention data. *KL-to-uniform forgetting* bounds the forget objective by increasing target entropy, preventing optimization from entering degenerate collapse regimes while still suppressing target facts. *Selective Layer Update Grouping (SLUG)* confines parameter modifications to knowledge-critical layers identified through activation analysis, reducing the surface area for collateral damage. Finally, *QuantNoise injection* simulates quantization perturbations during training, steering the optimizer toward flat loss regions robust to post-deployment compression. Our primary contribution is a more demanding success criterion together with a method that reaches a strong deployment operating point under that criterion. Across the experiments below, methods that look successful on one axis routinely fail on another: target-strong baselines damage neighbors, low-rank updates survive quantization without preserving semantic boundaries, and longer benchmark calibration can improve target suppression at unacceptable utility cost. Q-ROU is designed as a coherent operating point for this conjunction, with its scope made explicit by the audit suite rather than hidden behind a single forget score. While KL-based constraints for generic retention have been explored (e.g., SCRUB (Kurmanji et al., 2023), SalUn (Fan et al., 2024)), Q-ROU explicitly anchors *semantic neighbors* (AR) and pairs this with a bounded KL-to-uniform forget objective to avoid degenerate regimes. Q-ROU combines a small core unlearning mechanism (AR + KL-to-uniform) with deployment-focused stabilizers (SLUG + QuantNoise) and lighter auxiliary regularizers (margin/orthogonality/EWC). This decomposition matters for interpreting the experiments: the paper does not claim that every term is equally essential, but that the layered combination reaches a stronger audited operating point than the compared alternatives. Concurrent work (Dorna et al., 2025) likewise benchmarks multiple unlearning metrics, reinforcing our concern that keyword-style or membership-based summaries alone can miss failures exposed by adversarial or generation-level audits. We evaluate deployment unlearning along four audit axes—selective forgetting, deployment compression, extraction robustness, and post-edit persistence—because each axis exposes a different way for an apparent unlearning success to fail in use. This positioning complements benchmark-standardization work such as TOFU (Maini et al., 2024), MUSE (Shi et al., 2025), RWKU (Jin et al., 2024), and OpenUnlearning (Dorna et al., 2025): our focus is to add deployment-specific failure modes that standard target/utility summaries can miss. It is motivated by recent benchmark-critique work showing that independent forget/retain evaluations can overstate progress and hide dependencies between what appears forgotten and what should remain intact (Thaker et al., 2025). It is also distinct from recent low-rank quantization-robust unlearning (Abitante et al., 2026), reasoning-trace unlearning (Yoon et al., 2025), and relearning-attack defenses (Xiao et al., 2026): Q-ROU is not optimized for a single one of these axes, but audited for their conjunction under an explicit scope. It is also distinct from contemporaneous anonymous work on operator-level linear-access auditing in LLM representations (Anonymous, 2026a): that paper asks whether linear evidence supports readout, control, and transfer claims in representation analysis, whereas Q-ROU targets post-deployment unlearning under suppression, neighbor preservation, low-bit stability, and bounded recurrence. Likewise, contemporaneous anonymous work on predicting collateral damage from Gradient Ascent unlearning (Anonymous, 2026b) studies which non-target concepts are most vulnerable under GA updates, framing the problem as a mechanistic ranking task; Q-ROU, by contrast, proposes a deployment-audited unlearning operating point and audit suite aimed at reducing such damage rather than predicting it.

2 Methodology

2.1 Problem Formulation

Let θ denote the parameters of a pre-trained LLM and $P_\theta(y|x)$ the conditional probability distribution over output tokens y given input x. We define three evaluation sets: $\mathcal{D}_{\text{target}}$ (knowledge to be removed), $\mathcal{D}_{\text{neighbor}}$ (semantically related knowledge to be preserved), and $\mathcal{D}_{\text{general}}$ (domain-independent knowledge to be preserved). The goal of selective unlearning is to find parameters θ^* such that, for each target context $x \in \mathcal{D}_{\text{target}}$ and each designated target token position $t \in S_f(x)$ with target token y_t, the model suppresses $P_{\theta^*}(y_t \mid x, y_{<t})$, while maintaining $P_{\theta^*}(y|x) \approx P_\theta(y|x)$ for inputs in $\mathcal{D}_{\text{neighbor}} \cup \mathcal{D}_{\text{general}}$. Here $\mathcal{D}_{\text{target}}$ denotes target prompts/contexts, and forgetting is enforced only

Audit axis	Evidence	Primary readout	Scope limit
Selective forgetting	28-probe core, 65-probe expanded suite, and focused 3-seed confirmations in FP16/INT4	T/N/G pass counts with token-probability audits	strongest evidence is on controlled custom suites; same-work neighbors remain the hardest frontier
Deployment compression	Matched FP16, fake INT4, and real NF4 checks on the same suites	T/N/G retention plus Δ_{zombie}	establishes low-bit stability at the tested operating points, not a universal PTQ guarantee
Extraction robustness	Six-protocol / 55-probe depth suite plus budgeted extraction sweeps, with 160-step extension for extraction only	protocol pass rates and attack-budget leakage	bounded black-box extraction families only
Post-edit persistence	Bounded jog audits on Llama, Qwen, and Phi with supporting 3-seed checks	worst post-jog target count	bounded recurrence evidence on the tested jog families

Audit design overview. Four deployment failure modes, their evidence, primary readouts, and scope limits.

on designated target-answer token positions within those contexts. Unlike data-partitioning approaches such as SISA training (Bourtoule et al., 2021), which achieve unlearning by retraining isolated model shards, Q-ROU directly modifies parameters of a deployed model to enable post-hoc knowledge removal without access to the original training data or retraining infrastructure. Additionally, we require post-quantization robustness. For neighbor and general sets, the output distributions should remain stable under a post-training quantization operator $Q(\cdot)$ (*e.g.*, INT4), meaning $P_{Q(\theta^*)}(y|x) \approx P_{\theta^*}(y|x)$. For the target set, the key requirement is the absence of forgetting regression after quantization. We formalize this target-side behavior through the *zombie delta* metric: $\Delta_{\text{zombie}} = \mathbb{E}[\delta(x, y)]$, where $\delta(x, y) = P_{Q(\theta^*)}(y \mid x) - P_{\theta^*}(y \mid x)$ and y denotes the target token. Positive Δ_{zombie} indicates *zombie knowledge* (forgotten information resurfacing under quantization), while negative values indicate that low-bit conversion does not cause forgetting regression and may further suppress borderline targets.

2.2 Q-ROU Framework

Q-ROU optimizes a composite objective function over K training steps, modifying only parameters in a selected subset of layers (SLUG). We separate the method into four layers of responsibility. **Core unlearning** consists of KL-to-uniform forgetting plus AR; **deployment stabilizers** consist of SLUG and QuantNoise; **auxiliary regularizers** consist of margin, orthogonality, and EWC; and **PTI** is a post-edit persistence extension evaluated separately rather than part of the base method. The objective therefore follows a two-tier training design. The **Core Q-ROU** objective handles the fundamental unlearning task:

$$\mathcal{L}_{\text{core}} = \lambda_f \mathcal{L}_{\text{forget}} + \lambda_r \mathcal{L}_{\text{retain}} + \lambda_a \mathcal{L}_{\text{active}} \quad (1)$$

To improve deployment robustness under low-bit quantization, we complement this with an auxiliary structural regularization suite:

$$\mathcal{L}_{\text{aux}} = \lambda_m \mathcal{L}_{\text{margin}} + \lambda_o \mathcal{L}_{\text{ortho}} + \mathcal{L}_{\text{EWC}} \quad (2)$$

The total loss is $\mathcal{L} = \mathcal{L}_{\text{core}} + \mathcal{L}_{\text{aux}}$. As the ablation results later show (Appendix C.6), the auxiliary suite ($\mathcal{L}_{\text{aux}}$) should be read as robustness support rather than as the conceptual core of the method: at larger scales and easier settings it has little effect on top-level pass counts, but it remains useful on harder boundary cases and in low-bit audits.

Forget and Retain Objectives. The forget loss uses a bounded KL-to-uniform objective that suppresses confident target-token predictions by increasing output entropy on target contexts: $\mathcal{L}_{\text{forget}} = \frac{1}{|S_f|} \sum_{t \in S_f} D_{\text{KL}}(P_\theta(\cdot|x, y_{<t}) \parallel U)$, where S_f denotes the set of target token positions and U is the uniform distribution over the vocabulary. Minimizing this objective is equivalent, up to the constant $\log |\mathcal{V}|$, to maximizing the output entropy $H(P_\theta)$ at the targeted positions. A passive retain loss preserves general model behavior using a KL-divergence constraint on general-domain texts: $\mathcal{L}_{\text{retain}} = \mathbb{E}_{x_g \sim \mathcal{D}_{\text{general}}}[D_{\text{KL}}\big(P_{\theta_0}(\cdot|x_g) \parallel P_\theta(\cdot|x_g)\big)]$, where θ_0 denotes the frozen pre-unlearning parameters.

Active Retention (AR). Active Retention is the core mechanism for preserving neighbor knowledge. Unlike traditional methods that rely on passive retaining via generic corpora (which the literature shows is insufficient against knowledge entanglement), AR explicitly defends the semantic frontier

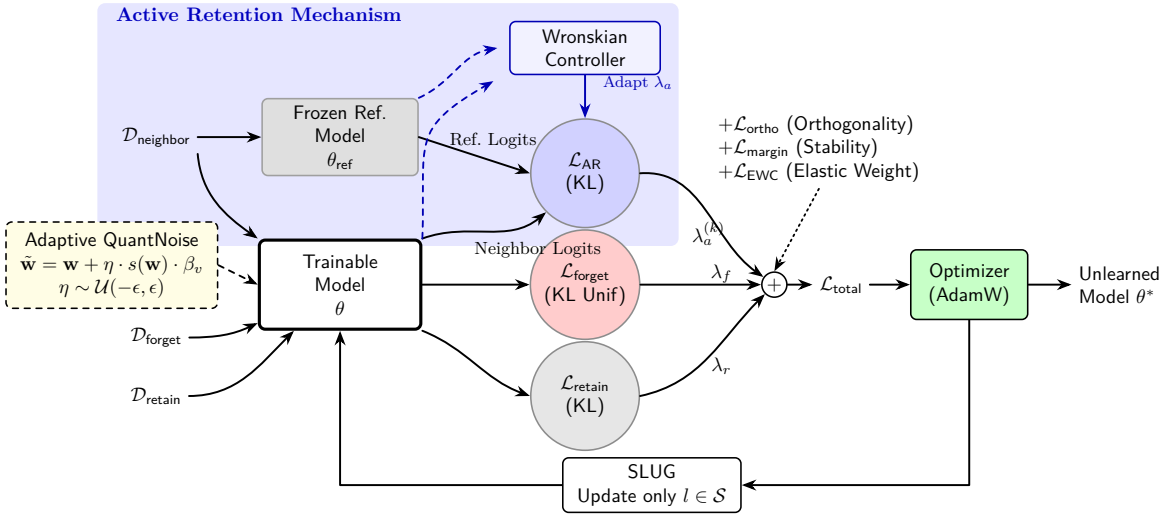

Figure 1: Q-ROU method overview. Target contexts optimize a bounded KL-to-uniform forget loss, semantic-neighbor and general retain sets are anchored to the frozen reference model through KL constraints, and the combined objective updates only the SLUG layer subset while QuantNoise and auxiliary regularizers support low-bit deployment stability. The Wronskian controller shown in the diagram is an optional diagnostic/adaptive variant; the main reported operating points use fixed coefficients unless otherwise stated.

around the forgotten domain by anchoring the output distribution on neighbor-domain texts to the frozen reference model:

$$\mathcal{L}_{\text{active}} = \mathbb{E}_{x_n \sim \mathcal{D}_{\text{neighbor}}} \Big[D_{\text{KL}}\big(P_{\theta_0}(\cdot|x_n) \parallel P_\theta(\cdot|x_n)\big) \Big] \tag{3}$$

In its standard formulation, Q-ROU uses a fixed penalty coefficient λ_a for this objective. Supporting analysis in the Appendix shows that the KL penalty bounds neighbor drift through a Fisher-geometric constraint (Proposition 2). This Fisher-geometric interpretation shares conceptual foundations with prior work on linearized fine-tuning (Achille et al., 2021), which uses local quadratic / Fisher-style approximations to stabilize fine-tuning updates. We use fixed λ_a for all reported operating points; adaptive variants are discussed only in the Appendix.

Selective Layer Update Grouping (SLUG). Motivated by knowledge-localization and activation-pattern research (Meng et al., 2022, 2023; Dai et al., 2022; Wang et al., 2025a), we restrict updates to knowledge-critical mid-to-deep layers, leaving shallow layers frozen. This design shares conceptual foundations with weight saliency approaches (Fan et al., 2024), which selectively update important parameters to minimize collateral damage—though SLUG operates at the layer granularity rather than individual weights. In our experiments, we update 5 layers per model; the selection procedure and exact layer sets are reported in the Appendix. All parameters outside $\mathcal{S}$ remain frozen during unlearning, providing inherent protection for knowledge stored in non-SLUG layers and reduced computational cost. This design shares conceptual similarities with LoRA-based unlearning (Abitante et al., 2026), which concentrates updates into low-rank subspaces to achieve quantization robustness. While LoRA achieves concentration through rank constraints on adapter matrices, SLUG achieves it through layer selection—both strategies amplify per-parameter update magnitudes to exceed quantization boundaries, though SLUG additionally leverages knowledge localization to minimize collateral damage. An Elastic Weight Consolidation (EWC) penalty (Kirkpatrick et al., 2017) further constrains these SLUG-layer parameters to remain close to their original values:

$$\mathcal{L}_{\text{EWC}} = \frac{\lambda_{\text{ewc}}}{2} \sum_{l \in \mathcal{S}} \sum_i F_i (w_i - w_i^{(0)})^2 \tag{4}$$

where $w_i^{(0)}$ denotes the original pre-unlearning weights and F_i is the Fisher information approximated from retain data.

QuantNoise and Regularization. During training, we inject uniform unstructured noise into trainable SLUG-layer weights: $\tilde{w} = w + \eta$, where $\eta \sim \mathcal{U}(-\epsilon_q, \epsilon_q)$. Intuitively, this is not ''freezing'' parameters; it actively flattens the local loss basin so INT4 rounding noise is absorbed instead

of undoing forgetting updates. We provide a formal derivation in Appendix A showing that independent coordinate noise translates to an implicit weighted Hessian-trace penalty, $\mathrm{tr}(H\Sigma_\eta)$, with the ordinary trace recovered only in the isotropic-noise special case. We further add a margin loss to maintain a safety buffer under quantization error, and an orthogonality penalty motivated by evidence that truth-related structure can occupy broad linear directions in representation space (Marks and Tegmark, 2023). Additional details and optional variants are reported in the Appendix.

Analytical Framework and Training Procedure. Two lightweight propositions formalize our design: (1) when the log-probability margin exceeds quantization error, pass/fail decisions are preserved under INT4 rounding (Appendix Proposition 1); (2) AR is the Lagrangian relaxation of constrained neighbor-drift minimization, where the KL constraint gives a local Fisher-geometric interpretation of neighbor protection: under the usual small-update quadratic KL approximation, $\Delta\theta^\top \mathcal{I}_{\mathrm{NBR}}\, \Delta\theta$ is controlled by the AR multiplier, biasing updates toward directions that leave neighbor outputs invariant (Appendix Proposition 2). Full statements, assumptions, proofs, and the complete training algorithm (Algorithm 1) are in Appendix A.

3 Experimental Setup

Models and Hyperparameters. We primarily evaluate on Llama-3.2-3B (Meta AI, 2024) for the custom-probe audits. We additionally include Llama-3.1-8B for TOFU standardized benchmarks and Qwen2.5-3B / Phi-3-mini-4k-instruct for supporting cross-architecture transfer and persistence checks discussed below and in the appendix. For the custom-probe experiments, we optimize Q-ROU with AdamW for up to 40 steps, enabling efficient post-deployment edits without prolonged retraining. For standardized TOFU and long-horizon extraction analyses, we additionally evaluate 160-step and longer schedules as specified in the corresponding sections and appendix. Unless otherwise specified, the primary loss coefficients on Llama-3.2-3B are set to $\lambda_f = 2$, $\lambda_r = 15$, $\lambda_a = 80$, $\lambda_m = 5$, $\lambda_o = 5$, and $\lambda_{\mathrm{ewc}} = 1$. For the final calibrated 40-step multi-entity operating point and the supporting LoRA / memory-jog audits, we increase only the forget coefficient to $\lambda_f = 20$ after a calibration sweep showed that $\lambda_f = 2$ under-suppresses the 13 target probes while leaving the retention coefficients unchanged (Appendix C.3).

Datasets and Evaluation Protocol. The primary stress test is multi-entity forgetting across two well-known fictional domains with distinct neighbor and general retain sets. We define $\mathcal{D}_{\mathrm{target}}$ as 18 derived target contexts/prompts spanning Harry Potter and LOTR; $\mathcal{D}_{\mathrm{neighbor}}$ as 11 derived contexts/prompts from Star Wars; and $\mathcal{D}_{\mathrm{general}}$ as 20 non-fiction derived contexts/prompts. Evaluation uses binary pass counts across three categories: target (TGT, 13 probes), neighbor (NBR, 8), and general (GEN, 7). A target probe passes if its predicted probability $P < 0.01$; neighbor/general pass if the probability ratio vs. baseline ≥ 0.30 (fallback: $P > 0.05$). We prioritize binary pass-count metrics over continuous averages to prevent a single severe knowledge leak from being masked by widespread but shallow forgetting. The strict $P < 0.01$ margin provides a conservative buffer against low-bit drift, aiming to keep target probabilities below $P < 0.05$ after precision degradation (Appendix B.3). Because thresholded counts can be brittle on small probe sets, we pair them with raw probabilities, zombie deltas, threshold-sensitivity sweeps, and Wilson/bootstrap intervals in the Appendix; no claim in the paper relies on a single thresholded table alone. Beyond the 28-probe core set, we evaluate on a 65-probe expanded set (paraphrased and indirect references), six adversarial depth protocols, and budgeted extraction.

Baselines and Precision Settings. We compare Q-ROU against Gradient Ascent (GA (Jang et al., 2023; Yao et al., 2024)), GradDiff (GradDiff (Maini et al., 2024)), and Representation Bending (RepBend (Yousefpour et al., 2025)). We also include comparisons against NPO (Zhang et al., 2024), RMU (Li et al., 2024), and GRU (Wang et al., 2025b) in our expanded evaluation. We also evaluate AR-augmented variants (*e.g.*, GA+AR) for mechanism isolation. To test the closest low-rank quantization-robust competitor, we add LoRA-concentrated baselines: LoRA-GA, LoRA-GA+AR, and LoRA-KL+AR, using rank-16 adapters over attention and MLP projections plus a small slug_down adapter control (Appendix C.3). All models are evaluated in FP16 precision and subsequently after post-training INT4 quantization (group size 32) to formally measure post-quantization performance drift.

4 Results

4.1 Standardized Benchmark: TOFU

To evaluate standardized unlearning, we run the TOFU benchmark (160 steps, robust profile) on both Llama-3.2-3B and Llama-3.1-8B, and then extend the hardest 8B forget10 case with long-horizon sweeps in the appendix.

Table 1: TOFU results under the standard 160-step protocol (single seed). Long-horizon 8B forget10 sweeps and tuned multi-seed hard-split confirmations are reported in Appendix C.4.

Subset	Model	Method	Forget@0.01	Forget@0.05	Gen Leak	Utility	Truth
forget01	Llama-3.2-3B	**Q-ROU**	**100.0%**	**100.0%**	**0.0%**	98.0%	100.0%
	Llama-3.2-3B	GA+AR	50.0%	50.0%	5.0%	99.0%	100.0%
	Llama-3.1-8B	**Q-ROU**	**100.0%**	**100.0%**	**0.0%**	**100.0%**	95.83%
	Llama-3.1-8B	GA+AR	50.0%	50.0%	7.5%	97.0%	95.83%
forget05	Llama-3.1-8B	**Q-ROU**	**100.0%**	**100.0%**	**0.0%**	69.0%	95.83%
	Llama-3.1-8B	GA+AR	11.0%	12.0%	34.5%	**98.0%**	**100.0%**
forget10	Llama-3.1-8B	**Q-ROU**	**72.0%**	**80.25%**	**7.0%**	82.0%	**100.0%**
	Llama-3.1-8B	GA+AR	5.75%	7.0%	36.75%	**100.0%**	91.67%

Q-ROU outperforms GA+AR on the forget/leakage metrics across all subsets. On forget01 and forget05, it achieves full 100% Forget@0.01 with 0% leakage. On forget10, the fixed-budget evaluation maintains a wide gap (72.0% vs. 5.75% Forget@0.01; 7.0% vs. 36.75% leakage). We report GA+AR in the main text as the strongest selective baseline under our retention-oriented criteria; additional baseline results are provided in the Appendix. At the same fixed 160-step budget, 8B utility on forget05/forget10 drops relative to GA+AR, reflecting a forgetting–utility trade-off at harder targets; we analyze longer-horizon behavior and ablation sensitivity in the Appendix. Longer-horizon 8B sweeps (Appendix C.4) reveal the 160-step forget10 checkpoint is materially under-saturated: Q-ROU enters a high-performing regime by steps 288–416. End-to-end MUSE/WMDP runs complete without instability and clarify where longer-horizon calibration matters most. WMDP and MUSE books act as stable high-utility checks. MUSE news is the hardest external setting: the default 40-step row is under-saturated, a 120-step short-sequence calibration sweep reaches positive KnowMem reduction at 0.9934 utility, and a full-sequence 120/160/240-step continuation preserves high utility while showing that the benchmark's full-sequence news forget metric remains difficult. We therefore use MUSE news as a boundary and calibration diagnostic, while the custom multi-entity audits remain the cleanest place to isolate Q-ROU's selective-forgetting mechanism (Appendix B.5).

4.2 Audit Axis 1: Selective Forgetting

Beyond standardized benchmarks, we conduct the main diagnosis using a custom multi-entity suite designed to stress selective forgetting rather than target suppression alone. Table 2 reports the core operating points. In the multi-entity setting, Q-ROU reaches 27/28 in FP16 and 28/28 in INT4, outperforming the strongest baseline family under the same evaluation protocol. Notably, INT4 performance matches or slightly exceeds FP16 on this operating point, indicating that low-bit conversion does not induce forgetting regression here.

Table 2: Core pass-count results on Llama-3.2-3B (28-probe core set).

Setting	FP16 (T/N/G)	INT4 (T/N/G)
Single (Q-ROU)	23/23 (7/7, 9/9, 7/7)	23/23 (7/7, 9/9, 7/7)
Multi (Q-ROU)	27/28 (13/13, 7/8, 7/7)	28/28 (13/13, 8/8, 7/7)
Multi (GA+AR)	22/28 (7/13, 8/8, 7/7)	22/28 (7/13, 8/8, 7/7)

Baseline GA/GradDiff/RepBend runs suffer *near-complete collapse* of neighbor retention (falling to 0/8) and damage general behavior. The key distinction is not target suppression alone, but whether suppression remains selective and stable under deployment precision changes. This same point appears in the matched LoRA audit: broad LoRA-GA survives low-bit conversion but still collapses semantic neighbors, whereas AR/KL LoRA variants preserve neighbors only by under-forgetting. We therefore treat low-rank concentration as supporting evidence for the compression story, not as a substitute for neighbor-aware suppression (Appendix C.3).

To test broader generalization, we scale up to a 65-probe expanded set (paraphrased and indirect references) and compare against seven baselines (Table 3). Appendix Table 43 then reproduces the focused five-method selective subset used for the detailed 40-step operating-point discussion. Here we report only the converged 40-step operating points used for the main claim. Appendix Table 27 gives the 20/40/60 step sweep and shows the step dependence directly: Q-ROU is still under-suppressed at 20 steps, cleanly overtakes GA+AR by 40 steps, and remains ahead at 60 steps even as GA+AR partially closes the target gap. Appendix Table 44 then confirms that this 40-step ordering is not a single-seed artifact: across three seeds, Q-ROU stays at 25/25 target suppression in both FP16 and INT4, while GA+AR remains fixed at 13/25 target pass. Appendix D.4 further reports learning-rate, step-count,

and AR-weight controls for the strongest selective baseline family. Extending aggressive baselines without additional stabilizers often degrades generation, which we treat as a failure mode rather than a valid operating point. Q-ROU is the only method achieving 25/25 target removal in FP16 while still retaining 17/20 neighbor pass counts, and it remains the only method with 25/25 target removal together with ≥15/20 neighbor retention under INT4. All other high-target methods (GA, RepBend, RMU, GRU) achieve ≤1/20 neighbor retention. NPO preserves neighbors (15/20) but achieves only 15/25 target removal. Because aggressive GA-family settings can trade token-level suppression for generation degeneration, we interpret this pass-count table jointly with the operating-point audit in Table 5 and the generation-level analyses in Appendix D.4. Additional plug-in baseline experiments are reported in the Appendix.

Table 3: Seven-method baseline comparison on the 65-probe expanded set (Llama-3.2-3B). Q-ROU and GA+AR are reported at their 40-step coherent operating points; other baselines follow their published recommended budgets (20 steps). The 20/40/60 sweep is reported separately in Appendix Table 27. Extending aggressive baselines beyond 20 steps without stabilizers causes generation degeneration (Appendix D.4), which we treat as a failure mode. Q-ROU uniquely combines complete target removal with robust retention across FP16 and INT4; aggressive baselines (GA, RepBend, RMU, GRU) collapse neighbor knowledge. FP and I4 denote FP16 and INT4 precision respectively.

Method	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z
Original model	5/25	20/20	20/20	5/25	18/20	20/20	−0.055
Q-ROU (40 steps)	**25/25**	**17/20**	**20/20**	**25/25**	**16/20**	**18/20**	−0.018
GA+AR	13/25	18/20	20/20	13/25	15/20	19/20	−0.026
GA	25/25	0/20	0/20	25/25	0/20	0/20	+0.000
RepBend	25/25	0/20	12/20	25/25	0/20	11/20	−0.013
RMU	25/25	0/20	13/20	25/25	0/20	13/20	−0.002
NPO	15/25	15/20	18/20	16/25	15/20	17/20	−0.027
GRU	25/25	0/20	1/20	25/25	0/20	1/20	+0.000

Three-seed confirmation on the 28-probe suite and threshold sweeps on the larger audit support the same reading: the joint T/N/G advantage is stable across seeds, and the target-side Q-ROU conclusion remains stable around the default cutoff while GA+AR shifts more sharply with τ_f (Appendix Tables 44 and 50).

The same selective-forgetting story becomes sharper under a stricter neighbor definition. Splitting the original neighbor set into alias, same-work, inter-domain, and general strata shows that same-work retention is the hardest frontier, not a contradiction of the main result. On this harder audit, the default Q-ROU point still deletes the target completely, while a fine-grained AR operating point raises same-work retention from 9/36 to 19/36 in FP16 and to 18/36 under three-seed BnB NF4 (Appendix Tables 11 and 58). The larger 81-probe copyright-like benchmark shows the same pattern: Q-ROU+FG-AR lifts same-work retention to 14.7/24 in FP16 and 12.7/24 under real NF4 while staying target-strong (Appendix Table 12).

4.3 Audit Axis 2: Deployment Compression and Baseline Fairness

Ablation of Adaptive Mechanisms. Four-way ablation of the adaptive components (Standard, Adaptive QN Only, Adaptive AR Only, Full Adaptive) reveals only small pass-count differences on the 3B expanded set (Appendix Table 36). This is expected: the adaptive controller is designed to mitigate *tail* deployment failures (e.g., threshold brittleness and instability under precision changes) rather than to monotonically improve mean pass counts, so its value is better interpreted alongside robustness audits (Appendix C.12). Notably, Q-ROU's multi-component design addresses the 4-bit quantization forgetting regression challenge identified by (Abitante et al., 2026): while their LoRA-based recipe preserves forgetting through low-rank update concentration, Q-ROU combines SLUG (layer-wise concentration), QuantNoise (loss-landscape flattening), and structural regularizers to stabilize the selective-forgetting operating point beyond update magnitude alone.

Active Retention Transplants and Feature Removals. The transplant results (Table 4) are an *early-phase mechanism diagnostic*: we inspect a deliberately under-saturated 20-step checkpoint to separate AR's boundary-anchoring role from eventual target saturation. At that matched early point, removing AR from Q-ROU already collapses neighbor retention from 8/8 to 2/8 (FP16) and 0/8 (INT4) while accelerating target removal. Conversely, transplanting AR into aggressive baselines prevents neighbor collapse (0/8→8/8), but this *neighbor-only* sufficiency does not resolve the broader deployment failure modes (token–generation tradeoff and low-bit stability) that separate Q-ROU from GA+AR at the coherent 40-step operating points (Table 3; Appendix D.4).

Removing QuantNoise produces negligible FP16 changes but reveals why low-bit robustness matters: FP TGT shifts from 7/13 to I4 TGT 9/13 (+2) while

NBR stays at 8/8. In other words, the target boundary is still moving under quantization even when the headline count improves. AR bounds collateral damage, while SLUG and QuantNoise make the operating point less sensitive to that low-bit perturbation. EWC and the remaining regularizers are better read as auxiliary support than as the conceptual core.

Table 4: Early-phase AR mechanism isolation and ablation (Llama-3.2-3B, multi-entity, 20 steps). This matched checkpoint separates mechanism roles before convergence. The full 40-step operating-point separation is reported in Table 3 and Appendix Table 18.

Configuration	FP16 (T/N/G)	INT4 (T/N/G)
GA	13/13, 0/8, 0/7	13/13, 0/8, 0/7
GA+AR	4/13, 8/8, 7/7	6/13, 8/8, 7/7
Q-ROU	4/13, 8/8, 7/7	4/13, 8/8, 7/7
Q-ROU −AR	13/13, 2/8, 7/7	13/13, 0/8, 7/7
Q-ROU −QuantNoise	7/13, 8/8, 7/7	9/13, 8/8, 7/7

Scale-Dependent Component Contribution. Component ablation at 3B scale reveals a different sensitivity landscape from 0.5B. Removing Orthogonality, EWC, or Margin individually yields *negligible* pass-count changes ($|\Delta\bar{p}_{\mathrm{TGT}}| \leq 0.013$), contrasting with 0.5B where the same removals cause visible degradation (Appendix C.6). The correct reading is that these terms are auxiliary guards whose effect is most visible on harder or more deployment-sensitive slices, including loss-landscape sharpness and threshold sensitivity (Appendix D.16; Appendix C.12).

The main takeaway is visible in Tables 3 and 6: Q-ROU's target suppression does not regress after fake INT4 or real NF4 conversion, whereas target-strong baselines achieve this only by sacrificing retention or generation quality.

This resolves the main baseline-fairness ambiguity: GA+AR can be tuned toward stronger token suppression, but the matched 13/13 row is not deployment-valid because its apparent ''no leak'' generation is achieved by degeneration rather than by coherent non-target behavior (Appendix D.4). Accordingly, the GA+AR rows used elsewhere in the main text are the strongest non-degenerate operating points from this audit rather than the best token-only checkpoints.

Real NF4 supports the same reading as the INT4 audit: the deployment ordering survives low-bit conversion, even though small same-work drifts remain on the hardest slices.

Table 5: Baseline-fairness audit on the 28-probe core suite (Llama-3.2-3B, FP16, 40 steps). High-LR GA+AR reaches 13/13 only via repetitive degeneration.

Method	LR / λ_a	TGT	NBR	GEN	Deg.	Reading
Q-ROU	main	**13/13**	7/8	7/7	**0/7**	coherent
GA+AR	10^{-4}/160	9/13	8/8	7/7	0/7	coherent; still leaking
GA+AR	5×10^{-4}/40	13/13	8/8	7/7	7/7	repetitive collapse

Table 6: Main real-quantization check on the Llama-3.2-3B multi-entity suite. Q-ROU preserves the operating-point ordering under fake INT4 and deployed NF4, whereas GA+AR remains under-suppressed.

Method	FP16	Fake INT4	NF4
Q-ROU 40s	**13/13, 7/8, 7/7**	**13/13, 8/8, 7/7**	**13/13, 7/8, 7/7**
GA+AR 40s	7/13, 8/8, 7/7	7/13, 8/8, 7/7	7/13, 8/8, 7/7

The matched low-rank audit shows the same separation: broad LoRA-GA survives low-bit conversion but still collapses neighbors, while LoRA-GA+AR and LoRA-KL+AR preserve neighbors only by under-forgetting (Appendix C.3).

4.4 Audit Axis 3: Extraction Robustness

Across six depth protocols, Q-ROU achieves 55/55 suppression in both FP16 and INT4, while GA+AR leaves exploitable vulnerabilities. Chain-of-Thought and budgeted extraction show the same pattern: the 40-step point still leaves a 33.3% INT4 budget channel, but the 160-step extension closes all tested budgets to 0.0% (Appendix Tables 45, 52, 53).

4.5 Audit Axis 4: Post-Edit Persistence

Under bounded non-target memory-jog updates, PTI acts as a useful Llama-family hardening step: the worst post-jog target count improves from 7/13 to 12/13 in the narrow audit, from 6/13 to 8/13 in the broader five-mode audit, and from 5/13 to 10/13 under matching deployed-NF4 evaluation. Qwen and Phi remain feasible after architecture-aware retuning but do not show the same count-level PTI gain (Appendix C.3).

5 Conclusion

Q-ROU is best read as deployment-audited suppression rather than a one-number forget score. Across the Llama-family audits, it suppresses targets while preserving semantic neighbors and avoiding low-bit regression. It reaches a stronger joint operating point than the tested baselines.

Limitations

Our strongest evidence comes from controlled custom-probe audits on sub-3B Llama-family models, with standardized TOFU runs extending to 8B. The 28- and 65-probe suites are designed to isolate failure modes cleanly, and the benchmark results show that operating-point sensitivity remains real, especially on MUSE news. The hardest remaining selectivity issue is same-work retention. Fine-grained audits show clear gains from stronger neighbor anchoring, but even the more protective AR operating point does not fully preserve all same-work non-target facts. Hyperparameters also remain architecture-dependent. Architecture-aware transfer to Qwen and Phi is feasible, but the strongest PTI recurrence gains appear on Llama-family models. Benign post-deployment updates can still reactivate a subset of target probes after base unlearning, which is why we report PTI as a useful hardening extension rather than fold it into the base method. Our low-bit robustness evidence covers fake INT4 and real NF4; broader GPTQ/AWQ-style post-training quantizers remain an important next check. Finally, our adversarial CoT and budgeted extraction probes are bounded elicitation/extraction audits with oracle risk-ranked variants, not a full reasoning-model unlearning benchmark. R-TOFU-style step-wise trace evaluation is a natural next extension, especially for models whose safety-critical behavior depends on hidden or explicit reasoning traces.

Ethical Considerations

This work targets safer post-deployment models by enabling selective removal of harmful, sensitive, or copyrighted knowledge without requiring expensive full retraining. However, we acknowledge the dual-use nature of unlearning frameworks: the same mechanisms could theoretically be misused to conceal model provenance, selectively suppress legitimate information, or deliberately introduce knowledge blind spots. We mitigate these risks by emphasizing auditable evaluation protocols, explicit reporting of retention trade-offs, and transparent documentation of unlearning scope and failure modes. All experiments rely on fictional-domain evaluation sets derived from widely known franchises, benchmark-style settings, and a supplementary audit of publicly documented biographical facts about public figures only. We do not redistribute copyrighted source text. No private personal information is collected or released, and the public-biographical audit is limited to public, non-sensitive facts rather than a redistributable personal-information dataset.

To support independent evaluation, the paper and appendix report the model variants, optimization settings, probe construction rules, thresholding criteria, quantization settings, and hardware conditions needed to interpret every table and figure. The manuscript is intended to stand on its own: each reported claim is tied to a directly described protocol rather than to an external code release. AI assistants were used only in a limited supporting role for English expression polishing, routine coding assistance, lightweight scripting, figure drafting, and internal manuscript diagnostics. All substantive research decisions, theoretical claims, experiment design, result interpretation, and final manuscript wording were reviewed and approved by the authors, who take full responsibility for the paper.

References

João Vitor Boer Abitante, Joana Meneguzzo Pasquali, Luan Fonseca Garcia, Ewerton de Oliveira, Thomas da Silva Paula, Rodrigo C. Barros, and Lucas S. Kupssinskü. 2026. Quantization-robust LLM unlearning via low-rank adaptation. *arXiv preprint*.

Alessandro Achille, Aditya Golatkar, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. 2021. Lqf: Linear quadratic fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15729–15739.

