FORGETTING THAT STICKS: QUANTIZATION-PERMANENT UNLEARNING VIA CIRCUIT ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard unlearning evaluations measure behavioral suppression in full precision, immediately after training, despite quantization being a common deployment transformation for language models. Prior works have shown that 4-bit post-training quantization can reverse machine unlearning. We show this is a systematic dual failure. Gradient-based methods that achieve meaningful forgetting lose it under compression, while methods that survive quantization barely change the model. Both failures have the same root cause. Across all baselines, per-parameter RMS updates are 47 to 828$\times$ below the canonical NF4-calibrated floor used in our implementation, so updates diffused across billions of parameters cannot clear quantization bin boundaries, a consequence we formalize as a sparsity-permanence tradeoff. We present **MANSU** (**M**echanistic-**A**ligned **N**ull-**S**pace **U**nlearning), which addresses this tradeoff by treating it as an update-budget allocation problem. Causal circuit attribution isolates a compact, causally implicated forget circuit, a diagonal-Fisher mask restricts updates to retain-safe coordinates within that circuit, and a quantizer-calibrated magnitude floor, applied post-hoc, makes the resulting update deployment-visible, which we verify directly on the trained checkpoints. Matched controls (the same floor applied to a random circuit, an inverse circuit, a globally projected update, and a Fisher-ratio selective baseline) show that neither circuit concentration nor magnitude flooring alone reproduces this behavior in the reported ablations. We additionally introduce **Circuit Attribution Divergence (CAD)**, a mechanistic verification metric providing evidence of circuit-level structural change beyond behavioral suppression, corroborated by independent linear-probe and activation-patching analyses. Across multiple model families and WMDP hazard domains, MANSU consistently yields a non-positive NF4 PTQ gap while retaining substantially higher general capability than aggressively optimized baselines, whereas gradient ascent, for instance, partially reverses under compression (BF16 forget 0.260 to NF4 0.310, PTQ gap $+0.050$). We scope our boundary-crossing analysis to NF4 and report GPTQ and INT8 results separately as empirical transfer checks.

1 INTRODUCTION

A model can pass an unlearning check in full precision (BF16) yet recover the targeted knowledge after deployment with 4-bit post-training quantization (NF4; separately, GPTQ and INT8). On our WMDP-bio setting, several conventional unlearning methods exhibit positive NF4 PTQ gaps: forgetting weakens after quantization.

Prior work established the failure. The design question is still open. Zhang et al. (2025) first documented 4-bit PTQ reversing unlearning (up to 83% recovery) and proposed a saliency-based mitigation (PTQ-LR/SURE); concurrent work proposes quantization-aware objectives (Mishra & Mehreen, 2026; Khan & Sohel, 2026). None localizes causally, although factual knowledge is known to live in sparse subgraphs (Meng et al., 2022; Hanna et al., 2024; Kasliwal et al., 2026), and Lee et al. (2025) and Guo et al. (2025) caution that localized parameters need not be the ones unlearning must change. *If selective or larger updates are necessary for quantization-robust forgetting, where should the update budget be spent if retained capability must also survive?*

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

Table 1: **Prior work against three primary evaluation axes and a mechanistic diagnostic for deployment-robust unlearning** (Section 3). **F** = forget, **R** = retain, **Q** = NF4 quantization stability ($\Delta_{\text{PTQ}} \leq 0$), **S** = mechanistic evidence beyond behavioral suppression (CAD $\gg 0$). ✓ satisfies, ✗ fails, ∼ partial. † evaluated in our experiments. Axes are diagnostics, not a completeness checklist.

Method	Family	F	R	Q	S	Key limitation
GA† (Jang et al., 2023)	Grad. ascent	✓	∼	✗	✗	Updates diffuse over d params; per-param $\ll \delta_i$
Surgical GA† (Jang et al., 2023)	Grad. ascent	✓	∼	✗	✗	Layer restriction reduces diffusion but cannot reach δ_i
NPO† (Zhang et al., 2024)	Pref. opt.	∼	✓	✓	✗	Frozen reference prevents updates $\geq \delta_i$
SimNPO† (Fan et al., 2025)	Pref. opt.	✗	✓	✓	✗	Reference-free, but updates stay sub-threshold
GU+SimNPO† (Huang et al., 2024; Fan et al., 2025)	Null-space + Pref.	✗	✓	✓	✗	Global projection reinstates diffusion problem
LUNAR† (Shen et al., 2025)	Repr. steer.	∼	✓	✗	✗	Edits non-circuit MLP projection; PTQ gap $\approx +0.01$ (Table 2)
PTQ-LR (Zhang et al., 2025)	Quant.-aware	✓	∼	∼	✗	Raises LR; retain constraint re-bounds max useful LR
MANSU (ours)	Circuit + floor	✓	✓	✓	✓	Combines causal circuit restriction, retain-aware masking, and explicit NF4 boundary targeting

The mechanism. Under a retain constraint, an unlearning method has a limited update budget, which unrestricted gradient-based methods spread across most of the d parameters. For Llama-3.1-8B ($d \approx 8 \times 10^9$), even a large-norm gradient induces per-parameter changes of order 10^{-6}, far below the NF4 floor of $\approx 8.4 \times 10^{-4}$ for these layers (Definition 1). At compression time, changes at this scale rarely cross a quantization boundary and the forgetting is erased (Proposition 1). Methods that instead keep updates within a KL-anchored neighborhood of the original model (preference-optimization and null-space methods) avoid this only by barely changing the model (a mean forget-set reduction of 1.6 pp on a four-way MCQ task, Table 14).