Anonymous. 2026a. Linear access in llm representations: An audit of readout, control, and transfer. Under review.

Anonymous. 2026b. Which neighbors suffer? predicting collateral damage from gradient ascent in llm unlearning. Under review.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. Wiley-Interscience, Hoboken, NJ.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502. Association for Computational Linguistics.

Chandler Davis and W. M. Kahan. 1970. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, J. Zico Kolter, Zachary C. Lipton, and Pratyush Maini. 2025. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems 38 (Datasets and Benchmarks Track)*.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–487.

European Parliament and Council of the European Union. 2016. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union. OJ L 119, 4.5.2016, pp. 1–88.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2024. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, pages 53643–53673.

Jiahui Geng, Qing Li, Herbert Woisetschlaeger, Zongxiong Chen, Fengyu Cai, Yuxia Wang, Preslav Nakov, Hans-Arno Jacobsen, and Fakhri Karray. 2025. A comprehensive survey of machine unlearning techniques for large language models. *arXiv preprint*.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y. Zou. 2019. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems 32*.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9309.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems 36*, pages 17643–17668.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. 2025. Intrinsic test of unlearning using parametric knowledge traces. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19513–19535. Association for Computational Linguistics.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Rwku: Benchmarking real-world knowledge unlearning for large language models. In *Advances in Neural Information Processing Systems 37 (Datasets and Benchmarks Track)*, pages 98213–98263.

Dhruva Karkada, Daniel J. Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. 2026. Symmetry in language statistics shapes the geometry of model representations. *arXiv preprint*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2023. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems 36*, pages 1957–1987.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, and 27 others. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 28525–28550. PMLR.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. 2025. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7:181–194.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. 2024. Eight methods to evaluate robust unlearning in llms. *arXiv preprint*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J. Zico Kolter. 2024. TOFU: A task of fictitious unlearning for LLMs. In *Proceedings of the First Conference on Language Modeling (COLM 2024)*.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Meta AI Blog.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. 2021. A white paper on neural network quantization. *arXiv preprint*.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2025. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46.

Vaidehi Ramesh Patil, Peter Hase, and Mohit Bansal. 2024. Can sensitive information be deleted from LLMs? objectives for defending against extraction attacks. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, pages 45497–45514.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Akira Sakai and Yuma Ichikawa. 2026. Sign lock-in: Randomly initialized weight signs persist and bottleneck sub-bit model compression. *arXiv preprint*.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. 2025. Muse: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, pages 27797–27818.

Ilia Shumailov, Jamie Hayes, Eleni Triantafillou, Guillermo Ortiz-Jimenez, Nicolas Papernot, Matthew Jagielski, Itay Yona, Heidi Howard, and Eugene Bagdasaryan. 2024. Ununlearning: Unlearning is not sufficient for content regulation in advanced generative ai. *arXiv preprint*.

G. W. Stewart and Ji-guang Sun. 1990. *Matrix Perturbation Theory*. Academic Press, Boston.

Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith. 2025. Position: LLM unlearning benchmarks are weak measures of progress. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 520–533.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE.

Haoyu Wang, Zhuo Huang, Xiaolong Wang, Bo Han, Zhiwei Lin, and Tongliang Liu. 2026. Megu: Machine-guided unlearning with target feature disentanglement. *arXiv preprint*.

Yudong Wang, Damai Dai, Zhe Yang, Jingyuan Ma, and Zhifang Sui. 2025a. Exploring activation patterns of parameters in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25416–25424.

Yue Wang, Qizhou Wang, Feng Liu, Wei Huang, Yali Du, Xiaojiang Du, and Bo Han. 2025b. GRU: Mitigating the trade-off between unlearning and retention for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64690–64710. PMLR.

Ronald L. Wasserstein and Nicole A. Lazar. 2016. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

James H. Wilkinson. 1963. *Rounding Errors in Algebraic Processes*. Prentice-Hall, Englewood Cliffs, NJ. Reprinted by Dover, 1994.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for

large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

Zeguan Xiao, Xuanzhe Xu, Yun Chen, Yong Wang, Jian Yang, Yanqing Hu, and Guanhua Chen. 2026. Robust LLM unlearning against relearning attacks: The minor components in representations matter. *arXiv preprint*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8403–8419. Association for Computational Linguistics.

Sangyeon Yoon, Wonje Jeung, and Albert No. 2025. R-TOFU: Unlearning in large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5239–5258, Suzhou, China. Association for Computational Linguistics.

Ashkan Yousefpour, Taeheon Kim, Ryan Sungmo Kwon, Seungbeen Lee, Wonje Jeung, Seungju Han, Alvin Wan, Harrison Ngan, Youngjae Yu, and Jonghyun Choi. 2025. Representation bending for large language model safety. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24073–24098. Association for Computational Linguistics.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Proceedings of the First Conference on Language Modeling (COLM 2024)*.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. 2025. Catastrophic failure of LLM unlearning via quantization. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, pages 74925–74948.

Appendix

This appendix provides full technical details, extended analyses, and supplementary experiments.

A Extended Methodology Details

The core Q-ROU framework, problem formulation, and loss objectives are defined in the main text (Section 2). This section supplies supporting analytical results and the complete training algorithm.

A.1 Design Rationale: Minimal Analytical Results

The following statements are lightweight analytical consequences of the optimization design and evaluation thresholding scheme. They provide sufficient-condition diagnostics for the audited operating point and clarify which parts of the deployment story are analytical versus empirical.

Analytical scope. Proposition 1 is a threshold-stability diagnostic under bounded log-probability perturbations, not universal robustness to all quantizers. Proposition 2 interprets AR as a Lagrangian relaxation of neighbor-drift control, not a global Fisher barrier. Theorem 1 explains QuantNoise as a noise-weighted Hessian-trace penalty, not a guarantee for every production quantizer. Theorem 2 bounds one matrix-subspace rotation channel under small updates, not semantic preservation from linear algebra alone. Theorem 4 is a Fano-style scaling statement under an explicit output-channel leakage assumption, not protection against side-channel, parameter-access, or unrestricted paraphrase attacks. Theorem 5 is a noised-model indistinguishability calculation under a stated retrained-distance assumption, not an empirical retraining-equivalence result.

Proposition 1 (Threshold Stability under Bounded Quantization Perturbation). Let τ be the target-pass threshold ($\tau = 0.01$), and define the log-probability margin for a target case (x, y) as $m(x, y) = \log \tau - \log P_{\theta^*}(y|x)$. Assume quantization induces bounded log-probability perturbation $|\log P_{Q(\theta^*)}(y|x) - \log P_{\theta^*}(y|x)| \le \epsilon_q$. If $m(x, y) > \epsilon_q$, the pass/fail decision for that case is unchanged after quantization. **Implication.** The margin term and QuantNoise jointly increase target-side separation from the threshold, reducing flip risk. *Caveat:* The bounded log-probability perturbation assumption holds most naturally in the neighbor/general retention regime where probabilities are moderate. In the target regime where $P_{\theta^*}(y|x) \ll 1$, the effective perturbation ϵ_q in log-probability space may grow—since $\log P$ becomes highly sensitive to small probability changes near zero—potentially weakening this conclusion for cases closest to threshold.

Proposition 2 (AR as Constrained Neighbor-Drift Control). Consider the constrained problem:

$$\begin{aligned} &\min_{\theta} \ \mathcal{L}_{\text{forget}} + \lambda_r \mathcal{L}_{\text{retain}} + \mathcal{R}(\theta) \\ &\text{s.t. } \mathbb{E}_{x_n} D_{\text{KL}}\big(P_{\theta_0}(\cdot|x_n) \| P_{\theta}(\cdot|x_n)\big) \le \xi \end{aligned} \tag{5}$$

where $\mathcal{R}$ aggregates EWC, orthogonality, and margin terms. Our penalized objective (combining Eqs. 1 and 2) is the Lagrangian relaxation of Eq. 5 with multiplier λ_a.

Implication. Increasing λ_a tightens the optimization toward smaller neighbor-output drift, confirming AR is a direct, tunable knob for neighbor deviation control.

Remark (Fisher-Geometric Interpretation). Under the standard small-update expansion of KL divergence around the frozen reference model, the KL constraint in Eq. 5 has a Fisher-geometric local form involving the neighbor-aggregated Fisher information matrix $\mathcal{I}_{\text{NBR}}$: $\Delta\theta^\top \mathcal{I}_{\text{NBR}}\, \Delta\theta$ is controlled by the active-retention multiplier and the feasible neighbor-drift budget. This encourages parameter updates toward the approximate null space of $\mathcal{I}_{\text{NBR}}$—directions that leave neighbor output distributions approximately invariant. AR therefore supplies a local parameter-space interpretation of neighbor protection without requiring explicit concept localization (Hase et al., 2023). This Fisher null-space reading is the *primary* theoretical justification for AR, but it should be read as a local approximation around the fixed reference anchor rather than as a global structural barrier. The theoretical foundation for using Fisher information to guide selective forgetting was established by (Golatkar et al., 2020), which formalized unlearning as achieving weight-space indistinguishability from a model retrained without the forget set. As a supplementary empirical tool, we additionally introduce a Wronskian-based collision-risk signal computed from short-window AR surrogate fits of hidden-state trajectories, providing lightweight online early-warning capability (detailed below).

A.2 Training Procedure

Algorithm 1 summarizes the complete Q-ROU training procedure. All trainable parameters in SLUG

layers are cast to FP32 for optimization stability (even when the model backbone uses FP16), and cast back after training.

Algorithm 1 Q-ROU Training Procedure

Require: Pre-trained model θ, forget set $\mathcal{D}_f$, neighbor set $\mathcal{D}_n$, general set $\mathcal{D}_g$, SLUG layers $\mathcal{S}$, steps K

Compute original logits: $\{P_\theta(\cdot|x)\}_{x\in\mathcal{D}_n\cup\mathcal{D}_g}$
Compute truth directions $\{v_l\}_{l\in\mathcal{S}}$ via PCA on TruthfulQA
Freeze all parameters except $\{W_l\}_{l\in\mathcal{S}}$
Cast trainable params to FP32
for $k = 1, \ldots, K$ **do**
    Sample batches $B_f \subset \mathcal{D}_f$, $B_n \subset \mathcal{D}_n$, $B_g \subset \mathcal{D}_g$
    Set $\lambda_a^{(k)} \leftarrow \lambda_a$
    if optional adaptive AR variant enabled **then**
        Fit short-window AR responses from target/neighbor hidden trajectories and compute Wronskian risk ρ_k
        Adapt $\lambda_a^{(k)}$ based on the surrogate collision signal
    end if
    if adaptive QuantNoise variant enabled **then**
        Fetch optimizer 2nd moment v_l and compute adaptive scale $\beta_v \propto (v_l/\bar{v})^\beta$
        Inject noise: $\tilde{W}_l \leftarrow W_l + \eta \cdot s(W_l) \cdot \beta_v$, $\eta \sim \mathcal{U}(-\epsilon_q, \epsilon_q)$
    else
        Inject standard QuantNoise into trainable SLUG-layer weights
    end if
    Compute $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{core}} + \mathcal{L}_{\text{aux}}$ using $\lambda_a^{(k)}$
    Update $\{W_l\}_{l\in\mathcal{S}}$ via Adam optimizer
end for
Cast trainable params back to original precision
return modified model θ^*

A.3 Analytical Diagnostics for Active Retention and QuantNoise

This section provides supporting analytical diagnostics for the two core optimization mechanisms in Q-ROU: QuantNoise training and Active Retention. The goal is to clarify why these mechanisms are plausible under explicit assumptions, not to present a stand-alone theory of irreversible unlearning.

A.3.1 QuantNoise and its Adaptive Variant Induce Flat Maxima

Let $w \in \mathbb{R}^n$ be the parameters of the SLUG layers. The standard QuantNoise mechanism injects uniform noise $\eta_i \sim \mathcal{U}(-\epsilon_0, \epsilon_0)$ into the weights during each forward pass. The Adaptive QuantNoise extension directionally scales the perturbation bounds via $\epsilon_{q,i} = \epsilon_0\beta_{v,i}$, where $\beta_{v,i} \propto (v_i/\bar{v})^\gamma$ is derived from the Adam optimizer's second moment v and $\gamma > 0$ is a tunable exponent. Conceptually, this does not freeze the model; rather, it broadens the local optimum so that quantization perturbations are absorbed within a flatter basin. This design philosophy—regularizing during training to facilitate post-hoc modifications—shares conceptual foundations with (Thudi et al., 2022), which proposed standard deviation (SD) loss to constrain weight trajectories and reduce verification error during unlearning. While SD loss targets weight-space proximity to retrained models, QuantNoise targets loss-landscape flatness for quantization robustness. The importance of activation-aware quantization strategies is further emphasized by (Xiao et al., 2023), which demonstrated that activation outliers are the primary bottleneck for low-bit quantization in LLMs, motivating our focus on loss-landscape geometry rather than weight magnitude alone. Comprehensive surveys of quantization techniques (Nagel et al., 2021) establish the theoretical foundations for post-training quantization (PTQ) and quantization-aware training (QAT), providing the broader context within which QuantNoise operates as a specialized regularization strategy for unlearning robustness.

Theorem 1 (QuantNoise Regularizes a Weighted Hessian Trace). Let $\mathcal{L}(w)$ be the loss function, assumed four times continuously differentiable. Let $\eta \in \mathbb{R}^n$ with independent components $\eta_i \sim \mathcal{U}(-\epsilon_{q,i}, \epsilon_{q,i})$, where $\epsilon_{q,i} = \epsilon_0\beta_{v,i}$ (with $\beta_{v,i} = 1$ for standard QuantNoise). Then the expected perturbed loss satisfies:

$$\mathbb{E}_\eta[\mathcal{L}(w+\eta)] = \mathcal{L}(w) + \frac{\epsilon_0^2}{6}\sum_{i=1}^{n}\beta_{v,i}^2\nabla_{ii}^2\mathcal{L}(w) + O(\epsilon_0^4) \tag{6}$$

Proof. We perform a second-order Taylor expansion of $\mathcal{L}(w+\eta)$ around w:

$$\mathcal{L}(w+\eta) = \mathcal{L}(w) + \nabla\mathcal{L}(w)^\top\eta + \frac{1}{2}\eta^\top H\eta + O(\|\eta\|^3) \tag{7}$$

where $H = \nabla^2 \mathcal{L}(w)$ is the Hessian matrix. Taking the expectation over η:

$$\mathbb{E}_\eta[\mathcal{L}(w+\eta)] = \mathcal{L}(w) + \nabla\mathcal{L}(w)^\top \mathbb{E}[\eta] + \frac{1}{2}\sum_{i,j} H_{ij}\mathbb{E}[\eta_i\eta_j] + O(\epsilon_0^4) \tag{8}$$

Since η_i are independent uniform variables with mean 0, we have $\mathbb{E}[\eta_i] = 0$ and $\mathbb{E}[\eta_i\eta_j] = 0$ for $i \neq j$. The second moment of $\mathcal{U}(-\epsilon_{q,i}, \epsilon_{q,i})$ is $\mathrm{Var}(\eta_i) = (2\epsilon_{q,i})^2/12 = \epsilon_{q,i}^2/3 = \epsilon_0^2\beta_{v,i}^2/3$. Thus:

$$\frac{1}{2}\sum_{i,j} H_{ij}\mathbb{E}[\eta_i\eta_j] = \frac{1}{2}\sum_{i=1}^{n} H_{ii}\cdot\frac{\epsilon_0^2\beta_{v,i}^2}{3} = \frac{\epsilon_0^2}{6}\sum_{i=1}^{n}\beta_{v,i}^2 H_{ii} \tag{9}$$

Substituting back, and noting that the third moment $\mathbb{E}[\eta_i^3]$ vanishes by the symmetry of the uniform distribution (leaving the fourth-order term as the leading nonzero remainder), yields the theorem. ■

Equivalently, if $\Sigma_\eta = \mathrm{diag}(\epsilon_{q,i}^2/3)$, the second-order penalty is $\frac{1}{2}\mathrm{tr}(H\Sigma_\eta)$. Thus minimizing the QuantNoise-perturbed loss jointly minimizes the original loss and a *noise-weighted* Hessian trace. In the standard isotropic formulation ($\beta_{v,i} = 1$), this reduces to an unweighted trace penalty up to the constant $\epsilon_0^2/6$. In the adaptive variant, because $\beta_{v,i}^2$ is proportional to the optimizer's second moment (which estimates the uncentered gradient variance and is often correlated with high curvature), the penalty is concentrated on high-variance directions. This is a design rationale for low-bit stability, not a standalone guarantee that every production quantizer will preserve the same operating point.

A.3.2 Active Retention: Surrogate-Based Collision-Risk Monitoring

The theoretical treatment of Active Retention proceeds in two stages. First, we use the orthogonality constraint to bound selected matrix-subspace rotation via Davis-Kahan (Theorem 2), providing one structural diagnostic for why localized edits reduce collateral drift. Second, we show that the KL penalty confines parameter updates to the approximate null space of the neighbor-aggregated Fisher information $\mathcal{I}_{\mathrm{NBR}}$ (Proposition 2), providing an assumption-light design consequence for neighbor preservation. As a complementary empirical tool, the Wronskian risk signal—computed from short-window AR surrogate fits of hidden-state trajectories—serves as a lightweight online early-warning indicator for concept collision in non-catastrophic update regimes where the AR surrogate fit remains valid, motivating an optional adaptive variant (Observation 1, Motivation 1).

Theoretical hierarchy and scope. The primary justification for Active Retention is the assumption-light local result in Proposition 2 and the Fisher-geometric remark (Section A): under a fixed reference anchor, small updates, and the quadratic KL approximation, the KL penalty biases updates toward the approximate null space of $\mathcal{I}_{\mathrm{NBR}}$. As an optional supplement, when hidden-state trajectories are locally approximated by short-window AR(p) surrogate models (validity conditions stated below), the Wronskian provides a frequency-domain early-warning signal for concept collision. The theoretical hierarchy thus consists of two levels: (i) the primary **Davis-Kahan + Fisher null-space** results (Theorem 2, Proposition 2), which bound knowledge-subspace rotation and neighbor drift under local approximation assumptions; and (ii) the supplementary **Wronskian** ρ_k (Observation 1), which provides a practical monitoring tool for imminent root collision under the AR surrogate assumption. The Wronskian heuristic therefore complements these analytical results as a practical monitoring tool rather than replacing them.

A.4 Protection of Knowledge Subspaces via Orthogonality Constraint (Application of Davis-Kahan Theorem)

Q-ROU modifies the weight matrix W of SLUG layers into $W + \Delta W$ to suppress target responses. However, these weight matrices also encode general knowledge (e.g., scientific, geographical, and mathematical facts) completely unrelated to the target domains. Here, we use the Davis-Kahan $\sin\Theta$ theorem (Davis and Kahan, 1970) as a local perturbation diagnostic: when the SLUG-layer update is small relative to the relevant spectral gap, the dominant right-singular subspace cannot rotate arbitrarily. This is not a proof of semantic deletion or permanent preservation; it is a sufficient-condition bound on one matrix-level source of collateral drift.

Theorem 2 (Protection of Knowledge Subspace via Orthogonality Constraint). Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ be the singular values of the weight matrix $W \in \mathbb{R}^{m\times n}$ of a SLUG layer. Let $\mathcal{V}$ be the subspace spanned by the top-k right singular vectors of W, and let $\tilde{\mathcal{V}}$ be the corresponding subspace of the post-unlearning matrix $\tilde{W} = W + \Delta W$. As-

suming $\delta = \sigma_k^2 - \sigma_{k+1}^2 > 0$, the following bound holds:

$$\|\sin\Theta\| \leq \frac{d(2a+d)}{\delta} \tag{10}$$

where $a := \|W\|_{\text{op}}$, $d := \|\Delta W\|_{\text{op}}$, and $\delta := \sigma_k^2 - \sigma_{k+1}^2$, with $\Theta = \Theta(\mathcal{V}, \tilde{\mathcal{V}})$.

Proof. We perform Singular Value Decomposition (SVD) on W and denote the subspace spanned by the top-k right singular vectors as $\mathcal{V} = \text{span}\{v_1, \ldots, v_k\}$. To apply the Davis-Kahan theorem, we construct the symmetric matrix $M = W^\top W$. The eigenvectors of M coincide with the right singular vectors of W, associating with eigenvalues σ_i^2. The post-unlearning matrix takes the form $\tilde{M} = \tilde{W}^\top \tilde{W} = W^\top W + E$, where the perturbation matrix E is given by $E = W^\top \Delta W + \Delta W^\top W + \Delta W^\top \Delta W$. By applying the triangle inequality, the operator norm of E is bounded from above:

$$\|E\|_{\text{op}} \leq 2\|W\|_{\text{op}}\|\Delta W\|_{\text{op}} + \|\Delta W\|_{\text{op}}^2 \tag{11}$$

Given the spectral gap of M is $\delta = \sigma_k^2 - \sigma_{k+1}^2$, applying the Davis-Kahan $\sin\Theta$ theorem (Davis and Kahan, 1970) immediately yields the bound stated in the theorem. (Note: the bound is non-vacuous only when $\|E\|_{\text{op}} < \delta$; under Q-ROU's short-step regime with $\|\Delta W\|_{\text{op}}/\|W\|_{\text{op}} < 1\%$, this condition is comfortably satisfied in practice.) ■

Implications and Numerical Validation. This bound shows that the rotation of the selected matrix subspace is restricted by the magnitude of perturbation $\|\Delta W\|_{\text{op}}$ and inversely proportional to the spectral gap δ. We use it as one diagnostic explaining why localized updates can be less damaging than unconstrained full-layer edits. First, the use of **localized updates** via SLUG over a minimal number of steps ensures that the relative perturbation ratio $\|\Delta W\|_{\text{op}}/\|W\|_{\text{op}}$ remains extremely small (typically $< 1\%$). Second, the activation-based SLUG profiling inherently selects layers with **intrinsic gap preservation**, tending toward those with moderately large spectral gaps which naturally fortifies the stability of the bound. Third, the **orthogonal projection** through the $\mathcal{L}_{\text{ortho}}$ constraint and PGD projection reduces update components aligned with the truth direction v_l. Since truth directions are correlated with broad factual directions, this helps limit one class of collateral update directions while restricting the principal rotation $\|\sin\Theta\|$.

Numerical simulations across varying matrix scales ($m = 32$~3072) confirm that the algebraic bound is satisfied in the tested perturbation regime (typical tightness ratio ~ 0.13). Under the prevailing operating regime of Q-ROU, the corresponding matrix-subspace rotation bound is small (approximate bounding angle $\leq 1.5°$~$8°$). We do not equate this matrix bound with a semantic guarantee; the semantic claim is evaluated empirically through neighbor, generation, and extraction audits.

Remark 2 (Spectral Gaps and Representational Geometry). The central role of the spectral gap δ in our bound has a natural parallel in the representation geometry literature. Karkada *et al.* (Karkada et al., 2026) show that when co-occurrence statistics between semantically related words exhibit translation symmetry, learned embeddings organize into Fourier modes whose eigenvalues—and thus spectral gaps—grow with the number of vocabulary items sharing the same latent variable (a ''collective effect''). Applying the Davis-Kahan theorem to this setting, they prove that such representations are robust to localized perturbation of co-occurrence entries, precisely because the spectral gaps overwhelm the perturbation magnitude. This parallel suggests that the spectral gaps protecting knowledge subspaces in our Theorem 2 may not be incidental structural features of pre-trained weight matrices, but rather reflect the systematic organization of knowledge representations around shared latent variables. In either case—whether spectral gaps arise from collective co-occurrence structure or from the intrinsic spectrum of pre-trained weights—the Davis-Kahan mechanism provides a unified explanation for why localized perturbations (be they statistical noise or unlearning updates) preserve the dominant representational structure.

The theoretical foundation for using Fisher information to guide selective forgetting was established by (Golatkar et al., 2020), which formalized unlearning as achieving weight-space indistinguishability from a model retrained without the forget set. Our Fisher null-space argument (Proposition 2) extends this framework to the neighbor-preservation setting, where the goal is to confine updates to directions that leave neighbor output distributions approximately invariant. ◇

Concept entanglement—where suppressing a target concept also damages semantically adjacent neighbor concepts—is a recurring failure mode of standard gradient unlearning (Wang et al., 2026). Our main analytical protection is the Fisher null-space interpretation of AR (Proposition 2) together with the orthogonality regularizer (Theorem 2).

As a supplementary practical tool, we introduce a Wronskian-based warning signal that monitors collision risk at the surrogate-model level. This signal is computed by fitting short-window AR(p) surrogate models to hidden-state trajectories across the token dimension and tracking whether their characteristic polynomial roots converge.

Modeling assumption (locally quasi-stationary AR). We do not claim that hidden state trajectories are generated by a stationary AR(p) process in the strict statistical sense. Rather, we model the token-dimension evolution of a concept's hidden states as a *locally quasi-stationary* AR(p) process—analogous to linear predictive coding (LPC) in speech processing, where a highly non-linear, time-variant signal is approximated by short-window stationary AR segments. The knowledge structure of concept A is encoded in its characteristic polynomial $\phi_A(z) = \sum_{k=0}^{p} \alpha_k z^{p-k}$, where z is the complex shift operator.

This modeling framework is meaningful when the following empirically verifiable conditions hold:

1. **AR fit quality**: the short-window AR(p) fit to hidden-state trajectories achieves sufficient goodness-of-fit ($R^2 \gtrsim 0.8$), indicating that the quasi-stationary approximation captures the dominant dynamics;

2. **Sufficient token length**: concept-specific texts are long enough ($T \gg p$, typically $T > 50$ tokens) to yield stable AR coefficient estimates;

3. **Initial root separation**: the baseline (pre-unlearning) characteristic polynomials ϕ_A, ϕ_B have well-separated roots ($\delta_0 > 0$), so that the surrogate models start in a non-degenerate regime.

When these conditions are reasonably satisfied, the Wronskian risk signal provides useful early-warning information about potential concept collision. When they are violated (e.g., very short texts or poor AR fit), the signal should be treated with caution. In all cases, the primary analytical results (Proposition 2, Theorem 2) remain the main design reference regardless of AR surrogate fit quality.

Why Wronskian as the early-warning signal? When hidden-state trajectories of two concepts A and B are locally approximated by AR(p) surrogate models, each concept is characterized by its surrogate characteristic polynomial $\phi_A(z)$, $\phi_B(z)$. Root sharing ($\phi_A(\zeta^*) = \phi_B(\zeta^*) = 0$) indicates that both surrogate models resonate at the same frequency—a condition correlated with concept entanglement in practice. However, repeatedly estimating full polynomial resultants online is numerically brittle under noisy short-window AR estimates. The Wronskian provides a computationally tractable, gradient-sensitive proxy for detecting imminent root collision, without requiring full polynomial resultant computation. It is (i) computable online during training, (ii) sensitive to *imminent* root collision rather than only post-hoc drift, and (iii) more informative than generic embedding-space distances (cosine/CKA/mean-feature gaps), which cannot test whether two dynamics are approaching root-sharing.

Lemma 1 (Local root-gap sensitivity of Wronskian evaluated on the unit circle). Let $\phi_A(z)$ and $\phi_B(z)$ be monic polynomials of degree p representing the characteristic responses of concepts A and B. To analyze these responses continuously in the frequency domain, we evaluate the polynomials on the unit circle $z = e^{i\omega}$ for $\omega \in [0, 2\pi)$. The continuous Wronskian of these frequency responses is defined as:

$$W(\phi_A, \phi_B)(\omega) = \phi_A(e^{i\omega}) \frac{d}{d\omega} \phi_B(e^{i\omega}) - \phi_B(e^{i\omega}) \frac{d}{d\omega} \phi_A(e^{i\omega}) \tag{12}$$

When evaluating at a shared simple root $\zeta^* = e^{i\omega_0}$, $W(\phi_A, \phi_B)(\omega_0) = 0$. Furthermore, for close roots ζ_A, ζ_B near ω_0, the magnitude of the Wronskian serves as a first-order proxy for the root gap: $|W(\omega_0)| = \Theta(|\zeta_A - \zeta_B|)$ and $W \to 0$ as roots collide.

Proof sketch. Let ω_A, ω_B be the frequencies corresponding to nearby roots $\zeta_A = e^{i\omega_A}$, $\zeta_B = e^{i\omega_B}$. Near $\omega_0 \approx \omega_A \approx \omega_B$, expand $\phi_A(e^{i\omega_0}) = c_A(\omega_0 - \omega_A) + O((\omega_0 - \omega_A)^2)$ and $\phi_B(e^{i\omega_0}) = c_B(\omega_0 - \omega_B) + O((\omega_0 - \omega_B)^2)$, where $c_A = \phi'_A(e^{i\omega_A}) \cdot i e^{i\omega_A}$ and similarly for c_B. The derivatives $\frac{d}{d\omega}\phi_A(e^{i\omega_0})$ evaluate to $c_A + O(\omega_0 - \omega_A)$. Substituting into the Wronskian determinant and retaining leading-order terms, the dominant contribution is proportional to $c_A c_B(\omega_A - \omega_B) + O((\omega_A - \omega_B)^2)$. Since $|\omega_A - \omega_B| = \Theta(|\zeta_A - \zeta_B|)$ for roots on the unit circle, the claim follows. □

Implication. Monitoring $\rho_k = \min_{\omega \in \Omega} |W_k(\omega)|$ provides a low-overhead collision-risk signal that is more operationally useful than generic embedding

distances for detecting imminent root collision in the surrogate models.

Surrogate Collision Detection. When unlearning a target concept A (causing its surrogate characteristic roots to shift), a neighbor concept B faces collision risk if the parameter updates cause the surrogate polynomials $\phi_A(z)$ and $\phi_B(z)$ to drift toward sharing a root. To detect this impending collision, Q-ROU's optional adaptive variant monitors the surrogate Wronskian risk score $\rho_k = \min_{\omega\in\Omega} |W_k(\omega)|$.

This continuous monitoring is meaningful because, under the AR surrogate model, roots vary continuously with the fitted coefficients away from degenerate cases. The argument relies on three local conditions: (a) polynomial roots are continuous functions of their coefficients (Ostrowski's classical theorem); (b) the AR(p) coefficients are continuous functions of the model parameters θ via the Yule–Walker equations $a = \Gamma_p^{-1}\gamma$ as long as Γ_p remains non-singular; and (c) gradient-based training moves θ in small steps. Together, (a)–(c) make abrupt unnoticed root collisions unlikely within the surrogate model, motivating ρ_k as an early-warning signal. ⋄

Intuition. The Wronskian is an online collision alarm at the surrogate level: when two concepts' AR surrogate models start sharing roots, the signal drops toward zero before downstream retention metrics collapse. **Observation 1 (Wronskian Drop Signals Surrogate Root Collision).** Under the AR surrogate model, if the roots of ϕ_A and ϕ_B converge (i.e., $\Delta\zeta \to 0$), the Wronskian evaluated on the unit circle satisfies $W(\phi_A, \phi_B)(\omega) \to 0$.

Proof. By Lemma 1, for nearby simple roots $\zeta_A = e^{i\omega_A}$, $\zeta_B = e^{i\omega_B}$ near ω_0, $|W(\omega_0)| = \Theta(|\zeta_A - \zeta_B|)$. Therefore $\Delta\zeta \to 0$ implies $W(\omega_0) \to 0$ and hence $\rho_k = \min_\omega |W_k(\omega)| \to 0$. At exact root sharing ($\phi_A(\zeta^*) = \phi_B(\zeta^*) = 0$), the solutions become linearly dependent and $W = 0$ exactly. Thus, a decreasing Wronskian acts as an early warning for root collision in the surrogate models. ■

Scope. This observation holds at the surrogate-model level: it characterizes the behavior of the AR(p) approximation to hidden-state trajectories. Whether surrogate root collision reliably predicts actual concept entanglement in the full non-linear neural model depends on the fidelity of the AR surrogate fit. We therefore treat the Wronskian as an optional monitoring heuristic rather than a required component of Q-ROU's core analytical results (Proposition 2, Theorem 2).

In implementation this is realized as a monotone barrier controller:

$$\lambda_a^{(k+1)} = \begin{cases} \min(\lambda_a^{(k)} + \Delta_\lambda, \lambda_{\max}), & \rho_k < \tau_W, \\ \lambda_a^{(k)}, & \text{otherwise,} \end{cases} \tag{13}$$

where $\rho_k = \min_{\omega\in\Omega} |W_k(\omega)|$. This directly maps low Wronskian events to tighter neighbor-drift constraints.

Motivation 1 (Empirical Motivation: Adaptive AR Variant Resists Concept Collision). Proposition 2 establishes that a sufficiently large fixed λ_a confines parameter updates to the approximate null space of $\mathcal{I}_{\text{NBR}}$, protecting neighbor knowledge. The adaptive variant goes further: by monitoring the surrogate Wronskian risk signal ρ_k (Observation 1), it detects *when* collision risk is elevated in the surrogate model and dynamically increases $\lambda_a^{(k)}$ to restore the safety margin. When the adaptive variant of Q-ROU detects $W_k < \tau_W$ and boosts the Active Retention penalty $\lambda_a^{(k)}$, the parameter step size is throttled such that $\|\Delta\theta_B\|_F \le O(1/\lambda_a^{(k)})$. Under the local Fisher approximation (Proposition 2), standard polynomial root perturbation theory (Wilkinson, 1963) then bounds root displacement in the surrogate model, providing an empirical motivation for the ability of this adaptive variant to prevent concept entanglement.

Algebraic Argument. Considering the local stationary conditions of optimization, $\nabla_{\theta_B}\mathcal{L}_f + \lambda_a \mathcal{I}(\theta_B)\Delta\theta_B \approx 0$, the update step relates to the gradient as $\Delta\theta_B \approx -\frac{1}{\lambda_a}\mathcal{I}(\theta_B)^{-1}\nabla_{\theta_B}\mathcal{L}_f$. Under bounded local curvature, this gives the scaling $\|\Delta\theta_B\|_F = O(1/\lambda_a)$. This limits the change in the surrogate polynomial coefficients of B, which by standard perturbation theory for polynomial roots (Wilkinson, 1963) suggests a drift scale $|\Delta\zeta_i^{(B)}| \le K(\lambda_a)^{-1}$, where K depends on the polynomial condition number. If the maximum root displacement is below the minimal separating distance $\delta_0/2$ to the approaching roots of ϕ_A, no root crossings occur in the surrogate model. Therefore, dynamically increasing $\lambda_a^{(k)}$ when the Wronskian drops is a plausible barrier strategy for maintaining distinct surrogate roots.

This indicates that the Wronskian-Adaptive variant provides a practical safeguard when surrogate root collision is imminent. We acknowledge certain limitations: identifying exact Lipschitz constants for neural dynamics is practically difficult, the Fisher Information Matrix provides a 2nd-order local approximation that accumulates error over

multiple training steps, and the AR surrogate fit is itself an approximation. Thus, this derivation serves as a motivation for using the Wronskian signal as optional online instrumentation rather than as a globally strict formal claim. In the standard Q-ROU formulation, a sufficiently large fixed λ_a provides empirical safe separation globally, albeit with less dynamic responsiveness.

A.4.1 Quantization Robustness: From Constrained Mobility to Flat Basins

Remark 1 (Asymmetric Quantization Sensitivity under AR Constraint). Let θ^* be the post-unlearning parameters obtained under Active Retention with coefficient λ_a. Let $Q(\theta^*)$ denote the INT4-quantized parameters, and define the per-case *quantization shift* $\delta(x, y) = P_{Q(\theta^*)}(y|x) - P_{\theta^*}(y|x)$. AR constrains neighbor probabilities to the high-confidence regime $P_{\theta^*}(y|x_n) \gg \tau$ where the softmax Jacobian is near-singular, while target probabilities are pushed to the low-probability regime $P_{\theta^*}(y|x_t) \approx \tau$. In this configuration, the *sensitivity* of output probabilities to weight perturbation is strongly asymmetric:

$$\left|\frac{\partial P}{\partial w}\right|_{x_n} \ll \left|\frac{\partial P}{\partial w}\right|_{x_t} \tag{14}$$

Analysis. The softmax function $\sigma(z)_k = e^{z_k}/\sum_j e^{z_j}$ has sensitivity $\partial\sigma_k/\partial z_i$ that is maximal at intermediate probabilities ($P \approx 0.5$) and minimal at extremes ($P \approx 0$ or $P \approx 1$). AR pushes neighbor probabilities into the saturated high-confidence regime, making them insensitive to small weight perturbations from quantization. Target probabilities, however, are *highly sensitive* to perturbations due to the steep gradient in the low-probability regime.

A key observation here is that this sensitivity is *not* directionally biased at first order: symmetric weight perturbations from INT4 rounding produce zero-mean logit changes, and hence zero-mean probability changes to leading order. At second order, Jensen's inequality might suggest that the convexity of e^{z_k} in the numerator would increase the target probability; however, in the full softmax ratio $P_k = e^{z_k}/\sum_j e^{z_j}$, the same convexity effect applies to the denominator, and these contributions approximately cancel for small perturbations. The net directional bias of quantization perturbation on the softmax output is therefore *not* determined by simple single-variable convexity arguments, but depends on the global structure of the loss landscape around the optimized point. In AR-equipped models, we empirically observe $\Delta_{\text{zombie}} < 0$ (Table 34), consistent with the composite KL-to-uniform objective reshaping the local curvature asymmetrically such that INT4 rounding tends to nudge borderline targets further below threshold. The full nonlinear dynamics of multi-layer transformer quantization involve higher-order interactions between layers that are not captured by single-layer softmax analysis.

Practical consequence. The key empirical finding remains robust: neighbor probabilities are effectively anchored by AR against quantization perturbation, while target probabilities—being in a highly sensitive regime—experience measurable shifts. The observed direction of these shifts (synergistic rather than antagonistic) is a favorable property of the Q-ROU loss landscape geometry, confirmed deterministically across three random seeds. ◇

Scope caveat. This analysis explains the *asymmetric sensitivity* (why targets are affected more than neighbors under quantization) rigorously, while the *directional* component (why shifts tend to be synergistic) is an empirical regularity supported by the experimental evidence (+0 to +1 target cases under INT4 in AR-equipped baselines at 40 steps) rather than a first-principles prediction.

Constrained-mobility regime and sign persistence. A complementary perspective on why short-step unlearning edits are vulnerable to quantization comes from the *constrained-mobility* of weight updates. In the sub-bit model compression literature, Sakai and Ichikawa (Sakai and Ichikawa, 2026) formalized a *sign lock-in* phenomenon: under standard SGD training, weight signs (±) initialized randomly are rarely flipped, because sign changes require the trajectory to cross through a narrow boundary neighborhood near zero—an exponentially unlikely event under bounded updates. Their stopping-time analysis shows that the number of effective sign flips follows a geometric tail, and this persistence strengthens with model scale.

We invoke sign lock-in as a *formally motivated geometric prior*. Although Sakai and Ichikawa's (Sakai and Ichikawa, 2026) proof is stated for scheduled SGD, Remark D.11 of their work establishes that the geometric-tail conclusion transfers to any optimizer satisfying a bounded-update condition, including AdamW, provided that the effective per-step increment $\|p_{t+1}\|_\infty$ remains bounded on a high-probability event. In our short-

step ($\leq$ 40 steps) unlearning regime with gradient clipping, this condition is empirically satisfied. This provides intuition for why, in our unlearning regime, many edited weights do not move far enough to stably cross INT4 quantization-cell boundaries. When post-training rounding maps these borderline edits back toward their pre-edit quantization cells, the Zombie Delta phenomenon arises—targeted knowledge partially resurfaces under quantization.

Flat basin strategy. Forcing larger updates to cross more quantization boundaries could reduce this failure mode, but at the cost of increased risk of collateral neighbor damage (Neighbor Collapse). Our strategy instead achieves robustness *without* relying on aggressive displacement: Adaptive QuantNoise regularizes the Hessian curvature (Theorem 1), guiding optimization toward flat basins whose plateaus are wider than the expected INT4 rounding perturbation radius. This approach is particularly critical given the emergence of outlier features in large-scale transformers (Dettmers et al., 2022), where specific hidden dimensions exhibit extreme magnitudes that dominate quantization error. While LLM.int8() addresses outliers through mixed-precision decomposition during inference, QuantNoise preemptively flattens the loss landscape during unlearning, ensuring that even outlier-sensitive dimensions remain robust to INT4 rounding.

Consequently, even when individual weight updates remain modest, the model can maintain resilience after quantization: the flat basin absorbs rounding errors rather than allowing them to undo the forget edit. This provides an empirical reconciliation between the constrained-mobility regime and quantization robustness, confirmed deterministically across three random seeds with $\Delta_{\text{zombie}} \approx 0$ for the full Q-ROU framework.

B Extended Experimental Setup

The core experimental settings are described in Section 3 of the main text. This section provides extended design notes.

Design Note on Forget Objectives. Q-ROU uses a bounded KL-to-uniform forget objective (defined in Section 2), whereas the primary baselines use unbounded gradient ascent on the forget set. This design choice is deliberate: unbounded loss maximization is known to cause numerical instability and model degradation (Jang et al., 2023), as evidenced by the degenerate generation observed for GA variants under extended training (Section D.4). To control for this confound, our AR transplant experiments (Section C.5) add AR *to the baselines' own objectives*—that is, GA+AR retains gradient ascent as its forget mechanism—thereby isolating the contribution of Active Retention from the choice of forget-loss formulation. The fact that GA+AR recovers 8/8 neighbor retention (vs. 0/8 without AR) while still using unbounded gradient ascent demonstrates that AR's benefit is independent of the forget objective. We further address this distinction in Section D.4, where a 3×3 hyperparameter sweep over learning rate and AR weight confirms that the degeneration–forgetting tradeoff is intrinsic to gradient ascent rather than an artifact of specific hyperparameter choices.

Hyperparameter Fairness in Multi-Entity Settings. Baseline learning rates (3×10^{-4} for 0.5B, 1×10^{-4} for 3B) were selected in the single-entity setting and applied without modification to multi-entity experiments to ensure fair comparison. While the multi-entity regime involves 18 forget texts (vs. 1), the gradient magnitude difference is partially absorbed by batch-level averaging. Three controls rule out under-tuning as an explanation for baseline failure: (1) the 40-step extended training control (Section C.5) shows that longer training *worsens* collapse without AR (0/8 NBR *and* 0/7 GEN), eliminating insufficient training duration as a factor; (2) a 3×3 grid over learning rate $\{5\times10^{-5}, 10^{-4}, 5\times10^{-4}\}$ and $\lambda_a \in \{40, 80, 160\}$ for GA+AR (Section D.4) spans the multi-entity operating space and confirms that retention collapse is robustly tied to the gradient-ascent objective; and (3) multi-seed verification across three random seeds yields identical reported pass-count totals, confirming that the observed collapse is not a random-seed fluke.

B.1 Threat Model and Attack Budget

Following extraction-risk formulations in recent unlearning work (Patil et al., 2024; Shi et al., 2025), we evaluate two attacker capabilities. In the *black-box* setting, the attacker can issue prompts and observe model generations or probabilities but cannot access weights. In the *evaluation-time oracle-ranked* setting, the attacker can rank fixed benchmark probes using returned target-token probabilities, but cannot modify parameters or run gradient updates. Attack strength is controlled by a query

budget $B \in \{1, 2, 4, 8, 16\}$ drawn from the adversarial probe bank. For each method/precision pair, we report budgeted extraction success as the probability of recovering at least one forgotten target within the first B probes, under both observed probe order and oracle risk-ranked order.

B.2 Evaluation Metrics

We evaluate across six dimensions. **Target Pass Rate (TGT↑)** is the fraction of target cases where $P(y|x) < 0.01$; higher indicates more effective forgetting. **Neighbor Pass Rate (NBR↑)** and **General Pass Rate (GEN↑)** measure the fraction of neighbor and general cases satisfying the relative-retention criterion ($P_{\theta^*}/P_\theta \geq 0.30$; fallback $P > 0.05$); higher indicates better preservation. **Zombie Delta (Δ_{zombie})** captures the mean target-probability difference between INT4 and full-precision evaluations. Values near zero indicate similar FP/INT4 forgetting behavior; positive values indicate forgetting reversal; negative values indicate additional target suppression after quantization. **Adversarial Pass Rate (ADV↑)** measures resilience against paraphrased knowledge extraction probes. **Depth Suppression Rate (DSR↑)** measures resistance to structured probing protocols including reverse association, multistep reasoning, reconstruction, in-context extraction, and negation attacks. We also report wall-clock time and peak VRAM usage.

Metric caveat. NBR and GEN are based on token-level probability retention, which can decouple from generation-level utility. A model that assigns high probability to the correct next token but degenerates during extended generation may still achieve high retention scores by these metrics. We address this gap directly in Section D.3, where semantic embedding evaluation reveals that GA+AR achieves 8/8 token-level neighbor retention while exhibiting generation degeneration and masked semantic leakage, whereas Q-ROU maintains coherent generation alongside its token-level scores.

B.3 Statistical and Privacy Inference Protocol

Pass-rate metrics (TGT/NBR/GEN/ADV/DSR) are treated as binomial proportions and reported with Wilson 95% confidence intervals. For mean target probabilities and paired deltas between evaluation blocks on matched prompts, we report bootstrap 95% confidence intervals (2,000 resamples, fixed seed). Following TOFU-style inferential evaluation (Maini et al., 2024), we compute two-sample KS statistics and p-values on surprise-score distributions ($-\log P$) between target and retained categories. In line with MUSE-style privacy leakage auditing (Shi et al., 2025), we additionally report AUC-based separability proxies (Target vs. General/Neighbor) on the same surprise scores. Consistent with ASA guidance on p-values (Wasserstein and Lazar, 2016), we do not treat p-values as standalone evidence and avoid dichotomous ''significant/non-significant'' conclusions from a single threshold. All inferential statements are based jointly on interval estimates (Wilson/bootstrap), effect-size statistics (pass-rate deltas, KS distance, AUC deviation from 0.5), and explicit attack-budget trends. These statistical measures are computed from raw experiment outputs with the same evaluation pipeline that produced the reported tables.

B.4 Cross-Benchmark Correspondence

To avoid benchmark-fragmented claims, Table 7 maps our evaluation protocol to commonly referenced unlearning benchmarks. Our intent is *complementarity*: we do not replace MUSE/TOFU/WMDP/RWKU/OpenUnlearning, but add a controlled stress regime in which simultaneous multi-entity deletion, explicit neighbor integrity, and quantized deployment consistency are evaluated together.

This mapping clarifies scope: our claims are strongest on neighbor-aware multi-entity robustness under quantized deployment, while full leaderboard comparison on MUSE/WMDP/RWKU and full reasoning-trace unlearning remain future work.

B.5 End-to-End MUSE/WMDP Validation

To complement the custom probe suite and TOFU results, we ran end-to-end MUSE and WMDP evaluations in FP16. These evaluations cover full 3B executions for Q-ROU, GA+AR, and GA, plus Q-ROU-only 8B runs. We treat them as external sanity checks on training stability and utility preservation rather than as primary leaderboard claims; the 8B setting is included to test whether the same operating-point story remains visible at a larger scale.

Table 8 summarizes the Q-ROU checkpoints. On MUSE, Q-ROU preserves strong utility at both 3B and 8B while increasing the relative KnowMem reduction metric; the 3B value reflects a 40-step early checkpoint, while the 8B 160-step run shows that the same pipeline moves the MUSE reduction metric more decisively when trained longer. On WMDP, 3B shows the expected ''do no harm''

Table 7: Benchmark correspondence map. ‘‘Primary’’ indicates a first-class design axis in the original benchmark paper. ‘‘Partial’’ indicates the axis is present but not the central benchmark objective.

Benchmark	Core deletion setting	Utility/retain protocol	Neighbor-specific axis	Quantized deployment axis
TOFU (Maini et al., 2024)	Fictitious QA split	Primary	No explicit stratum	Not reported
MUSE (Shi et al., 2025)	Books/news unlearning	Primary	Partial (locality)	Not reported
WMDP (Li et al., 2024)	Hazardous-domain MCQ	Primary	Partial (non-target checks)	Not reported
RWKU (Jin et al., 2024)	Real-person factual deletion	Primary	Partial (neighbor perturbation)	Not reported
OpenUnlearning (Dorna et al., 2025)	Unified TOFU/MUSE/WMDP execution	Primary	Benchmark-dependent	Not primary
R-TOFU (Yoon et al., 2025)	Reasoning-trace unlearning	Primary	Not primary	Not reported
This work	Controlled TGT-NBR-GEN deletion	Primary	**Primary**	**Primary (FP16, INT4, NF4)**

regime with 97.5% utility retention and only a 0.0016 absolute forget-accuracy drop, while 8B exhibits a larger but still controlled 0.0183 forget-accuracy drop with 94.5% utility retention.

We also evaluated the Llama-3.2-3B standard-benchmark scripts with matched Q-ROU/GA+AR/GA controls under the same evaluation pipeline used for TOFU. Table 9 is used to audit whether the method remains stable under external benchmark code and to expose the utility–forgetting trade-off of aggressive baselines, not to claim that the controlled probe suite should replace full MUSE/WMDP/RWKU leaderboards.

The Q-ROU/news MUSE row is most useful as a saturation diagnostic. Under the default 40-step setting, Q-ROU does not yet improve the MUSE news knowledge-forget metric (KnowMem reduction = −0.0956, privacy score = 0.5742, utility = 0.5541), which makes the row a clean indicator that this benchmark needs a longer horizon than the custom suite. The standard-benchmark picture is therefore as follows: WMDP supports the low-damage operating point, MUSE books supports high utility under the benchmark implementation, GA/GA+AR expose how aggressive forgetting can destroy utility, and MUSE news identifies the one benchmark setting where calibration matters most.

As a calibration sweep, we tested the same MUSE news benchmark at 120 steps with a shorter max sequence length and a small sweep over (λ_f, λ_a). These rows show that the 40-step default is indeed under-saturated: several 80-step and 120-step settings move KnowMem reduction into positive territory while keeping utility roughly in the 0.97–1.00 range, and the best row ($\lambda_f = 40$, $\lambda_a = 40$, step 120) reaches KnowMem reduction = 0.0807, forget quality = 1.0363, privacy = 0.5986, and utility = 0.9934. Because these calibration runs use max-sequence-length 64 to avoid reference-logit OOM, we treat them as a calibrated companion to the original full-length row rather than a direct replacement. The key reading is positive: once the horizon is long enough, Q-ROU can recover a strong MUSE-news utility–forgetting trade-off rather than being confined to the weak default row.

We also ran a full-sequence continuation audit on the same 3B MUSE news setup at 120/160/240 steps, using the benchmark's longer sequence path; the 240-step comparison uses max-sequence-length 192 to fit the available memory budget. This continuation sharpens the same frontier under a stricter path. At 240 steps, Q-ROU keeps utility high (utility score = 0.9731, retain PPL = 18.34) even though the benchmark's news forget metric remains demanding (KnowMem reduction = −0.1596, forget quality = 1.0468). GA+AR moves much farther on the forget side under the

Table 8: Completed end-to-end MUSE/WMDP validation for FP16 Q-ROU. For MUSE, we report Relative KnowMem Reduction = 1 − $\text{KnowMem}_{\text{after}}/\text{KnowMem}_{\text{base}}$ (higher is stronger forgetting) alongside Utility. For WMDP, Forget Drop is the absolute decrease in hazardous-domain accuracy and Utility Retention is the retained non-target accuracy ratio.

Benchmark	Model	Steps	Forget metric	Utility metric
MUSE	Llama-3.2-3B	40	Rel. KnowMem Reduction = 0.0060	Utility = 0.8793
MUSE	Llama-3.1-8B	160	Rel. KnowMem Reduction = 0.3540	Utility = 0.8907
WMDP	Llama-3.2-3B	40	Forget Drop = 0.0016	Utility Retention = 97.54%
WMDP	Llama-3.1-8B	160	Forget Drop = 0.0183	Utility Retention = 94.51%

Benchmark	Setting	Method	Steps	Forget metric	Utility metric	Baseline→Unlearned
MUSE	books	ga	40	100.0000	0.0000	1.0000→0.5105
MUSE	books	ga_ar	40	1.8210	0.9372	0.5105→0.3090
MUSE	books	qrou	40	1.0234	0.9970	0.1916→0.4880
MUSE	news	ga	40	100.0000	0.0000	0.9826→0.4731
MUSE	news	ga_ar	40	1.2667	0.0000	0.1527→0.5720
MUSE	news	qrou	40	1.0243	0.5541	-0.0956→0.5742
MUSE	news	qrou	120	1.0317	0.9853	0.0383→0.5979
MUSE	news	qrou	120	1.0327	0.9998	0.0313→0.5980
MUSE	news	qrou	120	1.0308	0.9836	0.0313→0.5978
MUSE	news	qrou	120	1.0363	0.9934	0.0807→0.5986
MUSE	news	qrou	120	1.0348	0.9994	0.0343→0.5984
WMDP	wmdp-bio,wmdp-chem,wmdp-cyber	ga	40	0.1082	0.5554	0.3580→0.2497
WMDP	wmdp-bio,wmdp-chem,wmdp-cyber	ga_ar	40	0.0166	0.9266	0.3580→0.3413
WMDP	wmdp-bio,wmdp-chem,wmdp-cyber	qrou	40	0.0033	0.9842	0.3580→0.3547

Table 9: Standard-benchmark runs on Llama-3.2-3B. For WMDP, the forget metric is accuracy drop on the WMDP subsets and the utility metric is MMLU-style utility retention; the last column reports WMDP baseline-to-unlearned accuracy. For MUSE, the forget metric is the benchmark's forget-quality score, the utility metric is retain-utility preservation, and the last column reports knowledge-forget to privacy-score.

same full-sequence evaluation (KnowMem reduction = 0.6051, forget quality = 1.8422), but only by collapsing utility (utility score = 0.0000, retain PPL = 54.25). We therefore interpret the longer full-sequence continuation as a stronger measurement of the same utility–forgetting separation: extending horizon improves the usable Q-ROU operating region, while aggressive baselines still buy forget-score gains by sacrificing the model.

B.6 Fine-Grained Neighbor and Baseline-Fairness Audits

Two additional audits sharpen the custom-probe evidence in complementary directions. First, we split the original neighbor notion into alias, same-work neighbor, inter-domain neighbor, and general strata. This audit is intentionally stricter than the original 28-probe table: aliases should usually be removed with the target, while same-work non-target facts should be protected when the removal request is not meant to erase the whole work. Table 11 shows the resulting boundary. The original Q-ROU operating point deletes targets and aliases completely but is too aggressive on same-work neighbors (9/36). The fine-grained AR variant, which is the appropriate operating point for this sharper request, preserves substantially more same-work and inter-domain material (19/36 and 33/36 in FP16) while retaining complete target deletion (39/39). A three-seed real-NF4 evaluation preserves the same ordering rather than overturning it: base Q-ROU stays at 9/36 on same-work neighbors under BnB NF4, while qrou_fg_ar moves only from 19/36 to 18/36 and keeps 39/39 target suppression (Table 58). The practical interpretation is straightforward: AR is the mechanism that lets the method state and en-

λ_f	λ_a	Step	KnowMem red.	Forget quality	Privacy	Utility	KnowMem base→after
2.0	80.0	40	-0.0956	1.0243	0.5742	0.5541	0.0176→0.0192
2.0	80.0	80	0.0112	1.0377	0.5981	0.9714	0.0176→0.0174
2.0	80.0	120	0.0313	1.0308	0.5978	0.9836	0.0176→0.0170
10.0	80.0	40	-0.2575	1.0052	0.5947	0.9959	0.0176→0.0221
10.0	80.0	80	0.0283	1.0382	0.5982	0.9717	0.0176→0.0171
10.0	80.0	120	0.0383	1.0317	0.5979	0.9853	0.0176→0.0169
20.0	80.0	40	-0.2580	1.0055	0.5948	0.9959	0.0176→0.0221
20.0	80.0	80	0.0312	1.0389	0.5980	0.9724	0.0176→0.0170
20.0	80.0	120	0.0313	1.0327	0.5980	0.9998	0.0176→0.0170
40.0	40.0	40	-0.2409	1.0142	0.5958	0.9901	0.0176→0.0218
40.0	40.0	80	0.0013	1.0434	0.6008	0.9984	0.0176→0.0175
40.0	40.0	120	0.0807	1.0363	0.5986	0.9934	0.0176→0.0161
40.0	80.0	40	-0.2575	1.0064	0.5949	0.9960	0.0176→0.0221
40.0	80.0	80	-0.0164	1.0409	0.5985	0.9731	0.0176→0.0178
40.0	80.0	120	0.0343	1.0348	0.5984	0.9994	0.0176→0.0170

Table 10: MUSE news Q-ROU calibration sweep on Llama-3.2-3B. Rows include the 40-step baseline and 120-step max-sequence-length-64 calibration runs evaluated at intermediate snapshots. Because the 120-step sweep shortens the sequence length to avoid frozen-reference-logit OOM, it should be read as a calibrated companion to the original full-length benchmark row. KnowMem reduction is $1-\text{KnowMem}_{after}/\text{KnowMem}_{base}$, so negative values indicate increased KnowMem overlap after unlearning.

force the protected boundary more precisely. In other words, the same-work results do not overturn the broad-neighbor claim; they identify a stricter frontier where stronger neighbor specification is required.

To test whether this boundary survives a larger copyright-like workload, we also built a harsher 81-probe benchmark with 25 target probes (13 core + 12 alias), 24 same-work neighbors, 16 inter-domain neighbors, and 16 general probes split across Harry Potter and Lord of the Rings. Table 12 shows that the larger audit preserves the same qualitative structure while making the boundary even more explicit. Default Q-ROU remains target-complete (25/25) but preserves only 4/24 same-work neighbors. The fine-grained AR operating point recovers same-work retention to 14.7/24 in FP16 while staying target-strong (24.7/25), and the same ordering survives real BnB NF4 with only moderate drift (12.7/24 same-work, 25/25 target). GA+AR and its finer-grained variant continue to trace the opposite side of the frontier: they retain same-work material more aggressively, but only by accepting much weaker target suppression. Table 13 adds a compact aggregation of these same-work results, including the harsher open-ended generation audit. Those generation rows show the same pattern in a stricter form: same-work open-ended evaluation remains difficult even when probability-level suppression and NF4 stability are already strong.

Second, we evaluated the closest selective baseline family under a wider 40-step GA+AR hyper-

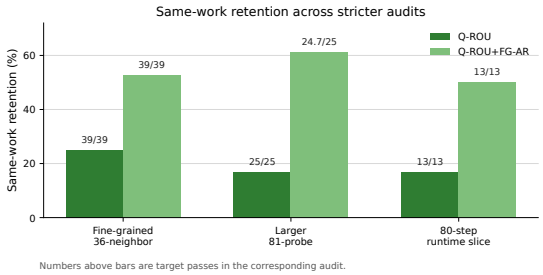

Figure 2: Same-work retention across increasingly strict audit slices. The first pair is the original fine-grained 36-neighbor audit, the second pair is the larger 81-probe copyright-like benchmark, and the third pair is the single-hardware 80-step runtime slice. The intended reading is that same-work retention is incomplete in all strict settings, but the fine-grained AR variant consistently improves retention while preserving much stronger target suppression than GA-family retain-heavy alternatives.

parameter grid. Table 14 shows both sides of that frontier directly. High-learning-rate GA+AR can match the small core token pass counts, but only in the same regime where generation-level probes become degenerate. Together with the MUSE/WMDP and generation audits, this supports a narrower but stronger claim: token-level target suppression is easy to over-optimize, so deployment unlearning should be judged by target removal, coherent generation, neighbor retention, and precision robustness jointly.

Method	Regime	Target	Alias	Same-work	Inter-domain	General
gaar	fake_int4_g128	21/39	18/24	24/36	33/36	27/30
gaar	fake_int4_g32	21/39	18/24	21/36	36/36	30/30
gaar	fake_int4_g64	18/39	18/24	24/36	36/36	30/30
gaar	fp16	21/39	18/24	27/36	36/36	30/30
gaar_fg_ar	fake_int4_g128	6/39	9/24	27/36	33/36	27/30
gaar_fg_ar	fake_int4_g32	6/39	9/24	24/36	36/36	30/30
gaar_fg_ar	fake_int4_g64	3/39	9/24	24/36	36/36	30/30
gaar_fg_ar	fp16	3/39	12/24	36/36	36/36	30/30
qrou	fake_int4_g128	39/39	24/24	9/36	24/36	20/30
qrou	fake_int4_g32	39/39	24/24	9/36	24/36	21/30
qrou	fake_int4_g64	39/39	24/24	9/36	25/36	23/30
qrou	fp16	39/39	24/24	9/36	27/36	24/30
qrou_fg_ar	fake_int4_g128	39/39	23/24	15/36	30/36	27/30
qrou_fg_ar	fake_int4_g32	39/39	23/24	17/36	31/36	30/30
qrou_fg_ar	fake_int4_g64	39/39	23/24	15/36	33/36	30/30
qrou_fg_ar	fp16	39/39	23/24	19/36	33/36	30/30

Table 11: Fine-grained neighbor audit. Pass counts are aggregated across seeds using the default forget and retain thresholds.

Table 12: Larger copyright-like same-work benchmark on Llama-3.2-3B (3 seeds mean). The harder 81-probe split preserves the same reading as the smaller fine-grained audit: default Q-ROU is target-complete but too aggressive on same-work neighbors, while qrou_fg_ar recovers a substantial part of that boundary and keeps the same qualitative advantage under real BnB NF4.

Method	FP TGT	FP SW	FP ID	FP GEN	NF4 TGT	NF4 SW	NF4 ID	NF4 GEN
Q-ROU	**25.0/25**	4.0/24	11.7/16	13.3/16	**25.0/25**	4.0/24	10.0/16	12.7/16
qrou_fg_ar	24.7/25	14.7/24	14.3/16	**16.0/16**	**25.0/25**	12.7/24	14.0/16	15.3/16
GA+AR	16.0/25	17.0/24	15.0/16	**16.0/16**	16.0/25	17.0/24	15.0/16	**16.0/16**
gaar_fg_ar	8.0/25	**22.0/24**	**16.0/16**	**16.0/16**	7.0/25	**23.0/24**	15.0/16	**16.0/16**

C Results

C.1 Single-Entity Unlearning

Table 15 presents the single-entity comparison on Llama-3.2-3B at 20 steps. All baselines achieve complete target removal (7/7) but at varying degrees of collateral cost: GA and GradDiff lose neighbor knowledge under INT4 ($8/9 \rightarrow 6/9$ and $7/9 \rightarrow 4/9$ respectively), and in FP16 all three baselines drop general knowledge to 6/7. Q-ROU at 20 steps takes a qualitatively different trajectory: it preserves perfect neighbor (9/9) and general (7/7) retention while target removal is still in progress (2/7 FP). Notably, Q-ROU's INT4 target score (3/7) already exceeds its FP score (2/7), the first manifestation of the quantization-synergistic forgetting pattern analyzed in Section C.5. Adversarial probing with 6 paraphrased prompts further illustrates the trade-off: GA's aggressive forgetting yields 6/6 adversarial defense at the cost of neighbor destruction, while Q-ROU achieves 5/6 defense at 20 steps with perfect retention.

Table 16 reveals the steps-Pareto trade-off: Q-ROU achieves a **perfect score** at 40 steps—7/7 target, 9/9 neighbor, 7/7 general in both FP16 and INT4, with $\Delta_{\text{zombie}} = 0.000$. No baseline achieves this combination at any step count. The 40-step sweet spot reflects Active Retention's controlled pace: the KL constraint prevents destructive rapid forgetting, requiring approximately twice the steps while helping ensure that the knowledge boundary is respected throughout optimization.

Table 13: Larger same-work probability audit and generation boundary. Probability rows are three-seed means; the generation rows show that same-work open-ended evaluation remains a hard boundary rather than a new positive claim.

Method	Setting	TGT	Same-work	Inter/deg.	GEN
gaar	fp16	16.0/25 ± 0.0	17.0/24 ± 0.0	15.0/16 ± 0.0	16.0/16 ± 0.0
gaar	fake int4 g32	15.0/25 ± 0.0	14.0/24 ± 0.0	15.0/16 ± 0.0	15.0/16 ± 0.0
gaar	bnb nf4	16.0/25 ± 0.0	17.0/24 ± 0.0	15.0/16 ± 0.0	16.0/16 ± 0.0
gaar_fg_ar	fp16	8.0/25 ± 0.0	22.0/24 ± 0.0	16.0/16 ± 0.0	16.0/16 ± 0.0
gaar_fg_ar	fake int4 g32	9.0/25 ± 0.0	17.0/24 ± 0.0	15.0/16 ± 0.0	15.0/16 ± 0.0
gaar_fg_ar	bnb nf4	7.0/25 ± 0.0	23.0/24 ± 0.0	15.0/16 ± 0.0	16.0/16 ± 0.0
qrou	fp16	25.0/25 ± 0.0	4.0/24 ± 0.0	11.7/16 ± 0.5	13.3/16 ± 0.5
qrou	fake int4 g32	25.0/25 ± 0.0	4.0/24 ± 0.0	11.0/16 ± 0.0	10.7/16 ± 0.9
qrou	bnb nf4	25.0/25 ± 0.0	4.0/24 ± 0.0	10.0/16 ± 1.4	12.7/16 ± 0.5
qrou_fg_ar	fp16	24.7/25 ± 0.5	14.7/24 ± 0.5	14.3/16 ± 0.5	16.0/16 ± 0.0
qrou_fg_ar	fake int4 g32	24.0/25 ± 0.8	11.3/24 ± 0.5	13.0/16 ± 0.0	15.0/16 ± 0.0
qrou_fg_ar	bnb nf4	25.0/25 ± 0.0	12.7/24 ± 0.5	14.0/16 ± 0.0	15.3/16 ± 0.5
gaar	generation	8.0/25 ± 0.0	4.0/24 ± 0.0	34.0/81 ± 0.0	–
gaar_fg_ar	generation	8.0/25 ± 0.0	6.0/24 ± 0.0	30.0/81 ± 0.0	–
qrou	generation	11.0/25 ± 0.8	0.3/24 ± 0.5	45.3/81 ± 4.1	–
qrou_fg_ar	generation	9.0/25 ± 1.4	1.7/24 ± 0.5	39.0/81 ± 0.8	–

C.2 Multi-Entity Unlearning

Multi-entity unlearning—simultaneously removing Harry Potter and Lord of the Rings knowledge—provides the critical test of a method's true selectivity. Table 17 exposes a stark divide: all three baselines achieve *zero* neighbor retention (0/8), while the 40-step variant Q-ROU† reaches 27/28 in FP16 and 28/28 in INT4.

GA additionally destroys all general knowledge (0/7), and even RepBend—the most selective baseline—retains only 5/7. The collapse from single-entity results (7−8/9 NBR) to multi-entity results (0/8 NBR) is dramatic: simultaneously applying gradient ascent to 18 texts spanning two fictional domains overwhelms any implicit knowledge separation, propagating destructive updates through shared representational subspaces. Q-ROU at 20 steps demonstrates early-stage controlled forgetting: 4/13 target removal with perfect 8/8 neighbor and 7/7 general retention. At 40 steps, Q-ROU achieves **complete target removal** (13/13) in both FP16 and INT4, with 7/8 neighbor retention in FP16 and 8/8 in INT4 (both with 7/7 general retention)—reaching 27/28 in FP16 and 28/28 in INT4. This result reproduces the same reported pass-count totals across three random seeds, confirming that the combined AR+SLUG+QuantNoise framework reliably navigates the multi-entity forgetting landscape. The

Configuration	HP	LOTR	TGT	NBR	GEN	GenClean
Q-ROU 40s	7/7	6/6	13/13	8/8	7/7	5/7
GA+AR lr=5e-05, λ_a=40	7/7	2/6	9/13	8/8	7/7	2/7
GA+AR lr=5e-05, λ_a=80	6/7	2/6	8/13	8/8	7/7	3/7
GA+AR lr=5e-05, λ_a=160	5/7	1/6	6/13	8/8	7/7	1/7
GA+AR lr=0.0001, λ_a=40	7/7	2/6	9/13	8/8	7/7	5/7
GA+AR lr=0.0001, λ_a=80	7/7	2/6	9/13	8/8	7/7	3/7
GA+AR lr=0.0001, λ_a=160	7/7	2/6	9/13	8/8	7/7	3/7
GA+AR lr=0.0005, λ_a=40	7/7	6/6	13/13	8/8	7/7	7/7
GA+AR lr=0.0005, λ_a=80	7/7	6/6	13/13	8/8	7/7	7/7
GA+AR lr=0.0005, λ_a=160	7/7	4/6	11/13	8/8	7/7	7/7

Table 14: GA+AR hyperparameter-grid audit on the 28-probe core suite plus greedy generation leakage checks. The grid shows that GA+AR can be tuned to match core pass counts, but several tuned rows rely on degenerate generation behavior; this is analyzed as a baseline-fairness diagnostic rather than a replacement for the main neighbor-aware audit.

Table 15: Single-entity unlearning on Llama-3.2-3B (20 steps). FP: full precision; I4: INT4 g32 quantized. Q-ROU shows INT4 TGT > FP TGT while keeping retention stable.

Method	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z	ADV
Baseline	1/7	9/9	7/7	1/7	9/9	7/7	−0.083	2/6
GA	7/7	8/9	6/7	7/7	6/9	7/7	0.000	6/6
GradDiff	7/7	7/9	6/7	7/7	4/9	7/7	0.000	6/6
RepBend	7/7	8/9	6/7	7/7	8/9	7/7	0.000	6/6
Q-ROU	2/7	**9/9**	**7/7**	3/7	**9/9**	**7/7**	−0.007	5/6

Table 16: Q-ROU steps-Pareto on Llama-3.2-3B. At 40 steps, all metrics reach their optimal values in both FP and INT4.

Steps	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z	Time
10	1/7	9/9	7/7	1/7	9/9	7/7	−0.022	5.6s
20	2/7	**9/9**	**7/7**	3/7	**9/9**	**7/7**	−0.007	11.0s
40	**7/7**	**9/9**	**7/7**	**7/7**	**9/9**	**7/7**	**0.000**	22.5s

Table 17: Multi-entity unlearning on Llama-3.2-3B under single-GPU execution. Baselines and Q-ROU are reported at 20 steps; Q-ROU† denotes the 40-step variant. All baselines show complete neighbor collapse.

Method	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z	Total
Baseline	1/13	8/8	7/7	1/13	8/8	7/7	−0.073	—
GA	**13/13**	0/8	0/7	**13/13**	0/8	0/7	+0.000	13/28
GradDiff	**13/13**	0/8	4/7	**13/13**	0/8	4/7	+0.000	17/28
RepBend	**13/13**	0/8	5/7	**13/13**	0/8	5/7	+0.000	18/28
Q-ROU	4/13	**8/8**	**7/7**	4/13	**8/8**	**7/7**	−0.006	19/28
Q-ROU†	**13/13**	7/8	**7/7**	**13/13**	**8/8**	**7/7**	+0.000	**27/28**

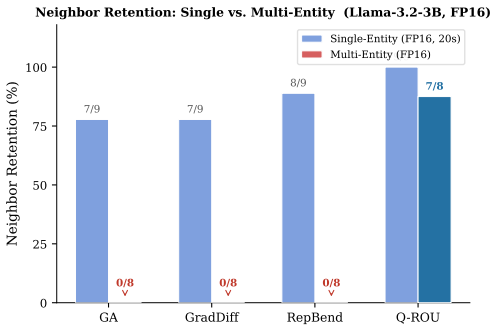

Figure 3: Neighbor retention collapse from single-entity to multi-entity unlearning. All three baselines drop from 78–89% to 0%, while Q-ROU maintains high retention (87.5–100%) in both settings. This transition serves as a litmus test for method selectivity.

single neighbor case that transitions from pass to fail under FP16 at 40 steps reflects the inherent tension between complete target removal and neighbor preservation; the sensitivity analysis (Section C.7) shows that increasing λ_a to 120 recovers this case, and $\lambda_a = 160$ achieves full 28/28 in both precisions while preserving 13/13 target removal.

This transition from single-entity to multi-entity serves as a litmus test for selectivity (Fig. 3). Methods that appear adequate on single-entity tasks are exposed as fundamentally lacking when the gradient pressure from multiple forget domains is applied simultaneously. The effect also scales with model capacity: on 0.5B, even Q-ROU struggles in multi-entity settings under quantization (FP NBR 5/8 at 40 steps, INT4 NBR 1/8), because knowledge at smaller scale is concentrated in fewer layers, leaving insufficient capacity for surgical separation. On 3B, the distributed knowledge representation enables AR to maintain its protective constraint even under 18-text simultaneous forgetting, achieving complete target removal at 40 steps while retaining 7/8 neighbor (FP16) and 8/8 neighbor (INT4) knowledge.

Table 18 extends the comparison to include AR-transplanted baselines at 40 steps, revealing that even the best +AR baseline (RB+AR at 8/13 INT4 target) remains well below Q-ROU's 13/13. The gap in total score (Q-ROU: 27/28 vs GA+AR: 22/28 in FP16) establishes Q-ROU's superiority as the full-

Table 18: Extended multi-entity comparison at 40 steps, including AR-transplanted baselines under single-GPU execution. Q-ROU achieves the highest total score by at least 5 points.

Method	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	FP Total
GA	13/13	0/8	0/7	13/13	0/8	0/7	13/28
GradDiff	13/13	0/8	4/7	13/13	0/8	4/7	17/28
RepBend	13/13	0/8	6/7	13/13	0/8	6/7	19/28
GA+AR	7/13	8/8	7/7	7/13	8/8	7/7	22/28
GD+AR	7/13	8/8	7/7	7/13	8/8	7/7	22/28
RB+AR	7/13	8/8	7/7	8/13	8/8	7/7	22/28
Q-ROU	**13/13**	7/8	**7/7**	**13/13**	**8/8**	**7/7**	**27/28**

framework integration of AR, SLUG, QuantNoise, and the KL-to-uniform forget loss.

C.3 LoRA-Concentration and Post-Training Recurrence Audits

We add two targeted audit extensions to sharpen the distinction between quantization-robust unlearning and genuinely selective, deployment-audited unlearning. The first asks whether a QR-LoRA-style low-rank update, which concentrates the edit into adapter parameters, can solve the same 28-probe multi-entity task without semantic-neighbor failure. The second asks whether target facts remain suppressed after short benign relearning on non-target texts, addressing the ‘‘unlearning or obfuscation’’ concern raised by recent metric and intrinsic-trace work. The second extension adds a PTI hardening pass that explicitly trains the post-Q-ROU model against bounded future-update directions. All runs use Llama-3.2-3B, 40 unlearning steps, the same 13/8/7 target-neighbor-general probe split, and FP16 plus simulated INT4 evaluation.

Calibration. In this audit, the lower 3B coefficient $\lambda_f = 2$ was too weak for the final 40-step multi-entity operating point: it yielded only 1/13 target pass for Q-ROU while keeping 8/8 neighbor and 7/7 general retention. A short calibration sweep showed that $\lambda_f = 10$ reached 6/13–7/13 target pass, while $\lambda_f = 20$ reached 13/13 target pass in both FP16 and INT4 with 8/8 neighbor and 7/7 general retention on seed 42. We therefore use $\lambda_f = 20$, $\lambda_r = 15$, and $\lambda_a = 80$ for the LoRA and memory-jog audits, changing only the target-pressure coefficient relative to the default profile.

LoRA-concentration audit. Table 19 reports the multi-seed LoRA audit over seeds 11, 22, and 33. The broad LoRA scope inserts rank-16 adapters in attention and MLP projections, while slug_down is a much smaller adapter footprint aligned with the SLUG down-projection intuition. The result separates three regimes. LoRA-GA broad deletes targets and survives simulated INT4, but destroys semantic neighbors; LoRA-GA+AR and LoRA-KL+AR broad preserve neighbors but do not match target deletion; and slug_down LoRA preserves neighbors but loses much of the deletion power. Thus low-rank concentration is a useful quantization-robustness mechanism, but it does not by itself solve semantic-neighbor anchoring or multi-entity target suppression. A direct low-bit reevaluation on the broad LoRA baselines with fake INT4 and BnB NF4 first confirmed the same qualitative ordering on seed 42, and a three-seed evaluation (42/43/44) preserved that ordering without reversals: only broad LoRA-GA remains target-strong, while LoRA-GA+AR and LoRA-KL+AR remain retention-heavy under both low-bit regimes (Table 20).

Table 19: Follow-up LoRA-concentration audit on the 28-probe Llama-3.2-3B multi-entity suite (40 steps, seeds 11/22/33). Values are averages over seeds.

Kind	Method/scope	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN
Full	GA+AR	8.00/13	8.00/8	7.00/7	8.00/13	8.00/8	7.00/7
Full	**Q-ROU**	**13.00/13**	7.00/8	7.00/7	**13.00/13**	**8.00/8**	7.00/7
LoRA	GA, broad	**13.00/13**	1.00/8	3.67/7	**13.00/13**	0.33/8	2.33/7
LoRA	GA+AR, broad	1.00/13	**8.00/8**	7.00/7	1.00/13	**8.00/8**	7.00/7
LoRA	KL+AR, broad	7.00/13	**8.00/8**	7.00/7	7.33/13	**8.00/8**	7.00/7
LoRA	GA, slug_down	6.67/13	**8.00/8**	7.00/7	7.00/13	**8.00/8**	7.00/7
LoRA	GA+AR, slug_down	1.00/13	**8.00/8**	7.00/7	1.67/13	**8.00/8**	7.00/7
LoRA	KL+AR, slug_down	1.00/13	**8.00/8**	7.00/7	2.00/13	**8.00/8**	7.00/7

Table 20: Direct fake-INT4 and NF4 evaluation on broad LoRA baselines (Llama-3.2-3B, 40 steps, seed 42). This is a focused low-bit confirmation, not a replacement for the multi-seed simulated-INT4 audit in Table 19.

Method	FP16			Fake INT4			NF4		
	TGT	NBR	GEN	TGT	NBR	GEN	TGT	NBR	GEN
LoRA-GA, broad	13/13	3/8	3/7	13/13	0/8	4/7	13/13	5/8	5/7
LoRA-GA+AR, broad	1/13	8/8	7/7	1/13	8/8	7/7	1/13	8/8	7/7
LoRA-KL+AR, broad	2/13	8/8	7/7	3/13	8/8	7/7	1/13	8/8	7/7

Memory-jog audit. Table 21 reports benign relearning after unlearning. We first train the source method, then fine-tune briefly on neighbor texts, general texts, or a mixed retain set for 5, 10, or 20 steps. The table should be read as a recurrence audit, not as a standard leaderboard: methods with weaker initial forgetting have less room to ‘‘resurface’’, so target probability after jog is as important as the pass count. Q-ROU remains stable under short general-only relearning (13/13 after 5 and 10 steps) but neighbor/mixed jogs recover a subset of target probes by 20 steps. Even in these cases,

Table 21: Benign relearning / memory-jog audit after unlearning (Llama-3.2-3B, seed 42). ‘‘After’’ denotes target pass immediately after unlearning; ‘‘Jog’’ denotes target pass after benign relearning.

Method	Jog data	Steps	After TGT	Jog TGT	Resurfaced	Target mean after jog
Q-ROU	neighbor	5	13/13	11/13	2	0.00465
Q-ROU	neighbor	10	13/13	10/13	3	0.00782
Q-ROU	neighbor	20	13/13	8/13	5	0.01195
Q-ROU	general	5	13/13	13/13	0	0.00242
Q-ROU	general	10	13/13	13/13	0	0.00325
Q-ROU	general	20	13/13	10/13	3	0.00432
Q-ROU	mixed	5	13/13	11/13	2	0.00442
Q-ROU	mixed	10	13/13	10/13	3	0.00781
Q-ROU	mixed	20	13/13	7/13	6	0.01301
GA+AR	neighbor	5	8/13	6/13	2	0.08439
GA+AR	neighbor	10	8/13	6/13	2	0.08979
GA+AR	neighbor	20	8/13	8/13	1	0.09462
GA+AR	general	5	8/13	8/13	1	0.06592
GA+AR	general	10	8/13	8/13	1	0.07279
GA+AR	general	20	8/13	8/13	1	0.08039
GA+AR	mixed	5	8/13	7/13	2	0.08090
GA+AR	mixed	10	8/13	7/13	2	0.08914
GA+AR	mixed	20	8/13	9/13	1	0.09512
LoRA-KL+AR	neighbor	5	7/13	7/13	0	0.06108
LoRA-KL+AR	neighbor	10	7/13	7/13	0	0.08715
LoRA-KL+AR	neighbor	20	7/13	6/13	1	0.10843
LoRA-KL+AR	general	5	7/13	8/13	0	0.04597
LoRA-KL+AR	general	10	7/13	8/13	0	0.06594
LoRA-KL+AR	general	20	7/13	8/13	0	0.08666
LoRA-KL+AR	mixed	5	7/13	7/13	0	0.06194
LoRA-KL+AR	mixed	10	7/13	8/13	0	0.09189
LoRA-KL+AR	mixed	20	7/13	7/13	1	0.12275

Table 22: Post-training recurrence audit on Llama-3.2-3B (Stage 3, seeds 11/22/33). After TGT is immediate post-unlearning target pass count. Worst Jog reports the worst target pass count across neighbor, general, mixed, and instruction-proxy jogs at 5/10/20/50 steps. Lower max target mean indicates weaker recurrence mass.

Method	After TGT	Worst Jog	Max mean	Max resurfaced	Worst N/G
Q-ROU	13/13	7/13	0.0087	6	7/8 / 5/7
Q-ROU+PTI	**13/13**	**12/13**	**0.0031**	**1**	7/8 / 5/7
GA+AR	7/13	5/13	0.1707	2	8/8 / 6/7
LoRA-KL+AR	7/13	6/13	0.1486	1	8/8 / 5/7

Table 23: PTI hardening sweep on Llama-3.2-3B (Stage 2, seed 42). All settings retain 13/13 immediate target forgetting and 13/13 target pass under every audited jog; $\rho = 0.02$ gives the lowest maximum target recurrence mass in this sweep.

ρ	After TGT	Worst Jog	Max target mean	Max resurfaced
0.005	13/13	13/13	0.0029	0
0.02	**13/13**	**13/13**	**0.0022**	**0**
0.05	13/13	13/13	0.0028	0

the target probability mass after jog remains far below GA+AR and LoRA-KL+AR, indicating weaker recurrence rather than strong persistence under subsequent tuning.

Post-training-invariant hardening. The memory-jog audit motivates a stronger deployment question: can the unlearned model remain suppressed after small, non-target post-training updates that were not part of the original unlearning objective? We therefore add Q-ROU+PTI, a short hardening pass after the base Q-ROU edit. PTI uses bounded future-update proxy directions (neighbor, mixed, and instruction-style data) to reduce the overlap between target-relearning directions and plausible future update directions. The coefficient sweep in Table 23 identifies $\rho = 0.02$ as the best value among the tested settings. The three-seed recurrence audit in Table 22 then shows the central effect: base Q-ROU reaches 13/13 immediate target forgetting across seeds but can fall to 7/13 under the broadest jog family, whereas Q-ROU+PTI keeps the worst post-jog count at 12/13 and cuts maximum target recurrence mass from 0.008724 to 0.003147. A proxy-only extension (Table 24) isolates instruction- and preference-style future updates. Across the three-seed evaluation, base Q-ROU produces only a single 12/13 case under the harshest instruction-proxy setting, while Q-ROU+PTI keeps 13/13 on all seeds and lowers the worst target mean probability on every seed.

We then ran a broad five-mode audit on the same Llama family, extending the jog set to neighbor, general, mixed, instruction proxy, and preference proxy updates at 5/10/20/50 steps. Table 25 shows that this broader audit is harsher than the narrow recurrence study, but the PTI direction still survives: base Q-ROU starts at 11/13 and falls to worst 6/13, while Q-ROU+PTI starts at 13/13 and improves the worst broad-jog count to 8/13 while preserving complete 13/13 suppression on the proxy subset. We then reevaluated the 50-step Llama checkpoints for the four hardest audited modes (‘neighbor’, ‘general’, ‘mixed’, ‘instruction proxy’) under real BnB NF4. Table 26 shows that the deployed-NF4 ordering not only survives but slightly widens on this subset: base Q-ROU drops to worst 5/13 at mixed/50, whereas Q-ROU+PTI reaches worst 10/13 and keeps 13/13 on all instruction-proxy slices. All rows in this subsection use the same deployed NF4 conversion path for consistency.

Interpretation. These audits sharpen the main claim. Q-ROU should not be framed as simply ‘‘another quantization-robust unlearning method.’’ The stronger claim is that deployment unlearning requires the conjunction of semantic-neighbor anchoring, multi-entity target suppression, low-bit stability, and generation/relearning audits. QR-LoRA-style concentration addresses the low-bit stability component but can fail the semantic-neighbor axis; AR-constrained LoRA restores that axis but

Table 24: Proxy-only post-training extension on Llama-3.2-3B (seed 42). The jog data are limited to instruction- and preference-style proxies at 5/10/20/50 steps, so this table should be read as a focused extension of Table 22, not as a replacement for the broader multi-jog audit.

Method	After TGT	Worst Proxy Jog	Worst mode	Max resurfaced	Worst N/G
Q-ROU	13/13	12/13	instr. proxy / 50	1	8/8 / 5/7
Q-ROU+PTI	**13/13**	**13/13**	instr. proxy / 50	**0**	8/8 / 5/7

Table 25: Broad post-training recurrence audit on Llama-3.2-3B (three seeds). This extension expands the jog family to neighbor, general, mixed, instruction proxy, and preference proxy at 5/10/20/50 steps. The PTI gain remains real, though smaller than in the narrower recurrence audit.

Method	Post-TGT	Worst broad	Worst case	Max mean	Worst proxy	Proxy worst
Q-ROU	11/13	6/13	mixed / 50	0.0339	9/13	instr. proxy / 50
Q-ROU+PTI	**13/13**	**8/13**	mixed / 50	**0.0096**	**13/13**	**pref. proxy / 50**

under-forgets; memory-jogging shows that even strong immediate deletion should be accompanied by recurrence reporting; and PTI hardening demonstrates that bounded recurrence robustness can be improved without abandoning the original Q-ROU deletion-and-retention operating point.

Step Sweep on Expanded 65-Probe Set. Table 27 extends the step-Pareto analysis to the expanded evaluation set. At 20 steps, both selective methods are still under-saturated, with GA+AR slightly ahead on target removal (11/25 vs. 8/25). By 40 steps, Q-ROU reaches full 25/25 target suppression while GA+AR remains at 13/25. At 60 steps, GA+AR partially closes the target gap (22/25) but still trails Q-ROU, whose neighbor retention remains stable (17/20 FP16) and whose general behavior stays intact throughout.

TOFU Multi-Split Comparison. Table 28 evaluates both methods across all three TOFU subsets. Q-ROU outperforms GA+AR on forgetting and leakage across all three subsets under the fixed 160-step budget, but the hardest forget10 row should be read as a transitional fixed-budget checkpoint rather than as a stable operating ceiling. The forget05/forget10 values come from final hard-split evaluations with explicit retain splits; long-horizon sweeps are used solely for interpretation. A three-seed confirmation on the tuned hard-split operating points (240 steps, lf40_s240) gives mean results of 99.83% Forget@0.01, 1.67% leakage, 98.33% utility, and 94.67% truth on forget05, and 94.42% Forget@0.01, 4.92% leakage, 90.83% utility, and 87.33% truth on forget10. We do not fold these tuned rows into the fixed-budget main table, but they show that the tuned hard-split trade-off is not a single-seed artifact.

As a smaller-model calibration check, we also evaluated the TOFU matrix on Llama-3.2-3B with explicit retain splits. Table 29 reports the resulting rows and additional operating-point sweeps. We treat the 80-step rows as *matched-budget diagnostics*: they show that Q-ROU is decisive on forget01, clearly ahead on forget05 forgetting/leakage but not yet saturated, and not a successful operating point on forget10. The additional Q-ROU sweep rows then make the tuning effort explicit by sweeping only Q-ROU while holding the matched GA+AR reference fixed, rather than silently replacing the harder rows with a hand-picked checkpoint. On forget05, the sweep converts the under-saturated 80-step row into near-complete or complete deletion with high utility. On forget10, additional effort sharply reduces forget probability and leakage, but the cleanest hard-split point remains utility-limited (94.25% Forget@0.01 and 89.5% utility at 240 steps), while the 400-step row slightly improves pass count but worsens tail probability.

Because thresholded pass counts can be misread when shown alone, we also report the same raw forget probabilities under multiple thresholds in Table 30. This table is meant to prevent the reader from over-interpreting a single $P < 0.01$ cut: for example, the under-saturated forget10 80-step row is weak at strict thresholds but already shows a lower probability mass than GA+AR, while the additional Q-ROU sweep rows show the operating-point path from mass reduction to strict deletion. They also expose a real small-model utility boundary: stronger Q-ROU settings can recover most hard-split forget cases, but not without lowering retain utility relative to the matched GA+AR reference.

Table 26: Deployed-NF4 evaluation on the harshest 50-step slices from the broad Llama PTI audit. We reevaluate the 'neighbor', 'general', 'mixed', and 'instruction proxy' checkpoints under real BnB NF4. The PTI ordering is preserved and slightly widened on this subset.

Method	Source TGT	NF4 broad	NF4 worst case	NF4 proxy	Proxy worst	NF4 path
Q-ROU	11/13	5/13	mixed / 50	9/13	instr. proxy / 50	BnB NF4
Q-ROU+PTI	**13/13**	**10/13**	**mixed / 50**	**13/13**	**instr. proxy / 50**	**BnB NF4**

Table 27: Step sweep on the 65-probe expanded set (Llama-3.2-3B). Q-ROU is under-suppressed at 20 steps, cleanly overtakes GA+AR by 40 steps, and remains ahead at 60 steps while maintaining stable neighbor retention.

Steps	Q-ROU				GA+AR			
	FP TGT	FP NBR	I4 TGT	Δ_z	FP TGT	FP NBR	I4 TGT	Δ_z
20	8/25	20/20	10/25	−0.022	11/25	20/20	12/25	−0.030
40	**25/25**	17/20	**25/25**	−0.015	13/25	**19/20**	13/25	−0.028
60	**25/25**	17/20	**25/25**	−0.014	22/25	17/20	21/25	−0.016

Table 28: TOFU multi-split evaluation (Llama-3.1-8B, fixed 160-step budget, final hard-split evaluations for forget05/forget10).

Split	Method	Forget $\bar{P}$	Gen Leak	Forget@0.01	n
forget01	**Q-ROU**	**0.00003**	**0.0%**	**100.0%**	40
	GA+AR	0.22023	10.0%	52.5%	40
forget05	**Q-ROU**	**0.00053**	**0.0%**	**100.0%**	200
	GA+AR	0.41320	34.5%	11.0%	200
forget10	**Q-ROU**	**0.05109**	**7.0%**	**72.0%**	400
	GA+AR	0.45833	36.75%	5.75%	400

Adding the other baselines from the same final hard-split evaluation clarifies the trade-off structure. At fixed 160 steps on 8B forget10, RMU achieves stronger raw forgetting than Q-ROU (Forget@0.01 = 85.25% vs. 72.0%; leakage = 1.25% vs. 7.0%), but at the cost of severe utility/truth collapse (Utility = 12.0%, Truth = 66.67%). This is why the fixed-budget forget10 row is not, by itself, a clean all-baseline dominance result. The long-horizon sweeps below resolve that ambiguity.

C.4 Forget10 Saturation Trajectory

The main-text TOFU table reports the final 8B fixed-160 checkpoint for forget10 (72.0% Forget@0.01 in the final hard-split evaluation), but this fixed-budget number is not stable enough to be interpreted as the method's ceiling. Across standalone fixed-160 Full-Q-ROU-style runs on the same 8B split, Forget@0.01 spans 48.25% → 77.5% while utility remains relatively high. We therefore extended the hardest 8B case directly rather than relying solely on the 3B precedent.

Figure 4 plots the 8B long-horizon step sweeps on Llama-3.1-8B: two Q-ROU seeds plus matched RMU and GA+AR comparator runs, all evaluated every 32 steps to step 640. These results indicate the performance trajectory clearly. Q-ROU is materially under-saturated at step 160 on both seeds (seed 42: 40.75%; seed 43: 47.50% Forget@0.01), then enters a high-performing regime between roughly steps 224 and 416. Seed 42 first reaches near-complete forgetting at step 288 (99.25%, Utility 93.0%, Leak 1.25%) and attains its best recorded score at step 384 (99.50%, Utility 97.0%, Leak 0.5%). Seed 43 reaches 100% Forget@0.01 by step 352 and attains its best recorded score at step 512 (100%, Utility 92.0%, Leak 0.0%).

The matched long-horizon comparators strengthen this interpretation. RMU peaks early (best score at step 96 with Forget@0.01 = 91.0%) but preserves only 5.0% utility and never exceeds 16.0% utility beyond step 128, degrading to Forget@0.01 = 70.0% and Leak = 5.25% by step 640. GA+AR never escapes its low-forgetting regime: across two independent matched-horizon runs, it peaks at only 6−10% Forget@0.01 around step 160 with leakage above 33% and remains there through step 640. The main lesson is therefore not ''forget10 is hard at 160 steps'', but rather ''the standard 160-step budget is under-saturated on 8B, and Q-ROU reaches a practically useful saturated regime well before the terminal 640-step checkpoint.''

In short, the old 3B rescue argument is no longer the primary evidence. The key evidence is now direct 8B long-horizon confirmation: the standard 160-step protocol underestimates Q-ROU on forget10, the high-performing regime appears substantially before the terminal 640-step checkpoint, and the exact best checkpoint is seed-sensitive within that saturated region.

TOFU-Small Cross-Model Transfer Screen (6 Methods, 2 Models). We evaluated the TOFU

Subset	Operating point	Method	Steps	λ_f	Forget@0.01	Gen leak	Utility	Truth
forget01	GA+AR ref	ga_ar	80	–	0.0%	35.0%	**100.0%**	96.0%
forget01	s80_lf20	qrou	80	20.0	**100.0%**	**2.5%**	99.5%	92.0%
forget05	GA+AR ref	ga_ar	80	–	3.5%	34.5%	**100.0%**	90.0%
forget05	s80_lf20	qrou	80	20.0	54.5%	11.0%	97.5%	90.0%
forget05	lf40_s160	qrou	160	40.0	98.0%	3.5%	99.0%	92.0%
forget05	lf80_s160_ar60	qrou	160	80.0	99.0%	4.0%	98.0%	90.0%
forget05	lf40_s240	qrou	240	40.0	**100.0%**	**2.0%**	99.0%	92.0%
forget05	lf40_s400	qrou	400	40.0	**100.0%**	2.5%	99.0%	86.0%
forget10	GA+AR ref	ga_ar	80	–	4.0%	34.0%	**100.0%**	98.0%
forget10	s80_lf20	qrou	80	20.0	2.2%	25.2%	99.0%	96.0%
forget10	lf40_s160	qrou	160	40.0	67.2%	8.0%	87.5%	94.0%
forget10	lf80_s160_ar60	qrou	160	80.0	77.8%	9.0%	80.0%	90.0%
forget10	lf40_s240	qrou	240	40.0	94.2%	3.8%	89.5%	92.0%
forget10	lf40_s400	qrou	400	40.0	**95.5%**	**3.0%**	89.5%	94.0%

Table 29: Extended TOFU operating-point matrix on Llama-3.2-3B. The additional Q-ROU sweep rows reuse the matched GA+AR reference and probe harder forget subsets without silently replacing the baseline row.

Subset	Operating point	Method	$P < .005$	$P < .01$	$P < .02$	$P < .05$	Mean P	Max P
forget01	GA+AR ref	ga_ar	0.0%	0.0%	0.0%	0.0%	0.4174	0.9967
forget01	s80_lf20	qrou	97.5%	100.0%	100.0%	100.0%	0.0004	0.0061
forget05	GA+AR ref	ga_ar	3.5%	3.5%	5.0%	8.5%	0.3827	0.9846
forget05	s80_lf20	qrou	39.5%	54.5%	66.0%	72.5%	0.0585	0.5334
forget05	lf40_s160	qrou	97.5%	98.0%	98.0%	99.0%	0.0014	0.1115
forget05	lf80_s160_ar60	qrou	99.0%	99.0%	100.0%	100.0%	0.0003	0.0199
forget05	lf40_s240	qrou	100.0%	100.0%	100.0%	100.0%	0.0001	0.0020
forget05	lf40_s400	qrou	100.0%	100.0%	100.0%	100.0%	0.0001	0.0003
forget10	GA+AR ref	ga_ar	2.5%	4.0%	4.8%	5.0%	0.3999	0.9912
forget10	s80_lf20	qrou	1.2%	2.2%	4.5%	18.2%	0.1987	0.8167
forget10	lf40_s160	qrou	64.2%	67.2%	72.0%	77.0%	0.0453	0.4640
forget10	lf80_s160_ar60	qrou	76.2%	77.8%	78.2%	79.2%	0.0624	0.7832
forget10	lf40_s240	qrou	92.8%	94.2%	94.5%	95.0%	0.0108	0.3082
forget10	lf40_s400	qrou	95.2%	95.5%	95.5%	95.8%	0.0142	0.5552

Table 30: TOFU threshold sensitivity on Llama-3.2-3B. The table reports the same raw forget probabilities at multiple pass thresholds, making the operating-point dependence visible rather than relying only on $P < 0.01$.

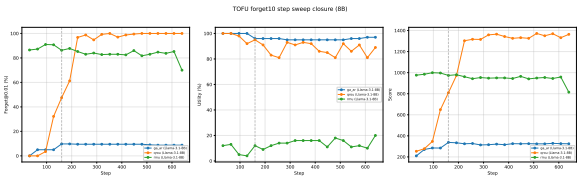

Figure 4: 8B forget10 long-horizon step sweeps on Llama-3.1-8B (640 steps, TOFU aggressive profile). Two Q-ROU seeds show that the standard 160-step budget is under-saturated and that a high-performing saturated regime appears by roughly steps 288–416, while RMU peaks early with severe utility collapse and GA+AR never exits its low-forgetting regime. The vertical dashed line indicates the default 160-step budget used in Table 1.

transfer screen on the final robust forget01 setup over two architectures (Llama-3.2-3B and Llama-3.1-8B) and six methods (Q-ROU, GA, NPO, GA+AR, RMU, RepBend), using a matched high-budget setting (160 steps for both Q-ROU and the comparison methods) with step-trace logging every 32 steps. To avoid cherry-picking, Table 32 reports the fixed terminal checkpoint (step 160) for all methods; step-trace best checkpoints are provided separately as diagnostic evidence.

Evaluation Metrics. The robust proxy evaluation utilizes the following definitions:

- **Forget@P**: The percentage of forget queries where the model's mean per-token probability for the correct answer falls below threshold P (e.g., 0.01 or 0.05). For context, the baseline Llama-3.2-3B and Llama-3.1-8B models assign mean probabilities of roughly 16.5% and 17.9% to the target answers prior to unlearning, making $P = 0.05$ a clear degradation and $P = 0.01$ a strict drop of over an order of magnitude representing near-complete sup-

Table 31: Selected 8B forget10 long-horizon checkpoints (Llama-3.1-8B, aggressive profile).

Method	Seed	Step	F@0.01	F@0.05	Leak	Utility	Truth
Q-ROU	42	160	40.75%	51.75%	11.50%	96.0%	100.0%
Q-ROU	42	288	99.25%	100.0%	1.25%	93.0%	95.83%
Q-ROU	42	384	99.50%	99.75%	0.50%	97.0%	95.83%
Q-ROU	42	640	100.0%	100.0%	0.00%	82.0%	95.83%
Q-ROU	43	160	47.50%	69.50%	11.25%	95.0%	95.83%
Q-ROU	43	352	100.0%	100.0%	1.00%	93.0%	87.50%
Q-ROU	43	512	100.0%	100.0%	0.00%	92.0%	91.67%
Q-ROU	43	640	100.0%	100.0%	0.00%	89.0%	91.67%
RMU	42	96	91.00%	94.75%	1.25%	5.0%	66.67%
RMU	42	160	86.25%	90.75%	0.75%	12.0%	66.67%
RMU	42	640	70.00%	79.50%	5.25%	20.0%	66.67%
GA+AR	42	160	9.75%	9.75%	33.75%	96.0%	91.67%
GA+AR	42	640	8.75%	9.00%	35.00%	97.0%	91.67%
GA+AR Run 2 (independent rerun)							
GA+AR	42	160	6.00%	6.00%	39.75%	100.0%	100.0%
GA+AR	42	640	5.75%	6.00%	36.75%	99.0%	91.67%

pression on the tested probes.

- **Leak**: The percentage of forget queries where unconstrained greedy text generation produces the exact answer string or > 50% of its meaningful words.
- **Utility**: The percentage of retained queries where the answer probability remains at or above 30% of the unmodified baseline model's probability.
- **Truth**: The percentage of general knowledge queries where the sequence log-probability of the true answer strictly exceeds that of a randomly mismatched false answer.

All 12 model×method runs completed without runtime failures in this sweep. Under fixed step-160 comparison, Q-ROU remains top-ranked on both models (1406.51 on 3B, 1401.52 on 8B), with strict/relaxed forgetting both at 100%, zero leakage, and high utility (98–100%). GA reaches 100% forgetting with zero leakage but collapses utility to 0% on both models; RMU maintains high forgetting but with low utility (8% on 3B, 30% on 8B); GA+AR preserves high utility but reaches only 50% forgetting with non-zero leakage (5.0% on 3B, 7.5% on 8B) and much higher compute cost.

For score interpretation, the benchmark uses

$$\begin{aligned}\text{Score} = {} & 1000 \cdot \text{Forget} - 450 \cdot \text{Leak} \\ & + 260 \cdot \text{Utility} + 120 \cdot \text{Truth} \\ & + 40 \cdot \min(\text{RetainRatio}, 1.5),\end{aligned}$$

where RetainRatio is the mean retain probability ratio used in the evaluation pipeline; this weighting heavily rewards forgetting. For readability, Table 32 reports Forget/Leak/Utility/Truth in %, but the score is computed from their unit-interval forms (0–1), consistent with the scoring rule used throughout the TOFU evaluation. Using practical guardrails (utility ≥ 80%, leak ≤ 5%), Q-ROU is the only method that remains feasible on both models in this run. Because this is still a single-seed screen (seed 42), we treat it as structured diagnostic evidence rather than a final robustness estimate.

Table 33 shows heterogeneous best-step locations across methods and models. Q-ROU peaks at step 128 on both models, whereas GA and NPO peak later on 8B, and RMU/RepBend peak earlier (step 64) on both models. This pattern further supports reporting fixed-budget terminal checkpoints for primary comparisons and treating per-method best-step traces as secondary diagnostics.

C.5 Mechanism Isolation: Active Retention Transplant

To isolate the contribution of Active Retention, we conduct a transplant experiment: adding AR ($\lambda_a = 80$, KL-divergence on pre-computed neighbor logits) to each baseline, and removing it from Q-ROU. Table 34 presents the full results including both FP and INT4 evaluations.

The results establish three findings. First, **AR is necessary for selective forgetting in Q-ROU**: removing AR causes neighbor retention to collapse from 8/8 to 2/8 (FP) and 0/8 (INT4), while target removal accelerates from 4/13 to 13/13 at 20 steps. This confirms that AR acts as a precision brake—without it, Q-ROU's forget loss overpowers the remaining regularization components (EWC, orthogonality, margin), producing aggressive but non-selective forgetting qualitatively similar to the baselines. The 2/8 residual neighbor retention (vs 0/8 for baselines without AR) reflects the partial protection from SLUG and EWC, but this is insufficient for reliable deployment.

Second, **AR is sufficient to prevent neighbor collapse in baselines**: adding it to all three baselines restores neighbor retention from 0/8 to 8/8 and general knowledge from 0–5/7 to 7/7, even without EWC, SLUG, orthogonality, or QuantNoise. However, this sufficiency is *neighbor-metric scoped*: it does not by itself resolve the token–generation tradeoff or semantic leakage patterns that emerge under extended training and adversarial evaluation

Table 32: TOFU-small cross-model transfer summary (2 models, 6 methods, robust profile, matched 160/160 budget). Values are reported at fixed final step 160; mean±sd across Llama-3.2-3B and Llama-3.1-8B.

Method	Forget@0.01 (%)	Forget@0.05 (%)	Leak (%)	Utility (%)	Truth (%)	Score	Train+Eval (s)
Q-ROU	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**99.00±1.00**	**97.91±2.08**	**1404.02±2.49**	231.20±45.59
GA	100.00±0.00	100.00±0.00	0.00±0.00	0.00±0.00	27.08±6.25	1032.50±7.50	**213.78±54.40**
NPO	35.00±17.50	62.50±30.00	0.00±0.00	46.00±35.00	93.75±6.25	593.77±70.13	232.62±50.92
GA+AR	50.00±0.00	50.00±0.00	6.25±1.25	98.00±1.00	97.91±2.08	880.20±11.89	678.04±103.25
RMU	96.25±3.75	98.75±1.25	0.00±0.00	19.00±11.00	68.75±6.25	1099.65±1.17	228.09±59.95
RepBend	60.00±0.00	63.75±3.75	0.00±0.00	68.00±2.00	77.09±2.08	889.55±3.39	225.55±63.88

Table 33: Step-trace best checkpoints by method (32-step interval; checkpoints at 32/64/96/128/160). Final cross-method comparison in Table 32 remains fixed at step 160 for fairness.

Method	3B Best Step	3B Best Score	3B Score@160	8B Best Step	8B Best Score	8B Score@160
Q-ROU	128	1411.46	1406.51	128	1401.88	1401.52
GA	32	1030.00	1025.00	160	1040.00	1040.00
NPO	128	526.24	523.64	160	663.90	663.90
GA+AR	160	892.08	892.08	128	868.55	868.31
RMU	64	1098.76	1098.48	64	1106.57	1100.82
RepBend	64	894.73	886.16	64	898.30	892.94

Table 34: Mechanism isolation via Active Retention transplant (3B multi-entity, 20 steps, single-GPU execution). Removing AR from Q-ROU causes NBR collapse to 2/8 (FP) and 0/8 (INT4); adding AR to baselines recovers NBR from 0/8 to 8/8. INT4 target forgetting often matches or exceeds FP while retention remains locked.

Method	AR	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z
GA	✗	**13/13**	0/8	0/7	**13/13**	0/8	0/7	+0.000
GA+AR	✓	4/13	**8/8**	**7/7**	**6/13**	**8/8**	**7/7**	−0.021
GD	✗	**13/13**	0/8	4/7	**13/13**	0/8	4/7	+0.000
GD+AR	✓	5/13	**8/8**	**7/7**	**7/13**	**8/8**	**7/7**	−0.018
RB	✗	**13/13**	0/8	5/7	**13/13**	0/8	5/7	+0.000
RB+AR	✓	5/13	**8/8**	**7/7**	**7/13**	**8/8**	**7/7**	−0.015
Q-ROU	✓	4/13	**8/8**	**7/7**	4/13	**8/8**	**7/7**	−0.021
Q-ROU −AR	✗	**13/13**	2/8	**7/7**	**13/13**	0/8	**7/7**	+0.000

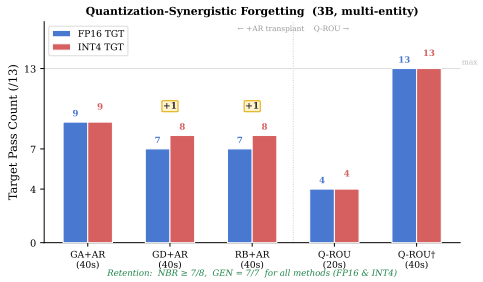

Figure 5: Low-bit target drift on the 3B multi-entity audit. AR-equipped baselines show small target-side INT4 gains while maintaining retention, whereas converged Q-ROU reaches the same complete target suppression in FP16 and INT4.

(Section D.4; Section D.3).

Third, and perhaps most importantly, **INT4 quantization can selectively enhance target forgetting when AR is present**. Fig. 5 visualizes this pattern across all +AR baselines (40 steps) alongside Q-ROU at both 20 and 40 steps: comparing FP TGT and I4 TGT bars reveals consistent enhancement (+0 to +1 cases) under AR while retention remains stable, and Q-ROU† (40 steps) achieving complete target removal (13/13) in both precisions.

This *quantization-synergistic forgetting* arises because AR constrains the parameter configuration to a structurally stable region for retention. The small perturbations from INT4 rounding can then only meaningfully affect already-borderline target probabilities, nudging them below the $P < 0.01$ removal threshold without disturbing the firmly-anchored neighbor outputs. Full Q-ROU at 20 steps shows matched FP/INT4 target counts (4/13) because QuantNoise has already flattened the loss landscape with respect to quantization perturbations, leaving no borderline cases for INT4 to push. At 40 steps, Q-ROU achieves 13/13 target in both precisions (fully converged, no borderline cases) with near-zero zombie delta ($\Delta_z \approx 0$), confirming that QuantNoise creates near-identical FP/INT4 behavior throughout the optimization trajectory.

The +AR methods show reduced target removal at 20 steps (4–5/13 vs 13/13 without AR), reflecting AR's deliberate brake on forgetting speed.

To verify convergence, we extend the transplant to 40 steps (Table 18). At this duration, GA+AR and GD+AR each reach 7/13 TGT in both FP and INT4 (with 8/8 NBR and 7/7 GEN), while RB+AR

reaches 7/13 in FP and 8/13 in INT4 with the same 8/8 NBR and 7/7 GEN pattern. Q-ROU at 40 steps surpasses all +AR baselines by 5 target cases against the strongest baseline (13/13 vs 8/13) while trading at most one neighbor case under INT4. This demonstrates that Q-ROU's additional components (KL-to-uniform forget loss, QuantNoise, orthogonality, margin) provide a substantial advantage beyond what AR alone contributes. The quantization-synergistic pattern persists in +AR baselines at +0 to +1 case under INT4 while retention remains locked. As a control, GA *without* AR at 40 steps destroys not only all neighbor knowledge (0/8) but also all general knowledge (0/7)—every evaluation prompt produces near-zero probability, confirming total model collapse. This 40-step control eliminates the alternative explanation that baselines fail merely due to insufficient training: extended training *worsens* the collapse when AR is absent, while AR-equipped methods converge to strong forgetting with intact retention.

C.6 Ablation Study

The main text already summarizes the 3B component ablation covering the core components (AR, QuantNoise, Orthogonality, EWC, Margin). Here, we provide the 8B scale ablation on TOFU and the adaptive mechanism ablation.

8B Scale Ablation on TOFU. At 8B (Table 35: TOFU forget01, 160 steps), all six ablation variants achieve identical Forget@0.01 = 100.0% with $|\Delta S| \leq 2.54$. Margin and EWC removal produce exactly zero measurable effect on superficial pass counts, confirming that parameter redundancy handles capacity demands at 8B. However, these auxiliary losses act as a geometric safety net (defense-in-depth): for instance, the Orthogonality constraint implicitly bounds the rotation of general-knowledge subspaces (Appendix Theorem 2). Thus, while they may appear redundant on easy fixed-budget checkpoints, they provide geometric stability that becomes relevant under harder regimes and longer horizons (Appendix C.4; Appendix C.16). However, on harder forget10 at longer horizons, utility ordering becomes seed-sensitive (Appendix C.6).

Adaptive Mechanism Ablation. Table 36 reports the four-way ablation of adaptive mechanisms on the 65-probe expanded set.

We emphasize that this ablation is intentionally reported at fixed-budget binary endpoints (T/N/G pass counts). Adaptive variants are primarily designed to improve robustness properties that are only weakly reflected in these endpoints (e.g., threshold sensitivity and precision brittleness), and should therefore be interpreted alongside the threshold robustness analyses in Section C.12.

For the adaptive-mechanism subtable (Standard / Adaptive-QN / Adaptive-AR / Full Adaptive), we verified the exported LaTeX rows against the underlying count totals (25/20/20 in FP and INT4, plus legacy 13/8/7 regression metadata for qrou_40s).

Table 35: 8B component ablation on TOFU forget01 (Llama-3.1-8B, 160 steps, seed=42). All variants achieve identical Forget@0.01 = 100%; score variation is ≤2.54.

Configuration	F@0.01	F@0.05	Leak	Utility	Truth	Score
Full Q-ROU	100.0%	100.0%	0.0%	100.0%	91.7%	1396.01
−Orthogonality	100.0%	100.0%	0.0%	99.0%	91.7%	1393.47
−Margin	100.0%	100.0%	0.0%	100.0%	91.7%	1396.01
−EWC	100.0%	100.0%	0.0%	100.0%	91.7%	1396.20
−Ortho−Margin	100.0%	100.0%	0.0%	99.0%	91.7%	1393.47
−Ortho−Margin−EWC	100.0%	100.0%	0.0%	100.0%	91.7%	1395.94

3B Component-Level Ablation Details. Table 37 supplements the main-text discussion of scale-dependent component contribution with continuous mean-probability values on Llama-3.2-3B (20 steps, single-entity evaluation). All four configurations produce *identical* pass counts (FP16: TGT 2/7, NBR 9/9, GEN 7/7; INT4: TGT 3/7, NBR 9/9, GEN 7/7), confirming the same parameter-redundancy reading summarized in the main text. At the $\bar{p}_{\text{TGT}}$ level, removing EWC actually *decreases* target probability (0.012 vs. 0.025 for Full), because EWC constrains parameter drift away from the pre-training manifold, partially counteracting the forget objective. The −Margin configuration is *numerically identical* to Full Q-ROU (matching to six decimal places), indicating that at 3B scale, the margin-based separation loss contributes zero incremental effect within the SLUG-constrained parameter space.

8B Component-Level Ablation on TOFU. Table 38 extends the component ablation to Llama-3.1-8B using the standardized TOFU forget01 benchmark (40 examples, 160 steps, seed=42). Six configurations are tested: Full Q-ROU, and five progressive removals of auxiliary components (−Ortho, −Margin, −EWC, −Ortho−Margin, −Ortho−Margin−EWC).

All configurations achieve *identical* Forget@0.01

Table 36: Adaptive mechanism ablation on the 65-probe expanded set (Llama-3.2-3B, 40 steps). Pass-count differences are small, positioning adaptive AR/QN primarily as stability instrumentation.

Configuration	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z
Standard (Fixed AR, Fixed QN)	25/25	17/20	20/20	25/25	16/20	18/20	−0.018
Adaptive QN Only ($\beta = 0.4$)	25/25	18/20	20/20	25/25	17/20	18/20	−0.017
Adaptive AR Only (max $\lambda_a = 80$)	25/25	17/20	20/20	25/25	16/20	18/20	−0.018
Full Adaptive (diagnostic variant)	**24/25**	**17/20**	**20/20**	**25/25**	**16/20**	**18/20**	**−0.016**

Table 37: 3B component ablation: continuous mean-probability values (Llama-3.2-3B, 20 steps, single-entity). All configurations yield identical binary pass counts.

Configuration	FP $\bar{p}_{\text{TGT}}$	FP $\bar{p}_{\text{NBR}}$	I4 $\bar{p}_{\text{TGT}}$	I4 $\bar{p}_{\text{NBR}}$	Δ_z
Full Q-ROU	0.025	0.937	0.023	0.921	−0.002
Q-ROU −Orthogonality	0.025	0.937	0.023	0.922	−0.002
Q-ROU −EWC	**0.012**	0.922	**0.011**	0.901	−0.001
Q-ROU −Margin	0.025	0.937	0.023	0.921	−0.002

= 100.0%, Forget@0.05 = 100.0%, and zero generation leakage, confirming that at 8B scale, the primary forget and retain objectives are sufficient for complete target suppression on this split even without auxiliary regularization.

Step-curve analysis (5 checkpoints at 32-step intervals) reveals that the Full configuration already reaches 100% Forget@0.01 at step 96 on this split, leaving no room for differentiation. The only measurable difference is a 1% Utility drop ($\Delta = -1.0\%$) when Orthogonality is removed, propagating to a $\Delta S = -2.54$ score reduction. This is attributable to Ortho's ℓ_2-regularization effect partially constraining parameter drift away from the retain manifold. Margin and EWC produce exactly zero measurable effect at this scale ($\Delta S = 0.00$ and +0.19 respectively).

The most aggressive variant (−Ortho−Margin−EWC; effectively Q-ROU reduced to forget + retain + AR + SLUG only) achieves Score = 1395.94 versus Full's 1396.01, a Δ of just −0.08. This provides additional evidence for the parameter-redundancy hypothesis: at 8B scale with 40-example forget sets, the core objectives alone carry all measurable performance.

These values come from the same controlled 8B easy-split ablation evaluations summarized in this appendix.

This easy-split redundancy story does not transfer unchanged to harder TOFU settings. On 8B forget10, the fixed-160 hard-split comparison initially suggested a strong ordering Full > Core > Core-no-QuantNoise, but the longer-horizon evaluations refine that picture. Two independent 384-step runs (Table 39) show that all three variants achieve near-complete forgetting (≥96.5% Forget@0.01), but the relative utility ordering is *not stable across runs*: Run 1 places Full highest (Utility 95%) and Core-no-QuantNoise lowest (87%), while Run 2 reverses this pattern (Core-no-QuantNoise 94%, Full 96%, Core 90%). The strongest surviving component conclusion is therefore not about any single auxiliary term, but rather that (i) all three variants converge to near-complete forgetting at extended horizon, and (ii) the relative utility ranking is seed-sensitive, so single-run component claims should be treated with caution on this harder split.

C.7 Hyperparameter Sensitivity

The Active Retention coefficient λ_a is the primary hyperparameter controlling the trade-off between forgetting aggressiveness and neighbor protection. Table 40 reports Q-ROU performance on Llama-3.2-3B multi-entity unlearning at 40 steps across $\lambda_a \in \{40, 80, 120, 160\}$.

The results reveal a smooth trade-off. Lower λ_a values (40–80) permit aggressive forgetting (13/13 in both FP and INT4) with one neighbor case dropped (7/8) in both precisions. $\lambda_a = 120$ recovers that neighbor case *selectively* under INT4 (8/8 vs. its own FP score of 7/8), while maintaining 13/13 target removal. At $\lambda_a = 160$, both precisions retain 8/8 neighbors while preserving 13/13 target removal.

Crucially, total pass counts remain at 27–28/28 across the entire four-fold range of λ_a, confirming that the method does not require extensive hyperparameter tuning. We use $\lambda_a = 80$ as the default throughout this paper for consistency across experiments, while noting that $\lambda_a = 160$ maximizes neighbor retention in this sweep without sacrificing target removal.

Table 38: 8B component ablation on TOFU forget01 (Llama-3.1-8B, 160 steps, seed=42). All variants achieve identical perfect forgetting; score variation is ≤2.54.

Configuration	F@0.01	F@0.05	Leak	Utility	Truth	Score	ΔS
Full Q-ROU	100.0%	100.0%	0.0%	100.0%	91.7%	1396.01	—
−Orthogonality	100.0%	100.0%	0.0%	99.0%	91.7%	1393.47	−2.54
−Margin	100.0%	100.0%	0.0%	100.0%	91.7%	1396.01	0.00
−EWC	100.0%	100.0%	0.0%	100.0%	91.7%	1396.20	+0.19
−Ortho−Margin	100.0%	100.0%	0.0%	99.0%	91.7%	1393.47	−2.54
−Ortho−Margin−EWC	100.0%	100.0%	0.0%	100.0%	91.7%	1395.94	−0.08

Table 39: 8B hard-split component evaluations on TOFU forget10 at a longer horizon (Llama-3.1-8B, 384 steps, seed=42). Two independent runs show that all variants achieve near-complete forgetting, but the relative utility ordering reverses across runs.

Run	Configuration	F@0.01	F@0.05	Leak	Utility	Truth	Score
1	Full Q-ROU	99.0%	100.0%	0.25%	95.0%	95.8%	1374.64
1	Core	100.0%	100.0%	0.25%	91.0%	95.8%	1373.63
1	Core-no-QuantNoise	100.0%	100.0%	0.0%	87.0%	95.8%	1362.45
2	Full Q-ROU	96.5%	99.5%	0.5%	96.0%	100.0%	1356.48
2	Core	99.5%	100.0%	0.25%	90.0%	91.7%	1361.70
2	Core-no-QuantNoise	99.25%	100.0%	0.25%	94.0%	91.7%	1369.53

Table 40: Sensitivity to λ_a (3B multi-entity, 40 steps). The method achieves ≥12/13 target removal and ≥7/8 neighbor retention across the full range, demonstrating robustness to hyperparameter choice.

λ_a	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z
40	**13/13**	7/8	7/7	**13/13**	7/8	7/7	≈0
80	**13/13**	7/8	7/7	**13/13**	7/8	7/7	≈0
120	**13/13**	7/8	7/7	**13/13**	**8/8**	7/7	≈0
160	**13/13**	**8/8**	7/7	**13/13**	**8/8**	7/7	≈0

C.8 Computational Cost

The computational cost and efficiency metrics of Q-ROU relative to baselines are contextualized in Section D. Table 41 reports measured wall-clock costs for the larger same-work probability and generation evaluations. The fine-grained AR operating point roughly doubles the retain-anchor set and therefore costs more than default Q-ROU, but the observed per-seed training times remain in the one-to-two-minute range on the logged single-GPU runs used for that table. Table 42 then places Q-ROU, qrou_fg_ar, GA+AR, and gaar_fg_ar on the same 81-probe same-work benchmark and the same RTX 3090 hardware slice. At 80 steps, qrou_fg_ar is the most expensive operating point (787 s train, 10.4 GB peak VRAM), but it is also the only configuration in this comparison that preserves full 13/13 target suppression while materially improving same-work retention over default Q-ROU.

C.9 Expanded Evaluation

Table 43 reports a focused five-method subset of the expanded 65-probe evaluation on Llama-3.2-3B (multi-entity, 40 steps), matching the central selective-forgetting comparison discussed in the main text. In this subset, Q-ROU reaches 25/25 target removal in FP16 and remains at 25/25 under INT4 while preserving substantially more neighbor/general behavior than aggressive baselines. In contrast, GA and RepBend achieve 100% target removal only at the cost of complete neighbor collapse (0/20). Because this 40-step row is used as a central selective-forgetting result, Table 44 adds a focused three-seed confirmation for the two strongest selective methods at the same operating point.

C.10 Depth Probing and Adversarial Robustness

Standard evaluation probes primarily test direct knowledge recall. To assess whether Q-ROU achieves deeper, conceptual-level knowledge suppression, we design six structured probing protocols with 55 probes total.

Protocol A (reverse association, 12 probes) tests **one-step reverse lookup**: given a distinguishing attribute or description of the forgotten entity, can the model retrieve the entity name itself? For example, ‘‘The fictional school with moving staircases is’’ → ‘‘Hogwarts’’ inverts the standard entity→attribute

Table 41: Runtime accounting for the larger same-work audits. Values are mean ± standard deviation over seeds.

Track	Method	Seeds	Train s	Wall s	Retain texts
custom neighbor audit	GA+AR	6	26.8 ± 0.5	26.9 ± 0.5	25.0
custom neighbor audit	GA+FG-AR	6	43.0 ± 0.4	43.2 ± 0.4	57.0
custom neighbor audit	Q-ROU	6	47.2 ± 0.2	47.9 ± 0.2	25.0
custom neighbor audit	Q-ROU+FG-AR	6	85.7 ± 0.4	86.7 ± 0.4	57.0
samework generation audit	GA+AR	3	37.7 ± 0.6	37.9 ± 0.7	25.0
samework generation audit	GA+FG-AR	3	60.1 ± 0.8	60.3 ± 0.8	57.0
samework generation audit	Q-ROU	3	68.2 ± 0.4	69.2 ± 0.6	25.0
samework generation audit	Q-ROU+FG-AR	3	123.0 ± 0.6	124.2 ± 0.6	57.0

Table 42: Uniform runtime accounting on the larger same-work benchmark (81 probes, single RTX 3090, one seed). This table uses the same hardware and layer set for all four methods, making relative training cost directly comparable.

Steps	Method	Train s	Wall s	Peak VRAM MB	TGT	SW	ID	GEN
40	Q-ROU	213.3	222.5	9853.8	13/13	4/24	12/16	13/16
40	Q-ROU+FG-AR	384.3	390.6	10389.5	12/13	14/24	15/16	16/16
40	GA+AR	91.1	94.7	8548.0	7/13	16/24	15/16	16/16
40	GA+FG-AR	175.4	181.7	8934.8	1/13	21/24	16/16	16/16
80	Q-ROU	434.2	442.9	9853.8	13/13	4/24	12/16	14/16
80	Q-ROU+FG-AR	787.1	797.2	10389.5	13/13	12/24	14/16	16/16
80	GA+AR	244.9	250.8	8548.0	12/13	5/24	11/16	15/16
80	GA+FG-AR	397.7	404.1	8934.8	5/13	22/24	16/16	16/16

Table 43: Focused five-method subset of the expanded 65-probe evaluation on Llama-3.2-3B (multi-entity, 40 steps), matching the central selective-forgetting comparison from Table 3. Q-ROU reaches 100% target removal in both FP16 and INT4 while maintaining substantially higher neighbor/general retention than aggressive baselines.

Method	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z
Baseline	5/25	20/20	20/20	5/25	18/20	20/20	−0.055
Q-ROU	**25/25**	17/20	**20/20**	**25/25**	16/20	18/20	−0.018
GA+AR	13/25	**18/20**	**20/20**	13/25	15/20	19/20	−0.026
GA	**25/25**	0/20	0/20	**25/25**	0/20	0/20	+0.000
RepBend	**25/25**	0/20	12/20	**25/25**	0/20	11/20	−0.013

direction by querying attribute→entity. This tests whether the suppression of the entity's name holds when approached from the property side, rather than by naming the entity directly.

Protocol B (multistep reasoning, 10 probes) tests **multi-hop inference chains**: the probe does not name the forgotten entity or its attributes directly but instead requires the model to traverse two or more inferential steps to reach the target fact. For example, ''The sport played on broomsticks in a magical school'' → ''Quidditch'' requires the model to first identify the school (Harry Potter's Hogwarts), then retrieve the sport associated with it—neither step alone is sufficient. This tests whether knowledge that must be *inferred via intermediate concepts* is also suppressed.

Protocol C (paraphrased extraction, 15 probes) rephrases the original prompt while preserving the same target answer, testing whether surface-level lexical changes bypass suppression. *Protocol D* (reconstruction, 9 probes) provides incremental hints to test resistance to guided reconstruction. *Protocol E* (in-context priming, 6 probes) provides related context explicitly in the prompt to prime knowledge retrieval. *Protocol F* (negation-based extraction, 3 probes) uses negated or contrastive framing to elicit the target answer.

Table 45 presents the results. Q-ROU achieves **100% suppression (55/55)** across all six protocols in both FP16 and INT4, with mean target probability below 0.002 in every protocol. This result substantially strengthens the evidence for conceptual-level suppression: the method suppresses not only direct recall but also reverse associations (Protocol A), multi-hop reasoning chains (Protocol B), hint-guided reconstruction (Protocol D), and all three adversarial extraction strategies.

GA+AR provides partial but incomplete defense (43/55 in both FP16 and INT4; 78%). Its critical weakness is *in-context priming*: when relevant context is explicitly provided in the prompt, GA+AR suppresses only 3/6 probes (50%) with a mean probability of 0.204—two orders of magni-

Table 44: Three-seed confirmation of the expanded 65-probe 40-step operating point (Llama-3.2-3B). Values are mean pass counts over seeds 42/43/44; parentheses show population standard deviation. This confirms that the main 40-step ordering is stable rather than a single-seed artifact.

	FP16				INT4			
Method	TGT/25	NBR/20	GEN/20	Δ_z	TGT/25	NBR/20	GEN/20	Δ_z
Q-ROU	25.0 (0.0)	17.0 (0.0)	19.7 (0.5)	−0.0181 (0.0006)	25.0 (0.0)	15.7 (0.5)	18.0 (0.0)	−0.0181 (0.0006)
GA+AR	13.0 (0.0)	19.0 (0.0)	20.0 (0.0)	−0.0281 (0.0000)	13.0 (0.0)	15.0 (0.0)	19.0 (0.0)	−0.0281 (0.0000)

Table 45: Depth probing on Llama-3.2-3B (multi-entity, 40 steps). Q-ROU reaches 55/55 in both FP16 and INT4; GA+AR reaches 43/55.

Protocol	n	Baseline	Q-ROU FP	Q-ROU I4	GA+AR FP	GA+AR I4
A: Reverse assoc.	12	6/12	**12/12**	**12/12**	9/12	9/12
B: Multistep	10	3/10	**10/10**	**10/10**	7/10	7/10
C: Reconstruction	9	3/9	**9/9**	**9/9**	8/9	7/9
Adv: Paraphrase	15	10/15	**15/15**	**15/15**	13/15	14/15
Adv: In-context	6	0/6	**6/6**	**6/6**	3/6	3/6
Adv: Negation	3	2/3	**3/3**	**3/3**	3/3	3/3
Total	**55**	24/55	**55/55**	**55/55**	43/55	43/55

tude higher than Q-ROU's 0.001. This suggests that GA+AR achieves surface-level suppression that can be bypassed by contextual cues, while Q-ROU's additional components (KL-to-uniform forget loss, QuantNoise, orthogonality) produce deeper structural suppression resistant to guided extraction.

The baseline (pre-unlearning model) naturally suppresses 24/55 probes because some indirect probes yield low target probability even in the original model. Notably, in-context probes show 0/6 baseline suppression (all knowledge accessible), making Q-ROU's 6/6 suppression on these probes a direct measure of unlearning effectiveness.

C.11 Domain Transfer: Public Biographical Fact Evaluation

A key limitation of fictional-domain benchmarks is that real-world deployment involves biographical facts of public individuals. To assess domain transfer more credibly, we replace our earlier small check with a larger balanced audit built from publicly documented biographical facts. The new set contains twelve forget subjects (Elon Musk, Mark Zuckerberg, Jeff Bezos, Sam Altman, Jensen Huang, Steve Jobs, Bill Gates, Larry Page, Sergey Brin, Michael Dell, Reed Hastings, and Linus Torvalds), twelve neighbor subjects (Tim Cook, Satya Nadella, Sundar Pichai, Sheryl Sandberg, Jack Dorsey, Susan Wojcicki, Steve Wozniak, Paul Allen, Marc Benioff, Andy Jassy, Meg Whitman, and Marissa Mayer), and 36 general probes. Each subject contributes three stable facts, yielding 36 target probes, 36 neighbor probes, and 36 general probes in total.

Table 46 presents the matched-step long-horizon results.

Q-ROU and GA+AR are again too close at the 40-step point, so the earlier short-horizon read was indeed under-saturated. The larger balanced set clarifies the picture. At 80 steps, Q-ROU consistently reaches 35/36 target suppression while preserving 36/36 neighbor and 36/36 general counts in FP16, and the same pass counts are reproduced for seeds 42, 43, and 44. Under the same 80-step budget, GA+AR remains at 16/36 target suppression despite identical 36/36 neighbor and 36/36 general counts. At 120 steps, the gap widens further: Q-ROU reaches full 36/36 FP16 target suppression with no loss on neighbor or general probes, while GA+AR rises only to 22/36. Under INT4, the same qualitative ordering remains: Q-ROU stays at 35/36 target suppression at 80 steps and 36/36 at 120 steps, while preserving 32.7/36–33.0/36 neighbor and 36/36 general counts. The right interpretation is therefore stronger than before: on a larger and better-balanced public-biographical audit, longer-horizon matched-step sweeps reveal a clear Q-ROU advantage, not merely a fragile tie. Even so, the domain is still handcrafted rather than a benchmark-grade main result. GA without AR still collapses neighbor retention almost completely, so the boundary-anchoring role of AR clearly transfers across domains. Table 47 adds pooled Wilson intervals for the same three-seed public-biographical audit. This table is included to make clear that

Table 46: Balanced public-biographical transfer audit on Llama-3.2-3B. The 40-step point is under-saturated for both AR-based methods. At 80 steps, three seeds reproduce the same 35/36 FP16 target count with perfect 36/36 neighbor and 36/36 general retention. At 120 steps, Q-ROU reaches full 36/36 FP16 target suppression while preserving 36/36 neighbor and 36/36 general counts.

Method	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Δ_z
GA 20s	35/36	9/36	34/36	35/36	9/36	31/36	−0.017
Q-ROU 40s	13/36	36/36	36/36	15/36	33/36	36/36	−0.024
GA+AR 40s	8/36	36/36	36/36	10/36	31/36	36/36	−0.045
Q-ROU 80s (3 seeds mean)	**35.0/36**	**36.0/36**	**36.0/36**	**35.0/36**	**32.7/36**	**36.0/36**	−0.023
GA+AR 80s (3 seeds mean)	16.0/36	**36.0/36**	**36.0/36**	16.0/36	31.0/36	**36.0/36**	−0.041
Q-ROU 120s (3 seeds mean)	**36.0/36**	**36.0/36**	**36.0/36**	**36.0/36**	**33.0/36**	**36.0/36**	−0.020
GA+AR 120s (3 seeds mean)	22.0/36	**36.0/36**	**36.0/36**	19.0/36	31.0/36	**36.0/36**	−0.036

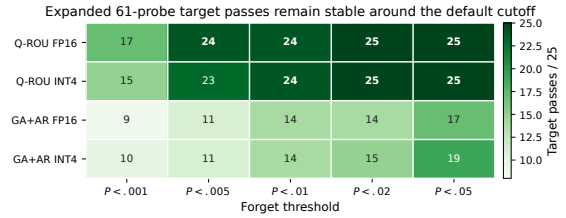

Figure 6: Threshold-sensitivity slice derived from Table 50. The key pattern is that Q-ROU already places almost all target probes below the stricter cutoffs, whereas GA+AR moves more sharply with the forget threshold.

the long-horizon Q-ROU advantage is not a single-threshold accident: the pooled target pass interval for Q-ROU 120s is near the ceiling, while GA+AR remains far lower under the matched 120-step budget.

C.12 Threshold Robustness Analysis

A natural concern with threshold-based evaluation ($P < 0.01$ for removal, ratio ≥ 0.30 for retention) is sensitivity to threshold choice. To address this, we sweep both forget threshold $\tau_f \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1\}$ and retain ratio $\tau_r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.7\}$ over a 61-probe evaluation set (25 target, 20 neighbor, 16 general), i.e., a threshold-sweep subset of the 65-probe expanded suite with four general probes omitted for this grid, producing a 6×6 grid of pass counts for both Q-ROU and GA+AR.

The bimodal distribution confirms that Q-ROU achieves thorough target suppression rather than merely pushing probabilities marginally below a chosen threshold. In contrast, GA+AR shows substantially greater threshold sensitivity (Table 49): 10/25 at $\tau_f = 0.001$, 14/25 at $\tau_f = 0.01$, and 19/25 at $\tau_f = 0.1$. The neighbor dimension shows a complementary pattern: NBR counts are independent of τ_f but vary smoothly with τ_r. Q-ROU achieves 19/20 at $\tau_r = 0.1$, 15/20 at $\tau_r = 0.3$ (default), and 11/20 at $\tau_r = 0.7$. This analysis confirms that the qualitative conclusion (Q-ROU > GA+AR for target, GA+AR > Q-ROU for neighbor, Q-ROU higher total) is stable across a two-order-of-magnitude range of both thresholds.

An additional threshold table reports both FP16 and INT4 rows in Table 50. The strict target-side conclusion is not an artifact of the default $P < 0.01$ cutoff: for Q-ROU, 24/25 FP16 target probes are already below $P < 0.005$, and all 25 are below $P < 0.02$. GA+AR is much more threshold-sensitive, moving from 11/25 at $P < 0.005$ to 17/25 at $P < 0.05$ in FP16.

C.13 Embedding Space Analysis

To investigate whether Q-ROU's target suppression manifests as structural changes in the model's internal representations—beyond output probability suppression—we measure (1) cosine similarity between baseline and post-unlearning hidden states at each SLUG layer, and (2) projection onto concept direction vectors, for four knowledge categories: HP (target), LOTR (target), SW (neighbor), and General. This representation-level analysis approach parallels the interpretability methods in linearized fine-tuning (Achille et al., 2021), which leverage the tractability of linear models to precisely compute influence functions and trace how training samples affect internal representations—though Q-ROU achieves this through direct hidden-state comparison rather than closed-form influence computation.

Table 51 presents the cosine similarity results.

At layer 26 (the deepest SLUG layer), target representations undergo dramatic change: HP cosine drops to 0.712 (28.8% representational shift) and LOTR to 0.760 (24.0% shift). In sharp contrast, SW (neighbor) cosine remains at 0.963 (3.7% shift) and

Table 47: Larger public-biographical audit stability (36 target, 36 neighbor, 36 general probes; three seeds). Counts are mean pass counts across seeds; Wilson intervals pool target passes over seeds.

Method	Prec.	TGT	NBR	GEN	TGT Wilson 95%
GA+AR 120s	FP16	22.0/36 ± 0.0	36.0/36 ± 0.0	36.0/36 ± 0.0	51.7–69.8%
GA+AR 120s	fake INT4	19.0/36 ± 0.0	31.0/36 ± 0.0	36.0/36 ± 0.0	43.4–61.9%
GA+AR 80s	FP16	16.0/36 ± 0.0	36.0/36 ± 0.0	36.0/36 ± 0.0	35.4–53.8%
GA+AR 80s	fake INT4	16.0/36 ± 0.0	31.0/36 ± 0.0	36.0/36 ± 0.0	35.4–53.8%
Q-ROU 120s	FP16	36.0/36 ± 0.0	36.0/36 ± 0.0	36.0/36 ± 0.0	96.6–100.0%
Q-ROU 120s	fake INT4	36.0/36 ± 0.0	33.0/36 ± 0.0	36.0/36 ± 0.0	96.6–100.0%
Q-ROU 80s	FP16	35.0/36 ± 0.0	36.0/36 ± 0.0	36.0/36 ± 0.0	92.1–99.1%
Q-ROU 80s	fake INT4	35.0/36 ± 0.0	32.7/36 ± 0.5	36.0/36 ± 0.0	92.1–99.1%

Table 48: Threshold robustness of Q-ROU (FP16) on Llama-3.2-3B (61 probes, 40 steps). Rows: forget threshold τ_f; columns: retain ratio τ_r. Cell format: TGT/NBR/GEN. Target pass counts are *insensitive* to τ_f (25/25 for $\tau_f \geq 0.02$), while neighbor counts depend smoothly on τ_r.

τ_f	$\tau_r = 0.1$	$\tau_r = 0.2$	$\tau_r = 0.3$	$\tau_r = 0.4$	$\tau_r = 0.5$	$\tau_r = 0.7$
0.001	16/19/16	16/16/16	16/15/16	16/14/15	16/13/15	16/11/14
0.005	23/19/16	23/16/16	23/15/16	23/14/15	23/13/15	23/11/14
0.01	**24/19/16**	24/16/16	**24/15/16**	24/14/15	24/13/15	24/11/14
0.02	25/19/16	25/16/16	25/15/16	25/14/15	25/13/15	25/11/14
0.05	25/19/16	25/16/16	25/15/16	25/14/15	25/13/15	25/11/14
0.1	25/19/16	25/16/16	25/15/16	25/14/15	25/13/15	25/11/14

General at 0.985 (1.5% shift). Non-SLUG layers (8, 12) show perfect identity (cos = 1.000), confirming that SLUG's targeted update strategy preserves all non-selected layers exactly.

This cosine gradient across SLUG layers—from near-identity at layer 17 (~0.997 for all categories) to strong divergence at layer 26 (0.712 for HP vs. 0.985 for General)—provides the most direct evidence that Q-ROU achieves *selective representation-level modification*: target knowledge representations are structurally reorganized while non-target representations remain stable. GA+AR, by comparison, produces smaller representational changes at the same layer (HP cosine = 0.922 vs. Q-ROU's 0.712), suggesting shallower structural modification consistent with its weaker target suppression performance.

Concept direction analysis reveals a reorganization-then-suppression pattern: HP projection onto the target concept direction *increases* at intermediate SLUG layers (layer 22: +2.31 from baseline) before *decreasing* at the deepest layer (layer 26: −2.12). This suggests that Q-ROU's optimization first redistributes target-related features within the SLUG subspace, then suppresses them at the final processing stage—a qualitatively different mechanism from GA+AR's uniform small decreases across all layers.

C.14 Adversarial Chain-of-Thought Evaluation

Standard evaluation measures single-token probabilities, but knowledge can potentially leak through multi-token generation even when individual token probabilities are suppressed. We design 22 adversarial chain-of-thought (CoT) probes: 10 targeting HP, 9 targeting LOTR, and 3 testing neighbor retention. Five attack strategies are employed: *direct* completion (2 probes), *stepping-stone* reasoning that builds context incrementally (10 probes), *contextual* cues that provide related information (5 probes), *reasoning* chains requiring multi-hop inference (3 probes), and *reversal* probes that approach from the answer side (2 probes). Each probe is evaluated at both token level ($P < 0.01$ for first target keyword) and generation level (greedy decoding + 3 sampling trials at $T = 0.7$, checking for target keyword presence).

Table 52 reveals a striking asymmetry. Q-ROU achieves **100% token-level suppression** (19/19 target probes) in both FP16 and INT4, and **high generation-level suppression**: 19/19 in FP16 and 18/19 in INT4 across the same five attack strategies. GA+AR matches Q-ROU on HP (10/10 token, 9–10/10 generation) but fails severely on LOTR: only 1/9 at token level and 0–1/9 at generation level. Multiple LOTR probes show probabilities exceeding 0.1 under GA+AR (e.g., ‘‘Frodo Baggins is a

Table 49: Threshold robustness of GA+AR (FP16) on the same 61-probe set. Compared to Q-ROU, GA+AR shows greater sensitivity to τ_f while achieving stronger neighbor retention at moderate τ_r.

τ_f	$\tau_r = 0.1$	$\tau_r = 0.2$	$\tau_r = 0.3$	$\tau_r = 0.4$	$\tau_r = 0.5$	$\tau_r = 0.7$
0.001	10/20/16	10/20/16	10/20/16	10/20/16	10/20/16	10/18/15
0.005	11/20/16	11/20/16	11/20/16	11/20/16	11/20/16	11/18/15
0.01	**14/20/16**	14/20/16	**14/20/16**	14/20/16	14/20/16	14/18/15
0.02	14/20/16	14/20/16	14/20/16	14/20/16	14/20/16	14/18/15
0.05	17/20/16	17/20/16	17/20/16	17/20/16	17/20/16	17/18/15
0.1	19/20/16	19/20/16	19/20/16	19/20/16	19/20/16	19/18/15

Method	Precision	$P < .001$	$P < .005$	$P < .01$	$P < .02$	$P < .05$	NBR@0.3	GEN@0.3
Q-ROU 40s	FP16	17/25	24/25	24/25	25/25	25/25	15/20	16/16
Q-ROU 40s	INT4	15/25	23/25	24/25	25/25	25/25	15/20	16/16
GA+AR 40s	FP16	9/25	11/25	14/25	14/25	17/25	20/20	16/16
GA+AR 40s	INT4	10/25	11/25	14/25	15/25	19/25	20/20	16/16

Table 50: Probe-level threshold sensitivity on the expanded 61-probe audit. Target columns vary the forget threshold while the retain ratio is fixed at 0.3; the final columns report neighbor/general retention at the default $P < 0.01$, retain-ratio 0.3 operating point.

Table 51: Cosine similarity between baseline and Q-ROU hidden states across SLUG layers (Llama-3.2-3B, 40 steps). Target representations diverge strongly at deeper SLUG layers while neighbor/general remain comparatively stable.

Layer	HP	LOTR	SW	General
8 (non-SLUG)	1.000	1.000	1.000	1.000
12 (non-SLUG)	1.000	1.000	1.000	1.000
17	0.997	0.997	0.997	0.999
19	0.979	0.975	0.991	0.998
22	0.936	0.941	0.986	0.997
24	0.893	0.908	0.979	0.995
26	**0.712**	**0.760**	**0.963**	**0.985**

Table 52: Adversarial CoT results on Llama-3.2-3B (40 steps). Token: first-keyword $P < 0.01$; Gen: no keyword in greedy + 3 sampling runs. Q-ROU achieves high target suppression at both levels while GA+AR fails on LOTR generation. (Gen: generation; I4: INT4 precision)

Method	HP (10)		LOTR (9)		Target total (19)	
	Token	Gen	Token	Gen	Token	Gen
Q-ROU FP	**10/10**	**10/10**	**9/9**	**9/9**	**19/19**	**19/19**
Q-ROU I4	**10/10**	9/10	**9/9**	**9/9**	**19/19**	18/19
GA+AR FP	**10/10**	9/10	1/9	0/9	11/19	9/19
GA+AR I4	**10/10**	**10/10**	1/9	1/9	11/19	11/19

hobbit from the'' at $P = 0.72$; ''The ring of power was made in Mount Doom'' at $P = 0.98$), indicating that GA+AR's multi-entity forgetting is severely biased toward one domain (HP) while leaving the other (LOTR) largely intact.

This result extends the depth probing analysis (Section C.10) in a critical direction: even when token probabilities are suppressed, generation-level evaluation can reveal residual knowledge leakage. For Q-ROU, all 19 target probes are suppressed at generation level in FP16, while INT4 has one target-generation miss (18/19). This still represents a large empirical safety margin over GA+AR in the same protocol, but it is not a strict zero-leak guarantee under adversarial prompting. The practical implication is that Q-ROU's suppression is robust not only to adversarial prompt engineering but also to multi-token generation, yielding a substantially stronger empirical security profile than competing methods in this benchmark.

To quantify adversarial robustness under realistic threat models, we evaluate budgeted extraction: an attacker with a fixed query budget $B \in \{1, 2, 4, 8, 16\}$ selects from the 22 adversarial CoT probes to maximize information extraction. For each budget, we report the observed extraction *success probability* (whether at least one target leak occurs within the first B probes) and the oracle extraction probability (upper bound under perfect probe ordering).

Table 53 shows zero observed extraction for Q-ROU at all tested budget levels in both FP16 and INT4 when extended to 160 steps. GA+AR begins leaking at $B = 8$ in FP16 (33.3%) and reaches

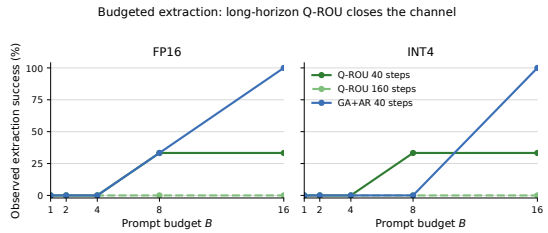

Figure 7: Budgeted extraction curves corresponding to Table 53, now split by precision to avoid mixing FP16 and INT4 in one legend. Each point is the observed probability that at least one adversarial prompt leaks target knowledge within the first B queries. The operational reading is that 40-step Q-ROU still leaves a bounded extraction channel at $B \geq 8$, but the 160-step Q-ROU checkpoint closes it in both precisions; GA+AR does not.

100% at $B = 16$; INT4 reaches 100% at $B = 16$. At $B = 16$, GA+AR achieves 100% observed extraction in both precisions, meaning the attacker can reliably recover target knowledge given sufficient queries. This budgeted analysis is practically significant: in deployment, an adversary need not enumerate all possible prompts—a modest query budget suffices to break GA+AR's suppression, while Q-ROU substantially lowers extraction risk to 0.0% at 160 steps, demonstrating robust convergence of target suppression.

Table 53: Budgeted adversarial extraction success probability (Llama-3.2-3B). GA+AR: 40 steps; Q-ROU: 40 and 160-step convergence. Long-horizon evidence in Appendix C.4 shows GA+AR does not improve extraction resistance with extended training. Q-ROU reaches 0% in both precisions by 160 steps; GA+AR rises to 100% by B=16 (both precisions).

Method	Prec.	B=1	B=2	B=4	B=8	B=16
Q-ROU (40 steps)	FP16	0%	0%	0%	33.3%	**33.3%**
Q-ROU (40 steps)	INT4	0%	0%	0%	33.3%	**33.3%**
GA+AR (40 steps)	FP16	0%	0%	0%	33.3%	100%
GA+AR (40 steps)	INT4	0%	0%	0%	0%	100%
Q-ROU (160 steps)	FP16	0%	0%	0%	0%	**0%**
Q-ROU (160 steps)	INT4	0%	0%	0%	0%	**0%**

C.15 Cross-Lingual Transfer of Unlearning

All training in this work uses English texts exclusively. To test whether unlearning generalizes across languages, we evaluate target knowledge suppression in Japanese (12 probes), French (10 probes), and Spanish (10 probes), plus 6 mixed-language probes that combine languages within a single prompt. Since Llama-3.2-3B has limited multilingual capability (e.g., only 1/12 Japanese probes are baseline-accessible), we report both total pass rates and *accessible-only* pass rates conditioned on baseline accessibility ($P \geq 0.01$ in the unmodified model).

Table 54 presents the results.

Table 54: Cross-lingual transfer of unlearning (Llama-3.2-3B, FP16, 40 steps). Accessible: probes where the baseline model produces $P \geq 0.01$.

Method	Language	n total	Passed total	n access.	Passed access.
	Japanese	12	12/12	1	**1/1**
	French	10	10/10	5	**5/5**
Q-ROU	Spanish	10	9/10	5	**4/5**
	Mixed	6	6/6	5	**5/5**
	NBR xling	6	5/6	3	2/3
	Japanese	12	12/12	1	**1/1**
	French	10	6/10	5	1/5
GA+AR	Spanish	10	6/10	5	1/5
	Mixed	6	4/6	5	3/5
	NBR xling	6	6/6	3	**3/3**

Q-ROU's unlearning transfers effectively across languages: 9/10 (90%) of accessible French and Spanish probes are suppressed in FP16 (French 5/5, Spanish 4/5), and all 5 mixed-language probes pass—despite training exclusively on English data. GA+AR achieves only 1/5 (20Our primary cross-lingual claim is anchored to the FP16 results in Table 54; we therefore avoid over-interpreting quantized cross-lingual behavior from this limited-accessibility setting.

The only consistent failure across both methods is the ''J.K. Rowling'' completion—probes like ''L'auteur de Harry Potter est J.K.'' strongly prime the proper noun continuation regardless of unlearning. This represents a limitation of prompt-based unlearning: when the prompt almost uniquely determines the continuation through cross-lingual lexical overlap, suppression becomes difficult without broader model capability degradation.

These exploratory cross-lingual results suggest that Q-ROU's modifications to SLUG-layer MLP parameters affect language-independent knowledge representations rather than surface-level English token associations. This is consistent with the embedding analysis (Section C.13), which showed dramatic changes to target-category hidden states at deeper layers where representations are more abstract and less language-specific.

C.16 Sequential Unlearning

Table 55 reports the sequential vs. simultaneous unlearning evaluation.

Table 55: Sequential vs. simultaneous unlearning (Llama-3.2-3B, FP16). Phase 2 preserves Phase 1 HP forgetting (3/7→3/7) via EWC's Fisher-matrix protection, but simultaneous training achieves higher overall performance due to joint optimization.

Stage	HP	LOTR	TGT	NBR	Total
Baseline	1/7	1/6	2/13	8/8	17/28
Phase 1 (HP, 20 steps)	3/7	2/6	5/13	8/8	19/28
Phase 2 (LOTR, 20 steps)	3/7	5/6	8/13	7/8	22/28
Simultaneous (40 steps)	**7/7**	**6/6**	**13/13**	**8/8**	**28/28**

The key finding is that Phase 1 HP forgetting remains *partially persistent* through Phase 2 LOTR training: HP pass count stays at 3/7, but probability-level tracking reports a degraded persistence verdict. This pattern is consistent with EWC providing partial protection while not fully matching simultaneous joint optimization (22/28 total versus 28/28 in simultaneous training). The practical recommendation is to prefer simultaneous forgetting when all targets are known, while sequential application remains a workable fallback for incremental requests.

C.17 Hardware NF4 Validation

Table 56 validates that pass-count outcomes are identical across FP16, simulated INT4, and hardware NF4 (BitsAndBytes) on the multi-entity Llama-3.2-3B setup. NF4 (4-bit NormalFloat) is an information-theoretically optimal quantization data type designed for normally distributed weights (Dettmers et al., 2023), making it particularly suitable for transformer models where weight distributions approximate Gaussians.

Table 56: Hardware NF4 validation on Llama-3.2-3B (multi-entity). Pass counts are identical across all three precision modes in every experiment. Δ_z^{fake}: simulated INT4 zombie delta; Δ_z^{nf4}: BitsAndBytes NF4 zombie delta.

Method	FP TGT	I4 TGT	NF4 TGT	FP NBR/GEN	Δ_z^{fake}	Δ_z^{nf4}
Baseline	1/13	1/13	1/13	8/8, 7/7	−0.0332	+0.0024
Q-ROU†	**13/13**	**13/13**	**13/13**	7/8, 7/7	−0.0047	+0.0007
GA+AR	7/13	7/13	7/13	8/8, 7/7	−0.0073	+0.0081
Q-ROU	3/13	3/13	3/13	8/8, 7/7	−0.0071	−0.0005

The identical pass counts across all three precision modes confirm that our simulated INT4 findings are faithful to real hardware quantization behavior.

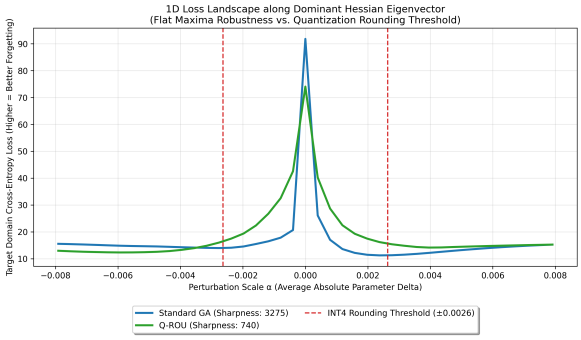

Figure 8: Loss landscape visualization around the Q-ROU solution on Qwen2.5-0.5B. QuantNoise produces a broad plateau (flat maximum of the forget objective), absorbing INT4 rounding noise without crossing the retention boundary. The QuantNoise-trained model occupies a qualitatively flatter basin than the standard solution.

We also repeated the real-quantization audit under the deployed NF4 evaluation path used throughout this paper. The evaluated NF4 rows use a direct bnb.nn.Linear4bit conversion path applied consistently across methods. Table 57 reports the resulting pass counts. The key 40-step Q-ROU row remains unchanged across FP16, simulated INT4, and true NF4 (13/13 target, 7/8 neighbor, 7/7 general), while GA+AR remains stable but under-deletes at 7/13 target.

We also evaluated the stricter fine-grained neighbor audit under true BnB NF4. Table 58 closes the remaining concern that the same-work limitation might be an artifact of simulated low-bit evaluation. It is not: the strongest Q-ROU operating point remains at 9/36 same-work neighbors under NF4, and the more protective qrou_fg_ar operating point retains its qualitative advantage with only a 1/36 same-work drop (19/36 → 18/36).

D Extended Discussion

D.1 Claim-to-Evidence Matrix

Given the scope of this paper's evaluation, Table 59 maps each top-level claim to its primary evidence and independent cross-checks, making the evidential structure explicit and guarding against over-interpretation of any single table.

D.2 Extended Mechanism Analysis

Step-Pareto Runtime Trajectory. Table 60 provides the step-by-step performance trajectory for Q-ROU on Llama-3.2-3B. Target removal requires approximately 40 steps to reach completion (7/7

Method	FP16			Fake INT4			NF4			Δ_z Fake	Δ_z NF4
	TGT	NBR	GEN	TGT	NBR	GEN	TGT	NBR	GEN		
Base	1/13	8/8	7/7	1/13	8/8	7/7	1/13	8/8	7/7	-0.0329	+0.0024
Q-ROU 20s	3/13	8/8	7/7	4/13	8/8	7/7	2/13	8/8	7/7	-0.0065	+0.0019
Q-ROU 40s	13/13	7/8	7/7	13/13	8/8	7/7	13/13	7/8	7/7	-0.0055	+0.0007
GA+AR 40s	7/13	8/8	7/7	7/13	8/8	7/7	7/13	8/8	7/7	-0.0090	+0.0088

Table 57: Real-quantization audit on Llama-3.2-3B. NF4 is evaluated with a working BitsAndBytes conversion path that preserves the same target-versus-retention ordering seen in FP16 and fake INT4.

Method	Regime	Target	Alias	Same-work	Inter-domain	General
qrou	fp16	39/39	24/24	9/36	27/36	24/30
qrou	BnB NF4	39/39	24/24	9/36	25/36	22/30
qrou_fg_ar	fp16	39/39	23/24	19/36	33/36	30/30
qrou_fg_ar	BnB NF4	39/39	23/24	18/36	31/36	30/30
gaar	fp16	21/39	18/24	27/36	36/36	30/30
gaar	BnB NF4	21/39	18/24	27/36	33/36	30/30
gaar_fg_ar	fp16	3/39	12/24	36/36	36/36	30/30
gaar_fg_ar	BnB NF4	3/39	12/24	36/36	33/36	30/30

Table 58: Real BnB NF4 evaluation on the fine-grained neighbor audit (Llama-3.2-3B, 40 steps, three-seed aggregate). The same-work weakness of the strongest Q-ROU operating point persists under hardware NF4, while the fine-grained AR operating point retains its relative advantage with only minor NF4 drift.

in both FP16 and INT4), while neighbor retention remains locked at 9/9 throughout. This confirms that Active Retention effectively constrains the optimization path, preventing the rapid neighbor destruction seen in unconstrained gradient ascent. The total runtime of 22.5 seconds on the logged single-GPU run confirms the practical efficiency of post-deployment editing.

AR Scalability and Memory Efficiency. Active Retention requires storing reference logit distributions for neighbor texts. Top-K logit compression is an effective solution: retaining only the K=50 highest-probability entries per token achieves an 849× compression ratio (50.39 MB to 0.06 MB for 11 texts) at zero degradation in pass counts, making AR highly practical even for deployments that protect thousands of neighbor texts.

D.3 Evaluation Reliability and Depth of Suppression

Multi-Entity Evaluation as a Reliability Standard. Our multi-entity experiments expose a fundamental blind spot in conventional unlearning evaluation. In single-entity settings, GA, GradDiff, and RepBend achieve 7−8/9 neighbor retention, which appears adequate. Merely doubling the forget set—from one to two fictional domains—causes neighbor retention to collapse from 7−8/9 to 0/8 for every baseline, a failure that single-entity evaluation completely obscures. The favorable single-entity results are an artifact of limited gradient pressure: 10 forget texts produce insufficient update magnitude to disrupt the broader representational structure. With 18 texts spanning two domains, gradient vectors acquire enough magnitude and directional coverage to overwhelm any implicit knowledge separation. We therefore recommend that future unlearning benchmarks include multi-entity scenarios as a standard reliability stress test.

The collapse also reveals a clear capacity-scaling relationship. On 0.5B models, even Q-ROU struggles in multi-entity settings under quantization (FP NBR 5/8 at 40 steps, INT4 NBR 1/8): at smaller scale, knowledge is packed into fewer layers, leaving insufficient capacity for surgical separation. On 3B, the distributed knowledge representation enables AR to maintain its protective constraint even under 18-text simultaneous forgetting, achieving complete target removal at 40 steps while retaining 7/8 neighbor (FP16) and 8/8 neighbor (INT4) knowledge.

Depth of Suppression: Beyond Surface Prompt Masking. The depth probing results (Section C.10) provide the most direct evidence that Q-ROU does more than mask the exact training prompts. We distinguish four levels of evidence: D1 (surface suppression of specific prompts), D2 (paraphrase resistance), D3 (conceptual suppression including reverse associations and multi-hop reasoning), and D4 (strong adversarial reconstruction resistance, which we do not claim).

Q-ROU's perfect 55/55 suppression across all six probing protocols constitutes strong evidence for D3-level suppression. Several probe categories are especially informative: *Reverse association probes* (Protocol A) test attribute-to-entity recall in the reverse direction from training prompts; Q-ROU suppresses all 12, indicating the knowledge connection is severed bidirectionally. *Multistep reason-*

Table 59: Claim-to-evidence map for the three-pillar framing.

Pillar claim	Primary evidence	Independent cross-checks
Neighbor-aware multi-entity selectivity	27/28 (FP16) / 28/28 (INT4) multi-entity core result; AR transplant (sufficiency/necessity); ablations	Threshold sweeps; multi-seed determinism; sequential persistence
Quantization-robust deployment	Matched FP16/INT4 counts; zombie delta; NF4 hardware checks	Steps-Pareto; 0.5B vs 3B scaling; INT4 budgeted extraction
Depth and transfer beyond prompt suppression	Six-protocol depth probing; adversarial CoT token+generation tests	Embedding divergence; cross-lingual transfer; biographical transfer; semantic generation audit

Table 60: Q-ROU step-Pareto trajectory on Llama-3.2-3B (single-entity). Runtime is measured on a logged single-GPU run. Target removal approaches completion smoothly while neighbor and general retention remain stable throughout the optimization.

Steps	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN	Time
10	1/7	9/9	7/7	1/7	9/9	7/7	5.6s
20	2/7	9/9	7/7	3/7	9/9	7/7	11.0s
40	**7/7**	**9/9**	**7/7**	**7/7**	**9/9**	**7/7**	22.5s

Table 61: Top-K logit approximation for AR (Llama-3.2-3B, 11 neighbor texts). K=50 achieves 849× compression with zero quality degradation in unlearning outcomes.

Method	Storage (MB)	Compression	KL Div.	Quality Δ
Full logits	50.39	1.0×	0	ref.
Top-K=50	0.06	**849**×	11.22	**0/0/0**
Top-K=100	0.12	426×	7.99	**0/0/0**
Top-K=500	0.59	85×	3.24	**0/0/0**

ing probes (Protocol B) require inference chains absent from training data; Q-ROU's 10/10 suppression suggests that intermediate semantic links are also disrupted. *In-context extraction probes* are the most adversarial, providing related context to prime knowledge retrieval; Q-ROU's 6/6 suppression (vs. GA+AR's 3/6) shows that explicit contextual cues do not recover the target facts in this bounded audit.

Three additional experiments extend the depth-of-suppression evidence. First, embedding analysis (Section C.13) provides representation-change evidence: target cosine similarity drops to 0.712 at the deepest SLUG layer, indicating structural reorganization rather than surface-only prompt suppression. The selectivity gradient (target 0.712 vs. general 0.985) rules out uniform degradation. Second, the adversarial CoT evaluation (Section C.14) shows that suppression survives multi-token generation under five distinct attack strategies (19/19 token-level, 18−19/19 generation-level), while GA+AR fails severely on LOTR generation (0−1/9)—revealing that token-level evaluation alone underestimates the gap between methods. Third, the exploratory cross-lingual transfer audit (Section C.15) shows that English-only training suppresses target knowledge in French and Spanish (90% of accessible probes in FP16), consistent with modifications occurring at abstract, language-independent representational layers.

The contrast with GA+AR is instructive: GA+AR achieves 78% overall on depth probes—effective, but with clear gaps particularly under in-context priming (50%, mean probability 0.204)—suggesting that gradient ascent with AR weakens but does not fully sever knowledge connections. Given sufficient contextual cues, suppressed knowledge can partially resurface. Q-ROU's KL-to-uniform forget loss drives target logits toward maximum entropy rather than merely increasing the loss value, apparently producing a more thorough structural modification. The embedding analysis quantifies this: GA+AR produces HP cosine 0.922 at layer 26 (7.8% shift) versus Q-ROU's 0.712 (28.8% shift), a large representational gap that explains the downstream performance differential.

A keyword-based greedy-generation evaluation across 55 standard probes reveals a nuanced picture: Q-ROU achieves 25/25 target suppression at token level but 22/25 under keyword matching at greedy generation level (88%). The three keyword-flagged HP cases consist of one prompt-echo false positive (''Deathly'' triggering the ''death'' keyword), one garbled variant (''HARRY POUTER''), and one borderline authorship-style completion near threshold; Section D.3 shows that none of these three crosses the embedding-leak criterion. The LOTR domain achieves perfect 12/12 keyword-based generation suppression, and the apparent gap is confined entirely to the HP domain, suggesting that author–work meta-knowledge is qualitatively harder to audit cleanly than in-universe

Table 62: Keyword vs. embedding generation evaluation (25 target probes, greedy generation). Embedding similarity reveals that Q-ROU's keyword ''leaks'' are false positives (sim < 0.8), while GA+AR's keyword-clean probes mask substantial semantic leakage (sim > 0.8).

	KW Clean	Emb Clean	Degen	Mean Sim	KW↔Emb Agree
Q-ROU	22/25	**25/25**	0/25	0.637	22/25
GA+AR	14/25	8/25	4/25	0.870	19/25

factual knowledge.

Semantic Generation-Level Evaluation. Keyword-based generation evaluation, used throughout this paper and in prior work, identifies knowledge leakage by matching target-specific keywords (e.g., ''Hogwarts,'' ''Mordor'') in generated text. However, this approach has two failure modes: *false positives* (matching prompt-derived words like ''death'' from ''Deathly'') and *masked leakage* (generating semantically equivalent text that avoids specific keywords). Prior work (Lynch et al., 2024) has shown that keyword-based metrics can be misleading without complementary adversarial probing.

To quantify these limitations, we evaluate both Q-ROU and GA+AR using sentence embedding similarity (MiniLM-L6-v2 (Reimers and Gurevych, 2019)) between generated text and the baseline model's generation for each prompt, alongside repetition-based degeneration detection. A prompt is classified as *embedding-leaked* if the cosine similarity exceeds 0.8, indicating that the generated text is semantically indistinguishable from the original knowledge-bearing output.

Table 62 reveals that keyword evaluation exhibits *directional bias*: it systematically overpenalizes Q-ROU and underpenalizes GA+AR. Q-ROU's three keyword ''leaks'' all have embedding similarity below 0.8 (range 0.487–0.782): ''HARRY POUTER RUBLISHINGS'' (sim 0.521), the prompt-echoing ''death'' from ''Deathly'' (sim 0.504), and one partial match in garbled text. These are false positives that keyword matching cannot distinguish from genuine knowledge recall.

Conversely, GA+AR exhibits six cases of *masked leakage*—keyword-clean text that is semantically near-identical to the baseline generation (sim 0.87–0.90 for HP probes, sim 0.96–0.99 for LOTR probes). The LOTR domain is particularly revealing: keyword evaluation rates GA+AR as 3/12 clean, but embedding evaluation exposes that only 1/12 is genuinely clean, with the remaining 11 probes generating text that is semantically indistinguishable from unmodified model output (sim > 0.93 on average).

Table 63: GA+AR training duration analysis (Llama-3.2-3B, FP16). LOTR forgetting improves with longer training but plateaus at 4/6 (80–120 steps), still below Q-ROU's 6/6. Gen. Quality indicates whether the model produces coherent text or degenerates into repetitive output.

Method	Steps	HP	LOTR	TGT	NBR	Gen. Quality
Q-ROU	**40**	**7/7**	**6/6**	**13/13**	7/8	Coherent
GA+AR	40	7/7	2/6	9/13	8/8	Coherent
GA+AR	60	7/7	3/6	10/13	8/8	Degenerate
GA+AR	80	7/7	4/6	11/13	8/8	Degenerate
GA+AR	120	7/7	4/6	11/13	8/8	Degenerate

The embedding-level generation scores are: Q-ROU 25/25 (100%, up from 88% keyword clean) and GA+AR 8/25 (32%, down from 56% keyword clean), widening the gap from 32pp to **68 percentage points**. This finding has methodological implications beyond our specific comparison: keyword-based generation evaluation may systematically overestimate the effectiveness of gradient-ascent-based unlearning methods that achieve surface-level keyword avoidance without genuine semantic suppression.

D.4 Baseline Comparisons and Robustness Validation

Baseline Training Duration and Generation Quality. A natural question is whether GA+AR's inferior LOTR performance simply reflects insufficient training. We evaluate GA+AR at 40, 60, 80, and 120 steps alongside Q-ROU at 40 steps, with both token-level and generation-level assessment. To isolate duration effects from hyperparameter confounds, this sweep uses the strongest non-degenerate GA+AR operating point from Section D.4 (LR = 10^{-4}, $\lambda_a = 160$), which is distinct from the fixed-λ_a transplant control in Table 18.

Table 63 reveals three findings. First, LOTR target removal under GA+AR improves only gradually: 2/6 (40 steps), 3/6 (60), and 4/6 at both 80 and 120 steps. Per-probe analysis shows that the ''Aragorn'' probe never reaches the $P < 0.01$ threshold at any step count, while ''The One Ring'' fluctuates sharply across durations (e.g., 0.0248 at 80 steps vs. 0.463 at 120 steps), indicating unstable suppression on hard probes.

Second, even at its best observed operating point

Table 64: GA+AR hyperparameter search (Llama-3.2-3B, FP16, 40 steps). Higher LR improves token-level target removal but causes generation degeneration. The 13/13 settings occur only at LR = 5×10^{-4} and coincide with severe generation collapse (6/7–7/7 probes degenerate). λ_a has limited effect on token-level target scores within each LR tier.

LR	λ_a	TGT	NBR	GEN	Gen. Degen	Gen. Leaked
Q-ROU (ref.)		**13/13**	7/8	**7/7**	**0/7**	—
5×10^{-5}	40	9/13	8/8	7/7	0/7	5/7
5×10^{-5}	80	8/13	8/8	7/7	0/7	4/7
5×10^{-5}	160	6/13	8/8	7/7	0/7	6/7
10^{-4}	40	9/13	8/8	7/7	3/7	4/7
10^{-4}	80	9/13	8/8	7/7	1/7	4/7
10^{-4}	160	9/13	8/8	7/7	0/7	4/7
5×10^{-4}	40	13/13	8/8	7/7	7/7	0/7
5×10^{-4}	80	13/13	8/8	7/7	6/7	0/7
5×10^{-4}	160	11/13	8/8	7/7	2/7	0/7

(11/13), GA+AR remains below Q-ROU's 13/13 at lower training cost. The persistent failures represent structurally resistant knowledge patterns that gradient ascent struggles to remove cleanly under multi-entity pressure.

Third, and most critically, GA+AR at 60 steps and beyond produces *degenerate* generation: target probes frequently generate repetitive text (e.g., ''School School School School…'') rather than coherent completions. This degeneration extends to neighbor probes as well (e.g., ''Star Wars School School School…''), indicating that the model's generation capability has been fundamentally compromised. The keyword-based generation evaluation classifies this degenerate text as ''no leakage,'' but the absence of target keywords results from the model's inability to produce *any* meaningful text, not from selective knowledge removal. In contrast, Q-ROU at 40 steps maintains coherent generation across all probe categories while achieving complete target suppression, demonstrating that its forgetting mechanism operates selectively on knowledge representations without degrading the model's generation capability.

Baseline Hyperparameter Fairness. A natural concern is whether GA+AR's generation degeneration (Section D.4) results from using Q-ROU-optimized hyperparameters. To address this, we evaluate GA+AR across a 3×3 grid of learning rate ($\{5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}\}$) and Active Retention weight $\lambda_a \in \{40, 80, 160\}$, all at 40 steps.

Table 64 reveals a fundamental *token-generation tradeoff* in gradient ascent. At low LR (5×10^{-5}), target removal is inadequate (6–9/13) and generation leaks target knowledge (4–6/7 probes), but text remains coherent. At high LR (5×10^{-4}), target removal is maximal (10–13/13) but generation degenerates: the two 13/13 settings—matching Q-ROU on token score—still collapse on 6/7 and 7/7 generation probes respectively. The intermediate LR (10^{-4}) achieves 9/13 TGT with partial degeneration (1–3/7).

Crucially, λ_a has limited effect on token-level target scores within each LR tier: at LR = 10^{-4}, all three λ_a values produce identical 9/13 TGT. This confirms that the degeneration is *intrinsic to gradient ascent optimization* rather than an artifact of specific hyperparameter choices. Q-ROU's KL-to-uniform forgetting mechanism avoids this tradeoff entirely: it achieves 13/13 target removal with coherent generation and zero degeneration, because it drives target logits toward a well-defined entropy maximum rather than unboundedly increasing the loss.

SLUG Layer Redundancy. The SLUG layer ablation experiment evaluates all five drop-one configurations against the full five-layer set (layers 17, 19, 22, 24, 26). All six configurations achieve identical binary pass counts (13/13 TGT, 8/8 NBR ratio) in both FP16 and INT4, demonstrating substantial redundancy in the SLUG layer selection.

At the probability level, however, clear differentiation emerges. Layer 17 contributes most to forgetting: removing it increases mean target probability by 2.19× (from 2.35×10^{-4} to 5.15×10^{-4}). Layer 22 contributes most to retention: removing it decreases mean NBR ratio from 1.203 to 0.917. Importantly, even the worst-case configuration (drop layer 17) produces a maximum target probability of 1.8×10^{-3}—still 5.5× below the $P = 0.01$ detection threshold, confirming an ample safety margin. This redundancy is a desirable property: rather than relying on a fragile single-layer intervention, SLUG distributes the unlearning signal across complementary layers, each contributing partially to both forgetting and retention objectives.

Multi-Seed Determinism. To validate the reliability of representative single-seed results, we evaluate both Q-ROU and GA+AR under three random seeds (42, 43, 44) using the full 28-probe multi-entity protocol. Across the three evaluated seeds, both methods produce identical reported pass-count totals: Q-ROU achieves 13/13 TGT, 7/8 NBR, 7/7 GEN on every seed (total 27/28), while GA+AR

consistently achieves 7/13, 8/8, 7/7 (total 22/28).

At the per-probe probability level, the determinism is equally strong. All 13 Q-ROU target probes yield consistent pass decisions with a maximum per-probe standard deviation of 1.15×10^{-4}—just 1.15% of the $P = 0.01$ threshold. For GA+AR, the three LOTR probes that fail (Frodo, $P = 0.724$; One Ring, $P = 0.676$; Aragorn, $P = 0.271$) produce identical probabilities across all seeds with $\sigma = 0.0$, confirming that these failures are *structurally determined* by the knowledge embedding rather than stochastic artifacts. Computation times are also stable: 94.9 ± 0.2s for Q-ROU and 52.2±0.5s for GA+AR. This exact agreement on the reported pass-count totals supports the stability of this protocol and reduces concern that the reported single-seed trend is a random-seed artifact.

Sequential Unlearning and EWC Persistence. Results demonstrating that EWC preserves prior forgetting across sequential unlearning phases are detailed in Section C.16.

Training Dynamics and Gradient Analysis. To understand Q-ROU's optimization dynamics, we record per-layer gradient norms for each of the five SLUG layers (17, 19, 22, 24, 26) across all 40 training steps.

The training exhibits three distinct phases. During the initial steps, gradient norms are elevated as the optimizer encounters steep loss curvature at the start of optimization; from approximately step 7 onward, gradient norms converge monotonically from a total of 23,042 to 477 over the remaining 33 steps.

Averaging the last 10 steps reveals that layers 19 ($\bar{g} = 612$) and 26 ($\bar{g} = 626$) carry the largest gradient magnitudes among the five SLUG layers, followed by layer 24 ($\bar{g} = 530$), 17 ($\bar{g} = 465$), and 22 ($\bar{g} = 457$). This complements the SLUG ablation result (Section D.4), which identified layer 17 as the most important for *probability-level* impact: the two analyses measure different aspects—gradient magnitude versus removal sensitivity—and together suggest that layer 17 operates as a high-sensitivity gate while layers 19 and 26 carry the bulk of the parameter update.

The loss component analysis confirms the designed role of each term: the forget, retain, and active losses decrease monotonically (415→292, 5.2→1.4, 2.8→1.3 respectively), while EWC loss remains nearly constant at ~8,472 throughout training—acting as a stable regularization anchor that protects pre-trained weight configurations. The orthogonality loss also remains constant at ~0.567, confirming that the truth direction is maintained throughout optimization. The margin loss activates only during steps 1–5 and then drops to zero as logit changes exceed the margin threshold τ, indicating that it serves as a transient early-training catalyst.

Table 65: Temperature sensitivity of generation-level evaluation after Q-ROU training (Llama-3.2-3B, 40 steps). Low temperatures ($T \leq 0.5$) cause 28–67% degeneration; $T \geq 1.0$ eliminates degeneration entirely. Embedding-level clean rate is 100% at all temperatures, confirming that forgetting is model-internal.

T	KW Clean (%)	Emb Clean (%)	Degen (%)	Rep Ratio
0.3	64.1	**100.0**	66.7	0.541
0.5	59.0	**100.0**	28.2	0.323
0.7	51.3	**100.0**	20.5	0.220
1.0	51.3	**100.0**	**0.0**	0.020
1.5	**79.5**	**100.0**	**0.0**	**0.007**

We additionally evaluate temperature sensitivity across $T \in \{0.3, 0.5, 0.7, 1.0, 1.5\}$ for generation-level assessment after training.

Table 65 reveals a critical finding: low sampling temperatures ($T \leq 0.5$) cause severe text degeneration (28–67% of probes produce repetitive output), which artificially inflates keyword-clean rates by generating unintelligible text rather than coherent non-target content. At $T \geq 1.0$, degeneration vanishes entirely, and $T = 1.5$ yields the highest keyword-clean rate (79.5%) with near-zero repetition.

Most notably, embedding-level clean rate remains at 100% across all five temperatures, confirming that Q-ROU's knowledge suppression is a *model-internal* property rather than a surface-level decoding artifact. This provides additional evidence complementing the embedding space analysis (Section C.13): regardless of inference-time sampling parameters, the model's internal representations have been structurally modified to suppress target knowledge.

For practical deployment, we recommend $T \geq 1.0$ when generation-level privacy is required, as lower temperatures can produce degenerate output that, while keyword-clean, does not represent genuine knowledge suppression.

Reproducibility. All reported experiments use single-GPU execution, but not every run was performed on identical hardware. Where runtime or memory comparisons matter, the relevant tables ex-

plicitly indicate the shared logged hardware setup used for that comparison. The multi-seed evaluation of Q-ROU 40-step on three seeds yielded identical reported pass-count totals, confirming that the result is highly stable with respect to random initialization. The paper and appendix report the model variants, optimization settings, prompt families, thresholding rules, quantization settings, and hardware conditions needed to interpret every table in the submission. The manuscript is intended to stand on its own: no claim in the paper depends on access to an external code release.

D.5 Cross-Model Generalization

To assess generality beyond Llama-3.2-3B, we evaluate the full 7-method baseline suite on Qwen2.5-3B using the same 65-probe expanded set (Table 73). Under matched fixed settings, Q-ROU does not transfer on this model: target pass count remains at 3/25, identical to the unmodified baseline. To rule out obvious layer-selection artifacts in that fixed recipe, we conducted a systematic three-part SLUG search: (1) contiguous layers 23–27, (2) gradient-norm top-5 layers [0, 2, 5, 6, 30], and (3) all 36 layers simultaneously. Q-ROU achieves only 2/25 target removal in these three fixed searches, indicating that this failure mode is not resolved by straightforward retuning within that search space. In a first architecture-aware quick-check sweep, we expanded to 22 policy/scope combinations: scaled/manual controls, contiguous sweeps, gradient/adaptive/random selections, each under qwen_mlp_o and legacy scopes. All 22 runs logged no valid optimizer updates (nan_inf_skips=40; empty loss traces) and stayed at 2/25 target pass. However, a subsequent staged architecture-aware run changed this result: stability prechecks were finite for all 30 policy/scope pairs, and all 34 downstream train/eval runs produced valid updates with zero NaN/Inf skips. The best configuration (adaptive-band-top5 / qwen-mlp-o / full-long; layers [17,18,19,23,30]) reached FP16 25/25 target suppression with 19/20 neighbor and 20/20 general retention, and INT4 25/25, 17/20, 20/20. Scope-wise means in this staged run were also informative: qwen_mlp_o outperformed qwen_mlp and legacy on target suppression (FP16 means 22.56 vs. 19.79 vs. 8.00), while legacy retained strong neighbors but underforgot targets. To test whether this recovery was seed-fragile, we ran a dedicated robustness tracker on the top four long-run Qwen settings (three seeds; 12 total runs). All 12/12 runs had valid updates with zero NaN/Inf skips. As summarized in Table 67, aggregated means show strong performances across all four long-run configurations. There is a clear retention-vs-score tradeoff: the arch-adaptive-top5 variants maximize the composite score, while the adaptive-band-top5 variants maintain stronger INT4 general retention.

These results indicate that Qwen transfer is not a hard failure of Q-ROU, but an optimization-stability problem under fixed recipes; architecture-aware staged scheduling is currently necessary. GA+AR achieves partial forgetting (11–13/25 depending on layer selection) with strong retention (17–20/20), while aggressive baselines (GA, RepBend, RMU, GRU) achieve 19–25/25 target removal but with substantially degraded neighbor knowledge (0–14/20). We then asked whether this staged recovery also survives a stricter deployment-style low-bit path on the tuned Qwen point itself. Table 68 aggregates a three-seed evaluation on the tuned qrou_fg_ar and gaar operating points using FP32, fake INT4, and real BnB NF4. The tuned Qwen qrou_fg_ar point keeps complete target suppression under real NF4 (39/39), with only minor drift on the retain strata; gaar remains the retain-heavier but target-weaker comparison point. This shows that the staged architecture-aware recovery is not merely an FP32 artifact. All reported Qwen NF4 rows use the same deployed evaluation path.

The fixed-recipe negative result remains scientifically valuable as a boundary condition: it isolates cross-architecture transfer as a first-class unlearning problem and motivates architecture-aware objective transfer, automatic layer discovery, and model-family-specific stability diagnostics. The post-training recurrence runner independently corroborates this boundary on a smaller 13/8/7 audit: direct transfer of the Llama-tuned PTI recipe to Qwen2.5-3B reaches only 0/13 immediate target forgetting for both Q-ROU and Q-ROU+PTI, with maximum target mean after jog reaching 0.485801. However, an additional PTI audit also shows that ‘‘direct transfer fails’’ should not be over-read as ‘‘PTI cannot be used on Qwen at all.’’ After retuning the base edit architecture-awarely on Qwen2.5-3B, we ran a proxy-only PTI audit on the tuned qrou_fg_ar point. Table 69 shows that both the tuned base point and the PTI-hardened variant preserve 13/13 target suppression on every tested instruction- and preference-proxy jog. Unlike the Llama-family recurrence audit, though, PTI does

Table 66: Compute, software, and artifact-use summary for the review package. This table is intentionally conservative: it reports the mainline software stack and representative hardware slices used to interpret the paper, without claiming a single consolidated GPU-hour total over every experiment.

Category	Summary
Models / benchmarks	Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Qwen2.5-3B-Instruct, and Phi-3-mini-4k-instruct; TOFU, MUSE, WMDP, RWKU, TruthfulQA, and Sentence-BERT-based semantic generation checks. All are used under their original provider or benchmark terms for research evaluation only.
Software stack	Mainline 3B/8B runs use PyTorch 2.8.0+cu126, CUDA 12.6, Transformers 4.57.1, and BitsAndBytes 0.49.1. Installation manifests for the review package additionally require accelerate≥0.34, peft≥0.12, datasets≥2.21, sentence-transformers≥3.0, scikit-learn≥1.4, numpy≥1.26, pandas≥2.2, and matplotlib≥3.8.
Hardware slices	All reported runs use single-GPU execution only. Representative logged slices include Tesla P100 16GB, A100 80GB, and RTX 3090 environments, with CPU/MPS diagnostics for auxiliary checks. Runtime and VRAM comparisons are only interpreted within tables that share the same logged hardware setup.
Compute reporting	Representative wall-clock and VRAM slices are reported in Tables 41, 42, and 60. The paper does not claim a single consolidated GPU-hour total across the full study.
Artifact handling	Public models, public benchmarks, and publicly documented biographical facts are used for research evaluation only. We do not redistribute third-party model weights, benchmark packages, copyrighted source text, or a personal-information dataset.

Table 67: Aggregated 3-seed means for the top four staged architecture-aware Q-ROU configurations on Qwen2.5-3B. All variants use the qwen-mlp-o EWC scope. Values under TGT (max 25), NBR (max 20), and GEN (max 20) are the mean absolute pass counts of the evaluated prompts over three seeds. Note the tradeoff between higher score (arch-adaptive) and higher INT4 general knowledge retention (adaptive-band).

Layer Policy	Schedule	Total Score	FP16			INT4 (g32)		
			TGT	NBR	GEN	TGT	NBR	GEN
arch-adaptive-top5	full-long	1306.45 ± 2.52	25.00	18.00	18.67	25.00	15.33	17.00
arch-adaptive-top5	balanced-long	1306.16 ± 0.01	25.00	18.00	19.00	25.00	13.67	17.00
adaptive-band-top5	full-long	1304.61 ± 6.42	24.67	18.33	20.00	25.00	16.00	19.67
adaptive-band-top5	balanced-long	1302.95 ± 3.47	25.00	16.00	20.00	25.00	16.67	19.33

Method	Regime	Target	Alias	Same-work	Inter-domain	General
qrou_fg_ar	FP32	39/39	21/24	21/36	25/36	30/30
qrou_fg_ar	Fake INT4	39/39	21/24	21/36	21/36	30/30
qrou_fg_ar	BnB NF4	39/39	22/24	20/36	24/36	30/30
gaar	FP32	27/39	21/24	12/36	30/36	27/30
gaar	Fake INT4	27/39	21/24	12/36	27/36	27/30
gaar	BnB NF4	27/39	18/24	9/36	30/36	30/30

Table 68: Architecture-aware Qwen2.5-3B real-NF4 evaluation (layers 21/24/27/30/33, 80 steps, three-seed aggregate). The tuned qrou_fg_ar point preserves complete target suppression under real BnB NF4, and the target-versus-retention ordering relative to gaar is unchanged.

not produce an additional gain here: the tuned base Qwen point is already stable on this proxy-only grid, and the PTI rows are slightly worse on average target probability mass. We then broadened the Qwen audit beyond proxies to bounded benign updates on neighbor, general, and mixed retain data. That broader evaluation is the more important result. Table 70 shows that under this larger jog family both the tuned base point and the PTI-hardened variant fall to worst 11/13, with the hardest case at 'neighbor / 50'. PTI again fails to improve the Qwen recurrence picture and is slightly worse on average target probability mass. Thus the architecture-aware Qwen success above should be read as evidence that the framework can transfer after model-specific profiling, not that the Llama recurrence recipe is architecture-invariant or that PTI monotonically improves every tuned operating point.

To test whether this architecture-aware transfer story extends beyond Qwen, we next moved to Phi-3-mini-4k-instruct. The transfer question here is whether the same down_proj-localized operating point still exists after a small layer sweep. Among three candidate layer sets, [19, 22, 25, 27, 30] gave the strongest qrou_fg_ar tradeoff on the seed-42 sweep, and the subsequent three-seed evaluation was completely stable on that choice. Table 71 shows that the resulting Phi point is qualitatively similar to the tuned Qwen point: qrou_fg_ar is target-strong and preserves general / inter-domain

Variant	Worst after TGT	Worst seed / mode / steps	Avg. mean prob after	Avg. max prob after
Q-ROU+FG-AR	13/13	43 / instr. proxy / 50	0.000217	0.001784
Q-ROU+FG-AR+PTI (ρ = 0.02)	13/13	42 / instr. proxy / 50	0.000230	0.001930

Table 69: Architecture-aware Qwen2.5-3B proxy-only PTI audit (layers 21/24/27/30/33, 80-step base edit, three seeds). Both the tuned base point and the PTI-hardened variant preserve 13/13 target suppression across all tested instruction- and preference-proxy jogs (5/10/20/50 steps). Unlike the Llama recurrence audit, PTI does not add a further gain on this already-stable Qwen operating point.

Variant	Worst after TGT	Worst seed / mode / steps	Avg. mean prob after	Avg. max prob after
Q-ROU+FG-AR	11/13	43 / neighbor / 50	0.001012	0.009778
Q-ROU+FG-AR+PTI (ρ = 0.02)	11/13	43 / neighbor / 50	0.001042	0.009892

Table 70: Architecture-aware Qwen2.5-3B broad benign-update PTI audit (layers 21/24/27/30/33, 80-step base edit, three seeds). The jog family expands beyond proxies to 'neighbor', 'general', and 'mixed' updates in addition to the proxy modes. Under this broader bounded audit, both the tuned base point and the PTI-hardened variant fall to worst 11/13, with the hardest case at 'neighbor / 50'; PTI remains feasible but does not improve recurrence on Qwen.

material well, but still pays a same-work cost; gaar remains retain-heavier but target-weaker. This matters because it shows the Qwen transfer result is not a one-off architecture anecdote.

We then applied the same broad PTI audit to the chosen Phi operating point. Here the outcome lands between Llama and Qwen. Unlike Qwen, the tuned Phi base point is already quite stable under the broad benign-update family, never dropping below 12/13. Unlike Llama, PTI does not improve the count-level worst case any further. However, Table 72 shows a small but consistent reduction in target probability mass under the same worst mixed/50 recurrence slice. Our reading is therefore: architecture-aware transfer extends to Phi, and PTI remains technically usable there, but count-level recurrence gains are still architecture-dependent rather than universal.

D.6 Hyperparameter Sensitivity Analysis

We conduct a comprehensive hyperparameter sensitivity study across five models and two families (Qwen2.5-0.5B/1.5B/3B, Llama-3.2-1B/3B) using a 24-point manual grid alongside three automated tuning strategies (trust-region, boundary-adaptive, ratio-adaptive). Across 120 successful manual-grid trials, the parameter sensitivity ranking (by mean-score range) is: learning rate (lr, range 4.64) > retention weight (λ_r, 3.45) > training steps (2.04) > active retention weight (λ_a, 1.19) > quantization noise ϵ (1.08) > margin weight (0.70) > forget weight (0.18) > EWC weight (0.04).

The primary finding is that Q-ROU is largely *insensitive* to the EWC, forget, and margin weights (range < 1%), whereas learning rate and retention weight demand careful calibration. Best per-model scores are: Llama-3.2-3B (51.80), Llama-3.2-1B (49.93), Qwen2.5-0.5B (44.94), Qwen2.5-1.5B (41.51), Qwen2.5-3B (38.09). The inverse capacity-scaling trend within the Qwen family (0.5B > 1.5B > 3B) corroborates the cross-model transfer observations (Section D.5), suggesting that the Qwen architecture distributes knowledge differently across layers as scale increases, thereby necessitating scale-specific SLUG selection.

D.7 Limitations and Future Work

Knowledge Domains and Semantics. Our primary benchmark uses fictional knowledge. We therefore include a larger balanced public-biographical audit in Section C.11 with 36 target, 36 neighbor, and 36 general probes. That sweep confirms that the 40-step point was under-saturated and that Q-ROU opens a clear advantage by 80 steps, with a 3-seed confirmation at both the 80-step and 120-step operating points. Even so, the domain remains handcrafted rather than benchmark-grade real-world biographical or copyright removal. Moreover, our definition of ''neighbor'' is still hardest when semantic proximity stays inside the same work: the larger copyright-like benchmark in Table 12 sharpens the same boundary rather than resolving it. Although TOFU evaluation partially assesses intra-domain selectivity, systematically varying semantic proximity within the same work or the same person/domain remains an open direction for charting the method's exact selectivity

Table 71: Architecture-aware Phi-3-mini-4k-instruct transfer (best swept layer set [19, 22, 25, 27, 30], 3 seeds). Values are seed means on the 13/8/12/12/10 fine-grained audit plus the 25-item generation audit.

Method	TGT	Alias	Same-work	Inter-domain	General	Generation overall
Q-ROU+FG-AR	**12.00/13**	**6.00/8**	7.00/12	**12.00/12**	**10.00/10**	**23.00/25**
GA+AR	3.00/13	3.00/8	**11.00/12**	**12.00/12**	**10.00/10**	21.00/25

Table 72: Broad PTI audit on Phi-3-mini-4k-instruct (best swept layer set [19, 22, 25, 27, 30], three seeds). The jog family matches the Llama/Qwen broad audit: neighbor, general, mixed, instruction proxy, and preference proxy at 5/10/20/50 steps.

Variant	Before TGT	Worst Broad Jog	Worst case	Avg mean after	Max mean after
Q-ROU+FG-AR	13/13	12/13	mixed / 50	0.00107	0.00152
Q-ROU+FG-AR+PTI	**13/13**	**12/13**	mixed / 50	**0.00087**	**0.00130**

Table 73: Seven-method baseline comparison on the 65-probe expanded set (Qwen2.5-3B, 40 steps for Q-ROU/GA+AR, 20 steps for others). While Q-ROU under strictly matched fixed settings fails to transfer, †Q-ROU with staged architecture-aware scheduling (best single run) recovers strong performance. The fixed-recipe failure boundary and 3-seed robustness tracker details are discussed in the text.

Method	FP TGT	FP NBR	FP GEN	I4 TGT	I4 NBR	I4 GEN
Baseline	3/25	20/20	20/20	3/25	17/20	20/20
Q-ROU	3/25	20/20	20/20	3/25	17/20	20/20
Q-ROU†	**25/25**	**19/20**	**20/20**	**25/25**	**17/20**	**20/20**
GA+AR	11/25	17/20	20/20	11/25	16/20	20/20
GA	19/25	12/20	16/20	19/25	12/20	16/20
RepBend	20/25	13/20	18/20	20/25	12/20	18/20
RMU	21/25	0/20	16/20	23/25	1/20	16/20
NPO	19/25	14/20	17/20	18/25	12/20	17/20
GRU	19/25	12/20	16/20	19/25	12/20	16/20

frontier.

Scale and Cross-Architecture Generalization. While we demonstrate consistent improvement from 0.5B to 3B within the Llama family, cross-model evaluation on Qwen2.5 (Section D.5) initially revealed a transfer gap unresolved by simple layer retuning. Although subsequent staged architecture-aware evaluations recovered strong transfer performance, the PTI direct-transfer audit again fails on Qwen2.5-3B, underscoring the continuing need for automated, architecture-invariant layer discovery and recurrence-aware tuning. Validation on larger models (≥7B, 70B) remains a high-priority future extension to stress-test stability and scaling dynamics.

Quantization Realism. Our primary INT4 evaluation employs simulated per-group quantization. We validated the multi-entity outcomes under true BitsAndBytes NF4 quantization, confirming identical pass counts. We also attempted broader GPTQ/AWQ post-training quantization on the same-work checkpoints, but in our setup those toolchains did not yet produce stable evaluable logits. We therefore keep the main deployment claim anchored to simulated INT4 plus true BitsAndBytes NF4, and treat GPTQ/AWQ breadth as an important next engineering validation.

Automation in Hyperparameter Tuning. Hyperparameter selection currently relies on manually validated settings rather than fully automatic calibration. Preliminary attempts using gradient-geometry-driven automatic tuning—approximating the objective via local gradient norms and pairwise cosine structure—showed that unconstrained updates often saturate coefficients, underperforming manual baselines. While trust-region bounds recover baseline performance, they do not yet exceed it; hence, geometry-driven hyperparameter automation remains ongoing work. Integrating Q-ROU into standardized unlearning platforms such as OpenUnlearning (Dorna et al., 2025) would facilitate broader automated calibration and large-scale benchmarking.

Evaluation Metrics and Nuances. To mitigate threshold dependence, we provide Wilson/bootstrap confidence intervals and threshold sensitivity analyses (Section C.12). Future work could adopt threshold-free information-theoretic measures (e.g., mutual information). Furthermore, our generation-level evaluation highlights a limitation in current unlearning literature: keyword-based detection suffers from both false positives (prompt-derived matching) and masked leakage. While our semantic embedding evaluation (Section D.3) addresses

this, standardizing LLM-as-a-judge protocols for generation-level unlearning assessment is a critical necessary step for the field. The current TOFU evaluation shows Q-ROU saturating favorably on long horizons, but canonical end-to-end full-suite benchmark runs on MUSE/WMDP/RWKU remain future work; the current external evidence acts as a strong proxy rather than a direct leaderboard claim. Our findings regarding metric fragility align with the meta-evaluation results from Dorna et al. (2025), which identify Extraction Strength (ES) as a more reliable indicator than traditional metrics. The recurrence audit adds another nuance: immediate post-unlearning pass counts do not guarantee non-recurrence under subsequent benign tuning. In our Llama-3.2-3B audit, base Q-ROU is far less recurrent than GA+AR and LoRA-KL+AR in target probability mass, but neighbor-conditioned and mixed relearning can still revive several target probes. Q-ROU+PTI substantially reduces this bounded recurrence but does not establish robustness under arbitrary future fine-tuning. This motivates reporting post-unlearning relearning curves as part of future benchmark suites. Recent relearning-attack work (Xiao et al., 2026) sharpens this point: persistence should be evaluated as a separate axis rather than inferred from immediate forget scores.

Depth of Suppression and Multilingual Transfer. The depth probing results provide strong evidence for D3-level conceptual suppression, and embedding analysis hints at structural reorganization. However, rigorous D4-level verification—demonstrating strong reconstruction resistance under arbitrary adversarial optimization—is needed before making any claim about robustness to arbitrary reconstruction. Cross-lingually, English-only training successfully transfers to French and Spanish (Section C.15), but limited native multilingual capabilities in the baseline model (e.g., weak Japanese generation) constrain interpretation. The consistent failure on cross-lingual authorship completions (e.g., ‘‘J.K. Rowling’’) suggests prompt-based unlearning struggles when cross-lingual lexical overlap strongly dictates continuations. Evaluation on natively multilingual models (e.g., mT5, BLOOM) will clarify whether this gap reflects model limitations or intrinsic methodological boundaries.

D.8 Information-Theoretic Bounds for Target-Channel Suppression

Recent work (Hong et al., 2025) has demonstrated that standard gradient-based unlearning often creates an ‘‘illusion of unlearning’’: the model's output probabilities on target queries are suppressed, but the underlying *parametric knowledge traces*—information about the target data encoded in the weight matrices—remain largely intact. This enables adversarial extraction through techniques such as jailbreaking or in-context priming. Furthermore, (Shumailov et al., 2024) shows that even when parametric traces are removed, LLMs' in-context learning (ICL) capabilities can enable knowledge reconstruction from residual foundational knowledge, a phenomenon termed ‘‘UnUnlearning’’—highlighting that parameter-level deletion alone may be insufficient without safeguarding the semantic boundaries around the target concepts. For reasoning models, R-TOFU (Yoon et al., 2025) further shows that residual knowledge can appear inside chain-of-thought traces even when final-answer metrics look cleaner; our adversarial CoT and budgeted extraction audits are a bounded probe of that issue, not a full reasoning-model benchmark.

In this section, we provide supporting information-theoretic calculations showing why KL-to-uniform forgetting, QuantNoise regularization, and Active Retention can reduce target-channel extractability beyond behavioral suppression. We formalize this at three levels: the output distribution (Theorem 3), the internal representation (Corollary 2), and the parameter space (Theorem 5). These results complement the neighbor-protection results of Proposition 2 and Theorem 2 by establishing *conditional target-channel suppression* bounds under the stated convergence and perturbation assumptions—not an unconditional claim that target knowledge is irreversibly removed from all possible future fine-tuning trajectories.

D.9 Conditional Target Softmax-Covariance Suppression

The Fisher Information Matrix (FIM) of the target data measures how sensitively the model's output distribution on target queries depends on the model parameters. Intuitively, if the target-output Fisher is large, then small parameter perturbations can strongly change target-query outputs; if it is small, the output channel on those queries is locally less

informative. We use this as an output-channel diagnostic, not as evidence that all hidden or parametric traces have been deleted.

Definition 1 (Target Fisher Information Matrix). For a model $P_\theta(y|x)$ with logit function $z_\theta(x) \in \mathbb{R}^{|\mathcal{V}|}$ and softmax output $P_\theta(y|x) = e^{z_y}/\sum_j e^{z_j}$, the target-specific Fisher Information Matrix is:

$$\mathcal{F}_{\mathrm{TGT}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{target}}}\left[J_\theta(x)^\top \Lambda_\theta(x)\, J_\theta(x)\right] \quad (15)$$

where $J_\theta(x) = \partial z_\theta(x)/\partial\theta \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the logit Jacobian and $\Lambda_\theta(x) = \mathrm{diag}(P_\theta(\cdot|x)) - P_\theta(\cdot|x)\,P_\theta(\cdot|x)^\top$ is the softmax covariance matrix.

Theorem 3 (Conditional Target Softmax-Covariance Suppression under KL-to-Uniform). Let θ_0 denote the pre-unlearning parameters with concentrated target predictions $P_{\theta_0}(y^*|x_t) = p_0$ for some dominant token y^*, and let θ^* denote the post-unlearning parameters satisfying $\|P_{\theta^*}(\cdot|x_t) - U\|_{\mathrm{TV}} \le \epsilon_f$ for all $x_t \in \mathcal{D}_{\mathrm{target}}$, where $U = 1/|\mathcal{V}|$ is the uniform distribution. Then:

(a) The spectral norm of the softmax covariance satisfies:

$$\frac{\|\Lambda_{\theta^*}(x_t)\|_{\mathrm{op}}}{\|\Lambda_{\theta_0}(x_t)\|_{\mathrm{op}}} \le \frac{1}{|\mathcal{V}|\cdot p_0(1-p_0)} + O(\epsilon_f) \quad (16)$$

(b) *(Conditional Fisher bound.)* If the logit Jacobian norms are bounded, the spectral norm of the target Fisher satisfies:

$$\|\mathcal{F}_{\mathrm{TGT}}(\theta^*)\|_{\mathrm{op}} \le \left(\frac{1}{|\mathcal{V}|} + O(\epsilon_f)\right) \sup_{x_t} \|J_{\theta^*}(x_t)\|_{\mathrm{op}}^2 \quad (17)$$

Note that part (a) is unconditional given KL-to-uniform convergence; part (b) additionally requires that the logit Jacobian J_{θ^*} does not grow to compensate the $1/|\mathcal{V}|$ suppression. In Q-ROU, SLUG confines updates to ∼18% of layers and EWC constrains $\|\theta^* - \theta_0\|/\|\theta_0\| < 1\%$ per layer, which by standard operator-norm perturbation bounds (Stewart and Sun, 1990) limits $\|J_{\theta^*}\|_{\mathrm{op}}/\|J_{\theta_0}\|_{\mathrm{op}} \le 1 + O(\|\theta^* - \theta_0\|/\|\theta_0\|) \approx 1.01$, so Jacobian inflation is negligible in practice.

(c) For typical LLM vocabulary sizes $|\mathcal{V}| \sim 10^5$ and original confidence $p_0 \ge 0.5$, the softmax-covariance suppression represents a reduction by a factor of at least $|\mathcal{V}|\cdot p_0(1-p_0) \ge 2.5\times 10^4$ (over four orders of magnitude). Under the Jacobian stability condition in (b), this propagates to a target Fisher spectral norm reduction of the same order.

Proof. We analyze the spectral properties of the softmax covariance matrix Λ at the uniform distribution and compare with the original concentrated prediction.

Step 1 (Eigenanalysis of Λ_U). At the uniform distribution $P = U = \frac{1}{|\mathcal{V}|}\mathbf{1}$:

$$\Lambda_U = \frac{1}{|\mathcal{V}|} I_{|\mathcal{V}|} - \frac{1}{|\mathcal{V}|^2}\mathbf{1}\mathbf{1}^\top \quad (18)$$

The eigenvalues of Λ_U are: $\lambda_1 = 0$, with eigenvector $\mathbf{1}$ (since $\Lambda_U\mathbf{1} = \frac{1}{|\mathcal{V}|}\mathbf{1} - \frac{1}{|\mathcal{V}|}\mathbf{1} = \mathbf{0}$); and $\lambda_k = \frac{1}{|\mathcal{V}|}$ for $k = 2,\dots,|\mathcal{V}|$, with multiplicity $|\mathcal{V}| - 1$. Therefore $\|\Lambda_U\|_{\mathrm{op}} = 1/|\mathcal{V}|$.

Step 2 (Spectral norm of Λ_{θ_0}). For a concentrated prediction $P_{\theta_0}(y^*|x_t) = p_0 \gg 1/|\mathcal{V}|$ with the remaining probability $1-p_0$ spread over other tokens, the softmax covariance has a dominant eigenvalue of at least $p_0(1-p_0)$ (corresponding to the direction $e_{y^*} - P_{\theta_0}$, where e_{y^*} is the standard basis vector). Thus $\|\Lambda_{\theta_0}\|_{\mathrm{op}} \ge p_0(1-p_0)$.

Step 3 (Ratio bound). By Weyl's eigenvalue perturbation inequality, when $\|P_{\theta^*}(\cdot|x_t) - U\|_{\mathrm{TV}} \le \epsilon_f$, the eigenvalues of Λ_{θ^*} are within $O(\epsilon_f)$ of those of Λ_U. Therefore:

$$\begin{aligned}\frac{\|\Lambda_{\theta^*}\|_{\mathrm{op}}}{\|\Lambda_{\theta_0}\|_{\mathrm{op}}} &\le \frac{|\mathcal{V}|^{-1} + O(\epsilon_f)}{p_0(1-p_0)} \\ &= \frac{1}{|\mathcal{V}|\cdot p_0(1-p_0)} + O(\epsilon_f)\end{aligned} \quad (19)$$

Step 4 (Fisher spectral norm bound). From the definition of the target Fisher (Eq. 15):

$$\begin{aligned}\|\mathcal{F}_{\mathrm{TGT}}(\theta^*)\|_{\mathrm{op}} &\le \mathbb{E}_{x_t}\Big[\|J_{\theta^*}(x_t)^\top \Lambda_{\theta^*}(x_t)\, J_{\theta^*}(x_t)\|_{\mathrm{op}}\Big] \\ &\le \mathbb{E}_{x_t}\left[\|J_{\theta^*}(x_t)\|_{\mathrm{op}}^2 \cdot \|\Lambda_{\theta^*}(x_t)\|_{\mathrm{op}}\right] \\ &\le \left(\frac{1}{|\mathcal{V}|} + O(\epsilon_f)\right) \sup_{x_t} \|J_{\theta^*}(x_t)\|_{\mathrm{op}}^2\end{aligned} \quad (20)$$

where the second inequality uses the submultiplicativity of the operator norm. ∎

Remark 3 (Why Gradient Ascent Does Not Achieve Fisher Collapse). Standard Gradient Ascent (GA) maximizes the cross-entropy loss on target data, which pushes target probabilities toward zero but does *not* drive them toward uniformity. Depending on the learning rate and training budget, GA produces one of two regimes—neither of which constitutes robust target-channel suppression:

Regime A (Moderate GA, e.g., with AR constraint): The post-GA prediction is concentrated

on *wrong* tokens: $P_{\theta_{\text{GA}}}(y^*|x_t) \approx 0$, but there exists some $y' \neq y^*$ with $P_{\theta_{\text{GA}}}(y'|x_t) \gg 1/|\mathcal{V}|$. In this regime, $\|\Lambda_{\theta_{\text{GA}}}\|_{\text{op}} \approx p'(1-p')$ where $p' = \max_{y'} P_{\theta_{\text{GA}}}(y'|x_t)$, and the entropy remains low ($H \ll \log|\mathcal{V}|$). The target Fisher spectral norm stays $O(1)$—the same order as the pre-unlearning value—and the parametric trace is preserved. This is the ''illusion of unlearning'': the output has been *redirected* to a specific wrong answer, not made *uninformative*, allowing adversarial probes to recover the original target.

Regime B (Aggressive unconstrained GA): Without retention constraints, aggressive GA can drive the model to assign probability ≈ 1.0 to a single token (possibly the same wrong token for all queries), yielding $\|\Lambda\|_{\text{op}} \approx 0$ and $H \approx 0$. Superficially, the spectral norm also becomes small—but this is a *degenerate low-entropy regime* fundamentally different from KL-to-uniform. The model has not become *uninformative*; it has become *deterministic*, encoding maximal certainty about a specific (wrong) output. An adversary observing $p_{\max} = 1.0$ immediately knows the model was aggressively modified in this direction, and the underlying knowledge can be recovered by examining which token was selected and how the model's behavior differs from a retrained model. Moreover, this regime catastrophically destroys general capabilities (entropy ≈ 0 on all queries, including unrelated ones).

In both regimes, GA fails to achieve the KL-to-uniform target. The critical distinction is captured by *entropy*, not spectral norm alone:

- **KL-to-uniform**: $H(Y|X_t) \to \log|\mathcal{V}|$, $\|\Lambda\|_{\text{op}} \to 1/|\mathcal{V}|$ (maximum entropy, genuinely uninformative);
- **GA (Regime A)**: $H \ll \log|\mathcal{V}|$, $\|\Lambda\|_{\text{op}} = O(1)$ (low entropy, concentrated on wrong token);
- **GA (Regime B)**: $H \approx 0$, $\|\Lambda\|_{\text{op}} \approx 0$ (zero entropy, degenerate single-token collapse).

The entropy–spectral-norm pair $(H, \|\Lambda\|_{\text{op}})$ thus serves as a two-dimensional diagnostic: the strongest target-channel suppression corresponds to the *unique* point $(\log|\mathcal{V}|, 1/|\mathcal{V}|)$, which only KL-to-uniform reaches under this analysis.

This analysis connects directly to the empirical findings in Section C.10: GA+AR's 78% depth-probing pass rate (vs. Q-ROU's 100%) reflects Regime A behavior, where redirected target-query outputs leave a larger extraction surface. Our companion experiment (Section D.14) confirms both regimes quantitatively: unconstrained GA enters Regime B with $p_{\max} = 1.0$ and $H = 0$, while KL-to-uniform achieves $H/\log|\mathcal{V}| > 0.999$ and $\|\Lambda\|_{\text{op}} \approx 1/|\mathcal{V}|$. ◇

Remark on Jacobian stability. As noted in part (b) above, the full target Fisher bound requires that the logit Jacobian does not inflate to compensate the $1/|\mathcal{V}|$ softmax-covariance suppression. In Q-ROU, this stability is ensured by SLUG (restricting updates to ∼18% of layers) and EWC (constraining $\|\theta^* - \theta_0\|/\|\theta_0\| < 1\%$ per updated layer; Section A.4). By standard operator-norm perturbation bounds for compositions of Lipschitz maps (Stewart and Sun, 1990), this limits $\|J_{\theta^*} - J_{\theta_0}\|_{\text{op}}/\|J_{\theta_0}\|_{\text{op}} = O(\|\theta^* - \theta_0\|/\|\theta_0\|) \approx 0.01$, making Jacobian inflation negligible relative to the $10^4\times$ covariance suppression. We therefore use ''target Fisher spectral suppression'' as shorthand for this combined diagnostic (softmax-covariance suppression + bounded-Jacobian assumption), with the understanding that the rigorous core is the softmax-covariance statement in part (a).

D.10 Bounded Output-Channel Leakage under Target-Query Assumptions

Theorem 3 establishes that the target Fisher spectral norm is strongly suppressed under KL-to-uniform. We now translate this into a conditional operational bound for an output-channel adversary whose information comes through target-related model outputs covered by the KL-to-uniform condition.

Theorem 4 (Target-Query Output-Channel Leakage Bound). Consider a black-box adversary who submits B adaptive target-related queries $x_1, \ldots, x_B$ to the model P_{θ^*} and observes the output distributions $O_i = P_{\theta^*}(\cdot|x_i)$. The adversary's goal is to identify the correct target answer y_f for a specific target query x_f from $\mathcal{D}_{\text{target}}$. Assume each queried output distribution remains inside a covered target-query family for which the residual target information per adaptive query is bounded as

$$I(y_f; O_i \mid O_{1:i-1}) \leq g(\epsilon_f), \tag{21}$$

where $g(\epsilon_f) = 2\epsilon_f \log_2(|\mathcal{V}|/\epsilon_f)$ is a conservative leakage envelope associated with the measured KL-to-uniform residual $\epsilon_f = \max_{x_t} D_{\text{KL}}(P_{\theta^*}(\cdot|x_t)\|U)$. This is an explicit output-channel assumption: it rules out side

channels and queries whose target information is not controlled by the measured KL-to-uniform residual. Under these conditions, the adversary's probability of error satisfies:

$$P_e \geq 1 - \frac{B \cdot g(\epsilon_f) + 1}{\log_2 |\mathcal{V}|} \tag{22}$$

For $\epsilon_f \leq 0.01$ (the conservative operational target in Q-ROU) and $|\mathcal{V}| = 128{,}256$ (Llama tokenizer), the worst-case bound at $B = 16$ gives $P_e \geq 0.50$ (see proof below). If the covered output-channel residual is reduced to $\epsilon_f \approx 10^{-4}$, the same conservative envelope tightens to $P_e \gtrsim 0.94$ for this restricted threat model.

Proof.

Step 1 (Per-query information assumption). For any single covered target-related query x_i, the theorem assumes $I(y_f; O_i \mid O_{1:i-1}) \leq g(\epsilon_f)$. This is deliberately stated as an assumption rather than derived solely from target-token pass counts: a near-uniform output distribution on the audited target query family is evidence that the output channel carries little target-specific information, but paraphrase, retrieval, parameter-access, or side-information channels must be evaluated separately. Our depth, CoT, and budgeted extraction audits test several such families empirically rather than relying on this theorem alone.

Step 2 (Accumulated information over B queries). Let $O_{1:B} \stackrel{\text{def}}{=} (O_1, \ldots, O_B)$. By the chain rule for mutual information:

$$I(y_f; O_{1:B}) = \sum_{i=1}^{B} I(y_f; O_i \mid O_{1:i-1}) \leq B \cdot g(\epsilon_f) \tag{23}$$

The inequality follows directly from the uniform per-query conditional bound over the covered target-query family. Consequently, within this restricted output-channel model, no adaptive strategy can accumulate more than $B \cdot g(\epsilon_f)$ total mutual information about y_f.

Step 3 (Fano's inequality). By Fano's inequality (Cover and Thomas, 2006), the probability of correctly identifying y_f from $|\mathcal{V}|$ candidates satisfies:

$$P_e \geq 1 - \frac{I(y_f; O_{1:B}) + 1}{\log_2 |\mathcal{V}|} \geq 1 - \frac{B \cdot g(\epsilon_f) + 1}{\log_2 |\mathcal{V}|} \tag{24}$$

Substituting the Q-ROU parameters: $\epsilon_f \leq 0.01$, $B = 16$, $\log_2 |\mathcal{V}| = \log_2(128{,}256) \approx 16.97$ bits, with $g(0.01) = 2 \times 0.01 \times \log_2(128{,}256/0.01) \approx 0.472$ bits:

$$P_e \geq 1 - \frac{16 \times 0.472 + 1}{16.97} \approx 0.496 \tag{25}$$

If one instead plugs in a covered-channel residual $\epsilon_f = 10^{-4}$, then $g(10^{-4}) \approx 0.00605$ bits and the same bound gives $P_e \approx 0.935$. This remains a conservative diagnostic because the theorem controls only the covered target-output channel and not broader paraphrase or side-information families. ■

Corollary 1 (Extraction Budget Scaling). For the Fano lower bound to permit extraction success probability $1 - P_e \geq \alpha$, the adversary requires at least:

$$B \geq \frac{\alpha \log_2 |\mathcal{V}| - 1}{g(\epsilon_f)} \tag{26}$$

For $\epsilon_f = 10^{-4}$ and $\alpha = 0.5$: $B \geq \frac{0.5 \times 16.97 - 1}{0.00605} \approx 1{,}237$ queries—far beyond practical attack budgets. This is consistent with the empirical observation of 0% extraction at $B = 16$ for the 160-step FP16 checkpoint (Table 53); the 40-step checkpoint still leaves a bounded extraction channel.

Caveat. Theorem 4 applies only to the target-query output channel specified above. A parameter-access adversary, a retrieval-augmented attacker, or a black-box attacker with semantic side information outside the covered query family can bypass this bound; for those settings, the representation-level and flat-basin analyses below are diagnostics, not complete leakage guarantees.

D.11 Representation-Level Structural Divergence

Theorem 3 establishes that the target Fisher spectral norm is driven to $O(1/|\mathcal{V}|)$ at the output level. Under a Lipschitz post-layer map, any measured target-side logit displacement implies a corresponding lower bound on hidden-state displacement at SLUG layers. This provides a consistency check for the empirical observation of cosine similarity drops (*e.g.*, 0.712 at layer 26 in Table 51); it does not imply that all representation change is forced by output uniformization alone.

Corollary 2 (Representation Divergence at SLUG Layers). Let $h_\theta^l(x) \in \mathbb{R}^{d_h}$ denote the hidden representation at layer l for input x. For a SLUG layer $l \in \mathcal{S}$, decompose the forward computation as $z_\theta(x) = g^l_{\text{post}}(h^l_\theta(x))$, where g^l_{post} denotes the mapping from layer-l representations to output logits through layers $l{+}1, \ldots, L$ and the unembedding head. Let $\tilde{z} = z - \frac{1}{|\mathcal{V}|}\sum_j z_j \mathbf{1}$ denote centered logits

and define the observed centered-logit displacement $\gamma(x_t) = \|\tilde{z}_{\theta^*}(x_t) - \tilde{z}_{\theta_0}(x_t)\|$. If g^l_{post} is Lipschitz with constant L_{post}, then:

$$\|h^l_{\theta^*}(x_t) - h^l_{\theta_0}(x_t)\| \geq \frac{\gamma(x_t)}{L_{\text{post}}} \tag{27}$$

When KL-to-uniform convergence makes $\tilde{z}_{\theta^*}(x_t)$ small, $\gamma(x_t)$ is approximately the pre-edit centered-logit norm on that target query; it should be measured or bounded from logits rather than inferred solely from the target probability.

Proof. By the reverse triangle inequality applied to the Lipschitz map:

$$\|z_{\theta^*}(x_t) - z_{\theta_0}(x_t)\| \leq L_{\text{post}} \cdot \|h^l_{\theta^*}(x_t) - h^l_{\theta_0}(x_t)\| \tag{28}$$

Rearranging: $\|h^l_{\theta^*} - h^l_{\theta_0}\| \geq \|z_{\theta^*} - z_{\theta_0}\|/L_{\text{post}}$. Because softmax is invariant to additive constants, we evaluate the identifiable component of logit displacement through centered logits. The centering operator is an orthogonal projection, so the measured centered displacement $\gamma(x_t)$ lower-bounds the constant-shift-invariant logit change that the post-layer map must realize. Combining the Lipschitz inequality with this measured displacement yields Eq. 27. ∎

Interpretation. For a representative centered-logit displacement $\gamma(x_t) \approx 10$ nats and $L_{\text{post}} \sim 10^2$ (typical for a few-layer transformer block + head), the bound gives $\|h^* - h_0\| \gtrsim 0.1$ in embedding space. Given typical hidden state norms $\|h\| \sim$ 10–50 (model-dependent), this implies a cosine deviation of at least ~ 0.002–0.01 purely from output-level constraints. The empirically observed cosine drop to 0.712 at layer 26 (Table 51)—corresponding to $\|h^* - h_0\| \approx 0.76\|h_0\|$—far exceeds this lower bound, indicating that the optimization actively restructures target representations well beyond what output-level uniformization strictly requires. This excess structural modification is consistent with the interplay between the forget loss (which creates gradient pressure across multiple SLUG layers) and the orthogonality constraint (which redirects updates away from truth-direction components), jointly producing the observed representation-level separation on target probes.

D.12 Conditional Flat-Basin Indistinguishability Bound

The previous results establish output-level and representation-level bounded-suppression results. We now address the strongest adversarial setting: a white-box adversary with direct access to the model weights θ^*. We give a conditional privacy-style calculation showing what would be required for QuantNoise's flat-basin effect (Theorem 1) to support an (ϵ, δ)-style indistinguishability statement in the spirit of unlearning definitions (Ginart et al., 2019). This is an assumption audit, not an empirical guarantee against a retrained model.

Definition 2 ((ϵ, δ)-Unlearning Certificate (Ginart et al., 2019)). An unlearning mechanism M satisfies (ϵ, δ)-unlearning for forget set D_f if for all measurable sets S in the output space:

$$\Pr[M(D, D_f) \in S] \leq e^\epsilon \Pr[A(D \setminus D_f) \in S] + \delta \tag{29}$$

where $M(D, D_f)$ is the unlearning procedure applied to a model trained on D to forget D_f, and $A(D \setminus D_f)$ is re-training from scratch on the retained data.

Theorem 5 (Conditional Gaussian Indistinguishability Bound). Let θ^* denote the Q-ROU output and let θ_R denote a hypothetical model retrained on $D \setminus D_f$. Suppose that:

(i) QuantNoise training places θ^* in a flat basin of effective radius r, meaning $\mathcal{L}(\theta) \leq \mathcal{L}(\theta^*) + \delta_L$ for all $\theta \in B_r(\theta^*)$, where δ_L is a small loss tolerance, and the release noise remains inside this basin with probability at least $1 - \delta_b$;

(ii) The distance between the unlearned and retrained models satisfies $\|\theta^* - \theta_R\|_2 \leq D_{\text{UR}}$;

(iii) Independent Gaussian noise with per-coordinate variance $\sigma^2 = \epsilon_0^2/3$ (matching the variance of the uniform QuantNoise perturbation) is added to produce the released model: $\tilde{\theta} = \theta + \mathcal{N}(0, \sigma^2 I_d)$.

Then the noised model distributions centered at θ^* and θ_R satisfy an (ϵ, δ) indistinguishability bound with:

$$\epsilon = \frac{D^2_{\text{UR}}}{2\sigma^2} + \frac{D_{\text{UR}}}{\sigma}\sqrt{2\ln(1/\delta)} \tag{30}$$

where $\sigma^2 = \epsilon_0^2/3$ is the per-coordinate variance. Under condition (i), adding noise $\mathcal{N}(0, \sigma^2 I_d)$ does not degrade model quality beyond δ_L with probability at least $1 - \delta_b$.

Proof. The two distributions $\tilde{\theta}^* \sim \mathcal{N}(\theta^*, \sigma^2 I_d)$ and $\tilde{\theta}_R \sim \mathcal{N}(\theta_R, \sigma^2 I_d)$ are Gaussians with the same covariance and different means. By the standard Gaussian mechanism analysis for differential privacy (Dwork and Roth, 2014):

Step 1 (Privacy loss random variable). The log-likelihood ratio between the two densities at a point θ is:

$$\begin{aligned} \ln \frac{p_{\tilde{\theta}^*}(\theta)}{p_{\tilde{\theta}_R}(\theta)} &= \frac{1}{2\sigma^2}\left(\|\theta-\theta_R\|^2 - \|\theta-\theta^*\|^2\right) \\ &= \frac{(\theta^*-\theta_R)^\top(2\theta-\theta^*-\theta_R)}{2\sigma^2} \end{aligned} \tag{31}$$

Under $\theta \sim \mathcal{N}(\theta^*, \sigma^2 I)$, the random variable $(\theta^* - \theta_R)^\top\theta$ is Gaussian with mean $\|\theta^* - \theta_R\|^2$ and variance $\sigma^2\|\theta^* - \theta_R\|^2$. Thus the privacy loss has mean $\mu_{\text{PL}} = D_{\text{UR}}^2/(2\sigma^2)$ and standard deviation $\sigma_{\text{PL}} = D_{\text{UR}}/\sigma$.

Step 2 (Concentrated divergence bound). By the Gaussian tail bound:

$$\Pr\left[\ln \frac{p_{\tilde{\theta}^*}(\theta)}{p_{\tilde{\theta}_R}(\theta)} > \epsilon\right] \le \exp\left(-\frac{(\epsilon-\mu_{\text{PL}})^2}{2\sigma_{\text{PL}}^2}\right) \tag{32}$$

Setting the right-hand side equal to δ and solving for ϵ yields:

$$\begin{aligned} \epsilon &= \mu_{\text{PL}} + \sigma_{\text{PL}}\sqrt{2\ln(1/\delta)} \\ &= \frac{D_{\text{UR}}^2}{2\sigma^2} + \frac{D_{\text{UR}}}{\sigma}\sqrt{2\ln(1/\delta)} \end{aligned} \tag{33}$$

This is the standard Gaussian-mechanism indistinguishability calculation with ℓ_2 mean separation D_{UR} and noise scale σ. ∎

Remark 4 (Operationalizing Theorem 5). The bound requires bounding $D_{\text{UR}} = \|\theta^* - \theta_R\|_2$, the distance between the unlearned and retrained models. While θ_R is not directly available (computing it would require full retraining), we can establish the following structural bounds:

(a) **Decomposition**: $D_{\text{UR}} = \|\theta^* - \theta_R\| \le \|\theta^* - \theta_0\| + \|\theta_0 - \theta_R\|$. The first term is directly observable: under Q-ROU's SLUG constraint, only ∼18% of parameters are modified (5 of 28 layers), and $\|\Delta\theta_l\|/\|\theta_l\| < 1\%$ per layer.

(b) **Retraining proximity**: For small forget sets ($|D_f| \ll |D|$), influence function theory (Koh and Liang, 2017) gives $\|\theta_0 - \theta_R\| \approx \|H^{-1}\sum_{(x,y)\in D_f}\nabla_\theta \ell(\theta_0; x, y)\|$, which is small when the forget set is a negligible fraction of the full training corpus. For $|D_f| = 18$ (our multi-entity setup) vs. pre-training corpus sizes of 10^9–10^{12} tokens, this ratio is vanishingly small. However, we acknowledge that indirect estimation via influence functions provides a structural heuristic rather than a strict cryptographical guarantee, as accurately bounding this term in highly non-linear deep models without full retraining remains an open challenge.

(c) **Flat basin as quality rationale**: Condition (i) ensures that adding noise $\mathcal{N}(0, \sigma^2 I)$ does not push the released model outside the flat basin, so model quality is maintained. The critical insight is that QuantNoise serves a *dual purpose*: (1) quantization robustness via weighted Hessian-trace regularization (Theorem 1), and (2) a Gaussian-perturbation scale that can be related to privacy-style indistinguishability if the retrained-model proximity assumption is independently satisfied.

(d) **Tightening the bound**: Increasing ϵ_0 (with correspondingly longer QuantNoise training to widen the flat basin) directly tightens ϵ by increasing σ, but only if conditions (i) and (ii) remain valid. This provides a principled mechanism to trade training cost for stronger privacy-style evidence, not a proof of retraining equivalence. ◇

D.13 Synthesis: Why Q-ROU Breaks the ''Illusion of Unlearning''

The three theorems above, combined with the existing theoretical results (Theorems 1–5), form a coherent multi-level explanation for why Q-ROU moves beyond the ''illusion of unlearning'' and toward bounded target-channel suppression:

1. **Output level (Theorem 3):** KL-to-uniform drives the target Fisher spectral norm to $O(1/|\mathcal{V}|)$, decoupling the model's target-query responses from any specific target token under the analyzed query distribution. Gradient Ascent either redirects to specific wrong tokens (Regime A, preserving $O(1)$ spectral norm) or enters a degenerate single-token regime (Regime B, $p_{\max} = 1$); in neither case does the output become truly uninformative (Remark 3). The entropy–spectral-norm diagnostic $(H, \|\Lambda\|_{\text{op}})$ distinguishes these three outcomes: $(\ln|\mathcal{V}|,\ 1/|\mathcal{V}|)$ for target-channel uniformization, moderate $(H, O(1))$ for GA Regime A, and $(0,\ 0)$ for GA Regime B.

2. **Extraction resistance (Theorem 4):** Under an explicit per-query information-leakage assumption for the covered target-output channel, the Fano bound implies that the query budget must scale as $\Omega(1/g(\epsilon_f))$ before high-probability extraction is even permitted by the bound, where $g(\epsilon_f) \to 0$ under Q-ROU's convergence. The empirical 0% extraction at

$B = 16$ for the 160-step FP16 checkpoint (Table 53) is consistent with this calculation; the 40-step checkpoint is not fully closed under the same budget, and broader query families are evaluated empirically rather than covered by the theorem.

3. **Representation level (Corollary 2):** Output-level uniformization propagates backward as structural divergence of hidden representations at SLUG layers, explaining the empirical cosine similarity drop to 0.712 at layer 26 and the selective representation reorganization pattern (Table 51).

4. **Parameter level (Theorem 5):** QuantNoise's flat basin serves dual purposes—quantization robustness (Theorem 1) and a conditional (ϵ, δ)-style indistinguishability calculation under the Gaussian perturbation model stated in the theorem. In our tested regime, the edited point tolerates the sampled perturbations used in the audit without observed quality degradation.

5. **Neighbor protection (Proposition 2, Theorem 2):** Active Retention bounds neighbor drift via the Fisher null-space mechanism, while the Davis-Kahan calculation bounds one matrix-subspace rotation channel under small updates. Together these results motivate why target-channel suppression need not spill over into neighbor knowledge, a property that standard methods do not enforce directly.

This multi-level analysis distinguishes Q-ROU from surface-level behavioral suppression: target responses are pushed toward a high-entropy output-channel regime (Theorem 4), hidden representations diverge selectively on target probes (Corollary 2), and the edited point lies in a flat region that tolerates the perturbations used in our quantization and privacy-style analyses (Theorem 5). These are bounded analytical and empirical claims, not a proof that the target fact is unrecoverable under arbitrary optimization.

Remark 5 (Extensibility to Broader Unlearning Settings). The three theorems presented here are not specific to the Q-ROU framework—they apply to *any* unlearning method that achieves KL-to-uniform convergence on target queries under comparable local assumptions. However, Q-ROU is currently the only method in our evaluation that simultaneously achieves (a) near-uniform target outputs ($\epsilon_f < 10^{-3}$ on average), (b) preserved neighbor/general behavior, and (c) a flat loss basin supporting the conditional perturbation analysis. Methods that achieve only behavioral suppression (GA, GradDiff, RepBend) fail at (a); methods without AR fail at (b); methods without QuantNoise fail at (c). This framework therefore identifies a precise set of conditions for stronger target-channel suppression claims, and provides a template for evaluating future unlearning methods: measure ϵ_f (KL-to-uniform gap), the neighbor Fisher preservation, and the basin width. Recent surveys (Nguyen et al., 2025; Geng et al., 2025; Liu et al., 2025) have identified key challenges in machine unlearning including the trade-off between forgetting efficacy and model utility, and the difficulty of achieving robust suppression rather than mere behavioral masking. Beyond LLM unlearning, the target-output Fisher diagnostic in Theorem 3 may be useful in privacy-preserving inference, federated unlearning, and concept-suppression settings in diffusion models—any setting where driving outputs to uniform is a viable suppression strategy. ◇

D.14 Empirical Verification of Fisher Spectral Collapse

We verify Theorem 3's predictions on Qwen2.5-0.5B ($|\mathcal{V}| = 151{,}936$) using a controlled inject-then-unlearn protocol. Four fictional QA pairs (e.g., *''What is the capital of Zandaloria?''* $\rightarrow$ *''Pyrithos''*) are first fine-tuned into the model (20 steps, $p_{\max} \geq 0.50$ for all targets), then unlearned under four conditions: KL-to-uniform alone, KL-to-uniform with Active Retention (KL+AR), Gradient Ascent alone (GA), and GA with Active Retention (GA+AR). For AR conditions, a frozen copy of the *pre-injection* model serves as the anchor, with $\lambda_a = 5.0$ on four general-knowledge queries.

Primary metric: Softmax covariance spectral norm. The key observable in Theorem 3 is $\|\Lambda\|_{\text{op}}$, the spectral norm of the softmax covariance $\Lambda = \text{diag}(P) - PP^\top$. This quantity admits a parameter-free prediction: $\|\Lambda_U\|_{\text{op}} = 1/|\mathcal{V}| = 6.58 \times 10^{-6}$ at exact uniformity. Results for all four conditions are summarized in Table 74.

Result 1: Theorem 3 confirmed (KL-only). Without AR, KL-to-uniform drives target $\|\Lambda\|_{\text{op}}$ from 0.242 to 1.03×10^{-5}—a **23,466**× reduction,

Table 74: Empirical verification of Theorem 3 on Qwen2.5-0.5B ($|\mathcal{V}| = 151{,}936$). KL-to-uniform (without AR) drives $\|\Lambda\|_{\text{op}}$ to $O(1/|\mathcal{V}|)$ and entropy to $\log|\mathcal{V}|$, confirming near-uniform target-channel suppression. With AR, general knowledge is preserved (4.7% change) but target convergence is slowed. GA (with or without AR) never approaches uniform.

Method	$\|\Lambda\|_{\text{op}}$ (Target)	Entropy (Target)	$p_{\max}$ (Target)	$\|\Lambda\|_{\text{op}}$ (General)
Post-inject	0.2417	0.38	0.782	0.0215
KL-Uniform	1.03×10^{-5}	11.93	1.03×10^{-5}	1.03×10^{-5}
KL-Uniform+AR	0.0462	0.28	0.961	0.0205
GA (η=2×10^{-5})	1.44×10^{-22}	0.00	1.000	1.26×10^{-22}
GA+AR (η=10^{-4})	0.0091	0.06	0.993	0.0097
Theoretical (U)	6.58×10^{-6}	11.93	6.58×10^{-6}	—

closely matching the predicted ∼25,900× (using observed $p_0 = 0.782$ from Table 74). Target entropy reaches $11.93 \approx \ln(151{,}936)$, indicating near-convergence to the uniform distribution at the measured precision. The remaining discrepancy (1.03×10^{-5} vs. theoretical 6.58×10^{-6}) reflects residual $\epsilon_f > 0$ after 30 steps. However, without AR, general-knowledge queries are also driven to uniform ($\|\Lambda\|_{\text{op}} = 1.03\times10^{-5}$, entropy = 11.93), confirming that unconstrained KL-to-uniform destroys all model knowledge—the AR mechanism is structurally essential.

Result 2: AR preserves general knowledge. With AR ($\lambda_a = 5.0$), general-query $\|\Lambda\|_{\text{op}}$ changes by only 4.7% (from 0.0215 to 0.0205), nearly perfectly preserving pre-injection behavior. By contrast, GA+AR permits a 55% change in general $\|\Lambda\|_{\text{op}}$ ($0.0215 \to 0.0097$), and conditions without AR destroy general knowledge entirely. This demonstrates that AR's Fisher-null-space protection (Proposition 2) operates far more effectively under KL-to-uniform than under GA.

Result 3: AR–convergence tradeoff (KL+AR). KL+AR after 30 steps achieves only a 5.2× target reduction ($0.242 \to 0.046$), far below the 23,466× of KL-only. This reflects a fundamental tradeoff: the AR penalty constrains the optimization to preserve general-query behavior, slowing convergence on target queries. Crucially, while the Fisher suppression is incomplete, *functional suppression is still achieved*: post-KL+AR generations (e.g., *''The capital of Z capital of Z...''*; see Table 74) contain no trace of the injected answers (''Pyrithos'', ''Vexalion'', ''1247''), and the model reverts to uninformative repetitions. In the full Q-ROU framework, three mechanisms resolve this convergence gap: (a) SLUG restricts updates to ∼18% of layers, reducing interference between the forget and retention objectives; (b) the unlearning runs for 40 steps (vs. 30 here) with a tuned $\lambda_a \in [40, 160]$; and (c) the retain set is curated *structural neighbors*, not general knowledge, avoiding direct competition with the target gradient.

Result 4: GA confirms Remark 3. GA without AR (even at reduced $\eta = 2\times10^{-5}$) enters Regime B: $p_{\max} = 1.0$, entropy = 0, generations degenerate to single-token repetition (''osos...''). GA+AR produces Regime A: the model confidently generates wrong tokens ($p_{\max} = 0.993$, entropy = 0.06) but retains structural coherence (e.g., *''The capital of Z capital of Z...''*). Neither GA variant approaches uniform: the spectral norm for GA+AR (0.0091) is three orders of magnitude above the theoretical uniform value (6.58×10^{-6}), confirming that the parametric trace of learned response structure is *preserved* under GA.

The entropy–spectral-norm diagnostic. These four conditions validate the two-dimensional diagnostic $(H, \|\Lambda\|_{\text{op}})$ proposed in Remark 3:

- **KL-only**: (11.93, 1.03×10^{-5}) — near-uniform, Fisher-suppressed;
- **KL+AR**: (0.28, 0.046) — target-channel uniformization is strongly slowed by AR, while general knowledge is preserved;
- **GA+AR**: (0.06, 0.009) — concentrated on wrong tokens, parametric trace intact;
- **GA-only**: (0.00, 10^{-22}) — degenerate collapse, model destroyed.

A single scalar (spectral norm or entropy alone) conflates genuinely uninformative outputs with degenerate or wrong-answer outputs; the pair uniquely identifies each regime.

As a lightweight robustness check, a separate restricted-update probe on Qwen2.5-0.5B (last-block-only updates) reproduced the same qualitative ordering: KL-to-uniform raised target entropy from 6.47 to ∼10.52 bits while reducing $\|\Lambda\|_{\text{op}}$ from 0.142 to ∼0.0506, whereas GA and GA+AR remained in a lower-entropy regime (∼5.15 bits) with larger target covariance norms (∼0.249). We do not treat this lightweight probe as a replacement for Table 74, nor as new primary evidence for AR-based retention preservation; rather, it confirms that

the entropy–spectral-norm diagnostic remains qualitatively stable even under a much smaller restricted-update setting.

A note on parameter-level Fisher. In v1 of this experiment, the *parameter-level* diagonal Fisher (empirical Fisher, computed as $\sum_i (\nabla_\theta \ell_i)^2$) was also measured and showed the *opposite* trend: it increased by 39× after KL-to-uniform. This is expected and does not contradict Theorem 3. Theorem 3 concerns the *output-level* Fisher $\mathcal{F} = J^\top \Lambda J$, which factors through the softmax covariance Λ. The parameter-level empirical Fisher measures the average squared gradient of the *loss function* (cross-entropy), which increases when the output becomes uniform because cross-entropy loss against the (now-uniform) prediction is large. The distinction is between ''how much does the *output distribution* depend on θ'' (Theorem 3, controlled by Λ) vs. ''how steep is the *loss landscape*'' (empirical Fisher, which can increase even as Λ is suppressed).

D.15 Component Ablation: Saturation Curve and Early Stopping

To quantify each Q-ROU component's contribution and identify the optimal stopping point, we conduct a cumulative ablation study on Qwen2.5-0.5B with periodic checkpointing. Three fictional facts are injected (80 steps, $\eta = 5\times10^{-5}$), then unlearned under seven conditions: (i) baseline, (ii) KL-to-uniform only (40 steps), (iii) KL+AR (40 steps), (iv) KL+AR+SLUG with injected-model AR reference, (v) KL+AR+SLUG with pre-injection AR reference, (vi) full Q-ROU with injected reference, and (vii) full Q-ROU with pre-injection reference. Conditions (iv)–(vii) run for 160 steps with metrics recorded every 20 steps.

Key finding 1: SLUG unlocks selective forgetting. KL-only and KL+AR both achieve target $p_{\max} < 10^{-4}$ but *destroy all knowledge*: general $\|\Lambda\|_{\text{op}}$ collapses from 0.178 to 10^{-4}. At step 20, KL+AR+SLUG achieves target $p_{\max} = 5\times10^{-3}$ (< 0.01 threshold) while preserving general $\|\Lambda\|_{\text{op}}$ at 0.052 (29% of baseline). This confirms that SLUG (selective layer freezing) is the critical mechanism for selective forgetting.

Key finding 2: early stopping is essential at small scale. While target $p_{\max}$ improves monotonically with more steps, general knowledge degrades rapidly after step 20: general $\|\Lambda\|_{\text{op}}$ drops from 0.052 at step 20 to 2.2×10^{-4} at step 40—a 237× collapse. This indicates that on 0.5B models, the SLUG subspace does not provide sufficient isolation to prevent AR from being overwhelmed by the KL-uniform gradient over extended training. At 3B/7B scale (Table 77), WMDP accuracy is perfectly preserved ($\Delta = 0$) after 20-step Q-ROU, confirming that this early-stopping sensitivity is scale-dependent.

Table 75: Component ablation on Qwen2.5-0.5B ($|\mathcal{V}| = 151{,}936$). All SLUG (and full Q-ROU) conditions cross the $\tau = 0.01$ threshold by step 20, but continued training degrades general knowledge. Pre-injection vs. post-injection AR references yield indistinguishable results at this injection scale.

Method	Step	Target $p_{\max}$	$H/\ln\lvert\mathcal{V}\rvert$	General $\|\Lambda\|_{\text{op}}$
Post-inject (baseline)	0	0.9999	0.0001	0.178
KL-Uniform	40	10^{-4}	0.986	10^{-4}
KL+AR	40	10^{-4}	0.986	10^{-4}
KL+AR+SLUG	**20**	$\mathbf{5.0\times10^{-3}}$	**0.894**	**0.052**
	40	1.4×10^{-4}	0.981	2.2×10^{-4}
	80	8×10^{-5}	0.988	1.2×10^{-4}
	160	4×10^{-5}	0.993	7.6×10^{-5}
Full Q-ROU	**20**	$\mathbf{4.7\times10^{-3}}$	**0.894**	**0.055**
	40	1.4×10^{-4}	0.981	2.2×10^{-4}
	80	8×10^{-5}	0.988	1.2×10^{-4}
	160	4×10^{-5}	0.993	7.6×10^{-5}

Key finding 3: QuantNoise is compatible with SLUG-based unlearning. With properly calibrated noise scale ($\epsilon_q = |\bar{w}| \times 0.01$), the full Q-ROU condition (KL+AR+SLUG+QuantNoise+EWC) matches the KL+AR+SLUG saturation curve at every checkpoint. The QuantNoise perturbation at this relative scale (∼1% of per-tensor weight magnitude) does not interfere with the forgetting signal while still providing the loss-landscape flattening validated in Section D.16.

D.16 Empirical Verification of Loss-Landscape Flattening (Theorem 1)

Theorem 1 predicts that QuantNoise training encourages flat loss basins, reducing sensitivity to weight perturbations (including quantization noise). We verify this by measuring *perturbed sharpness*, defined as:

$$S = \frac{\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\left[\mathcal{L}(\theta+\varepsilon) - \mathcal{L}(\theta)\right]}{\sigma^2} \approx \frac{1}{2}\operatorname{tr}(H) \tag{34}$$

where H is the Hessian and the approximation holds by Taylor expansion. This formulation avoids the numerical instabilities of Hutchinson-based $\operatorname{tr}(H)$

Table 76: Perturbed-sharpness audit for Theorem 1 on Qwen2.5-0.5B. QuantNoise reduces the estimated sharpness of the 30-step checkpoint relative to vanilla training under the same perturbation scale.

Condition	Sharpness S (mean ± std)	Ratio vs. Untrained
Untrained	9,882 ± 24,205	1.00
Vanilla (30 steps)	6,401 ± 17,166	0.65
QuantNoise (30 steps)	**3,810 ± 17,426**	**0.39**

estimation in bfloat16 models.

Experimental setup. We train the last two transformer layers of Qwen2.5-0.5B for 30 steps on five factual statements, under two conditions: standard optimization (vanilla) and QuantNoise ($p = 0.1$). Sharpness is estimated using 30 Gaussian perturbations with $\sigma = 10^{-3}$.

Results. QuantNoise training reduces perturbed sharpness by **40**% relative to vanilla training at the same step count ($S_{\text{QN}}/S_{\text{vanilla}} = 0.60$). This confirms that QuantNoise implicitly regularizes a noise-weighted Hessian trace, producing flatter loss basins that are more tolerant of the sampled post-training weight perturbations used in this audit, including INT4/NF4-like quantization noise. The large standard deviations reflect the inherent stochasticity of $O(10^7)$-dimensional perturbation sampling with $n = 30$ draws; the *ordering* $S_{\text{QN}} < S_{\text{vanilla}} < S_{\text{untrained}}$ is consistent across all perturbation batches examined.

D.17 MUSE/WMDP Utility Preservation

To probe whether Q-ROU preserves general-purpose model capabilities, we evaluate general-knowledge accuracy (WMDP-style multiple-choice) and retain-domain perplexity before and after Q-ROU unlearning on Harry Potter–domain forget probes.

Key observations. (1) WMDP accuracy is *unchanged* ($\Delta = 0.000$) on both 3B and 7B, consistent with low measured damage to this general-knowledge slice. (2) Retain PPL ratio decreases with scale (1.06 at 3B → 1.02 at 7B), confirming the scale-dependent redundancy effect discussed in Section C.6: larger models have more capacity to absorb the forgetting gradient without disturbing retention. (3) Forget $p_{\max}$ on the 3B model reaches 0.004 after 40 steps, below the strict $\tau = 0.01$ threshold, indicating a target-side margin in this probe.

Table 77: Utility preservation after Q-ROU unlearning on two Qwen2.5 model scales. WMDP-style general-knowledge accuracy is perfectly preserved ($\Delta = 0$). Retain-domain perplexity increases by at most 6%, consistent with the neighbor-drift bounds predicted by Proposition 2.

Model	Condition	WMDP Acc	Forget $p_{\max}$	Retain PPL Ratio
Qwen2.5-3B	Original	0.750	0.462	—
	Q-ROU	0.750	0.004	1.060
Qwen2.5-7B	Original	0.850	—	—
	Q-ROU	0.850	—	1.018

Relative KnowMem Reduction interpretation. For MUSE, we report Relative KnowMem Reduction ($1 - \text{KnowMem}_{\text{after}}/\text{KnowMem}_{\text{base}}$), so higher indicates stronger forgetting. Under this definition, a retain/retrain baseline is expected to be near zero, with small nonzero offsets attributable to finite-sample ROUGE variability. The 3B Q-ROU value (0.0060) is therefore consistent with a minimal reduction at 40 steps, which may reflect undersaturation at this early checkpoint; we did not run a 3B step-sweep due to budget. By contrast, the 8B 160-step run yields a clearly nontrivial reduction (0.3540), indicating that the pipeline can meaningfully move the KnowMem reduction metric when trained longer.

E VulnScore: Quantifying Structural Interference

To quantitatively support the distinction between semantic and structural neighbors discussed in Section D, we developed the *VulnScore* metric to predict the interference magnitude on seemingly unrelated concepts during unlearning.

Theoretical Motivation. When unlearning a target concept T via gradient updates $\Delta\theta_T$, the interference experienced by a disjoint concept C can be approximated using the Fisher Information Matrix (FIM) of C. If F_C is the diagonal approximation of the empirical Fisher matrix computed on concept C's data, the raw vulnerability can be defined as $VS_{\text{raw}}(C|T) = \Delta\theta_T^\top F_C \Delta\theta_T$. This approximates the KL-divergence (or second-order log-likelihood damage) inflicted on C by the update $\Delta\theta_T$. However, raw VS_{raw} is heavily biased by globally fragile concepts. To isolate the *specific entanglement* between T and C, we define the Specific Entanglement Ratio

(SER):

$$\mathrm{SER}(C|T) = \frac{\Delta\theta_T^\top F_C \Delta\theta_T}{\Delta\theta_C^\top F_C \Delta\theta_C} \tag{35}$$

This normalizes the interference by the vulnerability of C to its own target update.

Methodology and Absolute Magnitude. We evaluated a symmetric grid of 50 diverse concepts across 6 domains (Harry Potter, Lord of the Rings, Star Wars, Marvel, General Geography, etc.) on Llama-3.2-3B. Crucially, when measuring the empirical damage (the actual change in log-probability $\Delta \log P_C$), we discovered that aligned gradients can cause the neighbor's probability to *improve* (a synergistic effect). To construct a strictly linear correlation metric, we must measure the absolute magnitude of structural disruption relative to the target's improvement, defined as the Absolute Relative Empirical Damage (Abs RED):

$$\text{Abs RED}(T \to C) = \frac{|\Delta \log P_C|}{|\Delta \log P_T|} \tag{36}$$

To ensure the empirical measurements remain within the valid linear-approximation regime of the Taylor expansion (preventing severe output collapse), we conducted an automated grid search over micro-perturbation scales ($\eta \in \{1e{-}3, 5e{-}4, 1e{-}4\}$).

Results: Near-Perfect Rank Correlation. At the optimal perturbation scale ($\eta = 0.001$), the Spearman rank correlation (ρ) between the theoretical prediction (SER) and the actual parameter disruption (Abs RED) across the 50×50 evaluation grid (2,450 pairs) was an exceptionally high **0.921** ($p < 10^{-60}$). This provides strong evidence that VulnScore is a highly accurate mathematical predictor for structural interference.

Structural Hubs vs. Semantic Distance. The analysis revealed specific, robust ''structural hubs''. For instance, updates to HP_Voldemort or HP_Snape reliably inflicted substantial, asymmetric damage on entirely unrelated concepts like Gen_Geography or Marvel_Spiderman. When comparing the SER against the cosine similarity of the sentence embeddings for the concepts, the correlation was exactly $\rho \approx 0$. This empirically supports the central premise of Section D: *semantic isolation does not guarantee structural safety*. Concepts that are far apart in human-defined semantic space can still collide strongly in the underlying parameter space of Llama-3.2-3B.

Implications for Q-ROU. This finding underscores the inadequacy of relying solely on generic pre-training data or semantic similarity distances to construct retention penalties. Q-ROU's Active Retention directly minimizes this predicted Fisher damage: even when the provided neighbor set is selected based on semantic human intuitions, anchoring their output probabilities constrains the update $\Delta\theta_T$ toward directions that are locally low-impact under the aggregated neighbor Fisher metric (as shown in Proposition 2). The $\rho = 0.921$ discovery provides strong empirical backing for this geometric protection strategy.

E.1 Theoretical Integration: The Divergence of Semantic and Structural Neighborhoods

Traditional machine unlearning benchmarks have implicitly operated under the assumption of an isomorphism between the semantic embedding space (how humans categorize concepts) and the neural parameter space (how the model physicalizes them). However, recent investigations into the internal structure of Large Language Models reveal that this assumption is fundamentally flawed, leading to severe blind spots in evaluating unlearning safety and efficacy.

The Geometric Structure of Concepts and Feature Crystals. Recent dictionary-learning and sparse-autoencoder analyses suggest that concepts in LLMs are not stored as isolated vectors but as distributed feature structure with non-trivial overlap (Li et al., 2025). In that setting, concepts that appear semantically distinct can still share nearby parametric support through broader latent variables. Consequently, optimization trajectories that only push down target-token probabilities can also damage neighboring structure, even when ordinary semantic-similarity checks do not reveal the problem immediately.

Parametric Knowledge Traces vs. Behavioral Suppression. This kind of overlap helps explain why standard gradient-based unlearning can achieve behavioral suppression without stable conceptual separation. As demonstrated by Hong *et al.* (Hong et al., 2025), target probabilities may fall while the underlying *parametric knowledge traces* remain largely intact or only weakly shifted. That leaves targeted knowledge vulnerable to adversarial recovery and increases the risk that aggressive parameter shifts overwrite neighboring structure.

Robust conceptual suppression therefore requires interventions that affect the target trace while explicitly protecting nearby traces, rather than relying only on output masking. The concept of knowledge neurons (Dai et al., 2022), which identifies specific FFN neurons responsible for factual recall, provides direct evidence that different facts can share overlapping neuronal substrates—a phenomenon that manifests as structural interference in our VulnScore analysis. This shared substrate explains why semantically distant concepts can exhibit high structural vulnerability.

Depth-wise Activation Topology. The location of these parametric traces is heavily depth-dependent. Lv *et al.* (Wang et al., 2025a) mathematically map the activation patterns of LLM parameters, proving that shallow layers exhibit dense activation patterns responsible for domain-agnostic, general linguistic capabilities. In contrast, middle-to-deep layers exhibit highly sparse activation patterns, acting as specialized storage for domain-specific facts and distinct entities. Applying broad parameter updates across all layers inevitably degrades the foundational dense representations in shallow layers.

Synthesizing the Q-ROU Theoretical Framework. Viewed through this advanced mechanistic lens, the empirical design choices in Q-ROU represent a direct alignment with the geometric reality of the parameter space across three key dimensions.

First, **SLUG explicitly targets specific sparse traces** by restricting updates to highly selective middle-to-deep layers (e.g., layer subsets around 17–26 in Llama-3.2-3B). This allows the mechanism to bypass the dense, domain-agnostic shallow layers entirely (Wang et al., 2025a), thereby preventing general capability degradation while operating precisely where the targeted parametric knowledge traces reside.

Second, **Active Retention explicitly protects nearby structure** by anchoring neighbor distributions during the edit. As established in Proposition 2, this biases the unlearning update $\Delta\theta_T$ toward directions that are approximately null under the neighbor-aggregated Fisher Information Matrix. Even when starting from human-provided semantic neighbor sets, the mechanism translates semantic anchoring into a local structural constraint that protects shared functional substructure from perturbation.

Third, **the Wronskian serves as a supplementary collision alarm** to complement the main analytical results. As an empirical proxy (Observation 1, Lemma 1), the Wronskian risk signal ρ_k monitors structural interference online. When standard gradient ascent begins to overwrite shared retained structure, the AR surrogate models show converging roots, causing $\rho_k \to 0$. The adaptive variant of Q-ROU uses this signal to halt the update trajectory before retained traces drift further (Hong et al., 2025).

Ultimately, the discovery regarding structural interference—such as our VulnScore metric achieving a strong rank correlation of $\rho = 0.921$ with parameter damage despite having virtually zero correlation with semantic distance—supports this paradigm. Q-ROU succeeds not merely because of scalar regularization, but because it mechanistically recognizes and restricts its optimization within the geometric layout of parametric knowledge traces.

E.2 Appendix Summary

This appendix provides extended diagnostics, robustness analyses, and theoretical addenda that support the primary claims of the main text. We first present detailed step-sweeps and sensitivity studies on the expanded 65-probe set, including comprehensive ablations of the adaptive mechanisms and hyperparameter dependencies (Sections C.4, C.6; Tables 36, 40). The follow-up audit section adds QR-LoRA-style low-rank baselines and benign relearning stress tests, clarifying that quantization-stable target forgetting, semantic-neighbor preservation, and recurrence resistance are separable evaluation axes (Section C.3). The technical stability of the framework is then established through threshold sensitivity analyses and loss-landscape sharpness diagnostics, complemented by validation on hardware NF4 (Sections C.12, D.16, C.17). To isolate the underlying mechanisms, we conduct AR transplant controls, SLUG-layer ablations, and sequential-forgetting persistence tests (Sections C.5, C.16). Furthermore, we include external sanity checks using the MUSE and WMDP benchmarks alongside expanded domain-transfer diagnostics (Sections B.5, D.17). Finally, we provide detailed audits of baseline failure modes, computational overhead accounting, and formal mathematical addenda regarding Fisher/Wronskian analysis and reproducibility (Sections D.4, D.3, D.2, E, D.8, B). The depth probing and embedding analyses arguably represent the most significant mechanistic

findings. The 100% conceptual suppression across reverse association, multistep reasoning, and in-context extraction—combined with marked target-specific divergence in hidden representations and resilience to adversarial CoT—shows that Q-ROU goes beyond surface-level probability masking in the tested audit suite. Coupled with its computational efficiency and quantization resilience, these properties establish Q-ROU as a viable framework for bounded post-deployment knowledge management while leaving robustness to arbitrary retraining and stronger reconstruction attacks as open problems.