Why naive fixes are not enough. Enlarging the update globally trades this away in the other direction. Global GA's MMLU collapses to 0.235 from a 0.603 zero-shot baseline (§6). Localizing to an arbitrarily chosen or gradient-saliency-selected region concentrates magnitude, but our random-circuit and inverse-circuit controls (Ablation C, §6) show identity matters, not just size. A magnitude floor alone is not sufficient either. Applying the same floor to a Fisher-ratio selective baseline (Ablation E) or to an unrestricted global projection (Ablation D) fails retain or quantization survival in every domain we tested: all three ingredients are jointly required.

MANSU. We present MANSU (**M**echanistic-**A**ligned **N**ull-**S**pace **U**nlearning), which treats this as an update-budget allocation problem. (1) EAP-IG (Hanna et al., 2024) identifies a compact circuit $\mathcal{C}$ causally implicated in forget-set answers. (2) Gradient updates within $\mathcal{C}$ are masked along a retain-protective Fisher threshold (Theorem 1). (3) Every cumulative update below the NF4 bin size is rescaled to a quantizer-calibrated floor (Lemma 1). On Llama-3.1-8B-Instruct / WMDP-bio, MANSU achieves a PTQ gap of -0.040 while preserving MMLU within 0.030 of the zero-shot model. NF4 does not reverse the forgetting (Proposition 2). Results replicate on Qwen-3-8B and MUSE.

Contributions: (I) **A mechanistic account** of the sparsity-quantization tradeoff in deployment-robust unlearning (Proposition 1), documented across 94 non-MANSU evaluation configurations (84 WMDP cells in Table 14 plus 10 MUSE cells in Table 2) over three model families, three hazard domains, and two benchmarks. (II) **MANSU**, analyzed in Theorem 1 and Lemma 1 with full proofs in Appendix D. (III) **Evidence that the combination matters**: matched circuit and floor-only controls, independent linear-probe and activation-patching analyses, boundary-crossing audits, recovery attacks, and cross-quantizer tests (§6).

2 Background and Related Work

Table 1 summarizes prior work by family, and Appendix B extends this discussion.

Gradient ascent and variants (Jang et al., 2023; Liu et al., 2022) maximize forget-set loss. Effective in full precision, they distribute updates over all d parameters, pushing per-parameter magnitudes far below quantization bin widths. Surgical variants (Jang et al., 2023) reduce the active parameter count, but under the measured retain budget remain below the deployment floor in our experiments (Table 20).

Preference optimization (NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025)) adapts DPO (Rafailov et al., 2023) to treat forget-set responses as dis-preferred. NPO's frozen reference model, and SimNPO's reference-free length-normalized objective, both keep per-parameter updates small, giving good retain scores but negligible structural change.

Null-space projection (GU, Huang et al., 2024) projects gradient updates onto the null space of the retain Hessian, giving a formal retain-safety bound, but its global scope reinstates the diffusion problem. MANSU inherits the projection idea and proves a strictly tighter bound by restricting both the update and the projection to the causally identified circuit (Theorem 1).

Selective Fisher-based dampening (SSD, Foster et al., 2024) selects and dampens parameters by Fisher-ratio saliency, without a causal-circuit or quantization constraint. MANSU instead selects by causal attribution, using Fisher only as a retain-protective mask threshold within the circuit. Ablation E (Section 6) shows that applying our floor to SSD's selection does not reproduce MANSU's combined permanence and retain behavior.

Representation steering (LUNAR (Shen et al., 2025), RMU (Li et al., 2024)) suppresses forget-set outputs by redirecting activations at inference time, leaving the causal knowledge circuit intact while passing behavioural metrics (CAD ≈ 0, Section 4.1). We evaluate LUNAR and discuss RMU as an adjacent baseline.

Quantization robustness: Zhang et al. (2025) (Section 1) mitigate the PTQ reversal they document using a saliency-based unlearning strategy with a large learning rate ("PTQ-LR" in Table 1). We show (Proposition 1) that the retain constraint independently caps the useful learning rate, so the root cause remains unaddressed. Two concurrent 2026 methods, Mishra & Mehreen (2026) (QUAIL) and Khan & Sohel (2026) (DurableUn), corroborate the dual-failure diagnosis without causal localization. MANSU instead combines causal localization with the floor, rather than a uniform penalty across all parameters.

Mechanistic interpretability and knowledge localization: ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) established via causal patching that factual associations are stored in middle MLP layers. EAP-IG (Hanna et al., 2024) extends this to circuit-level attribution. Concurrently, Kasliwal et al. (2026) apply circuit-restricted weight arithmetic to embed refusal into checkpoints. We apply the same principle to unlearning, adding the orthogonal constraint of quantization permanence. Following Lee et al. (2025) and Guo et al. (2025), we do not assume localization alone determines effective unlearning. Instead, we use it to allocate a retain-constrained update budget, and test separately whether circuit *identity* matters using matched random and inverse controls (Ablation C).

Gap. Neither the quantization-aware objectives nor the localization critiques above say *where* to spend a retain-constrained update budget so that it survives deployment quantization, the question our mechanistic diagnosis (§3) and MANSU (§4) address.

3 PROBLEM FORMULATION

Let $\theta \in \mathbb{R}^d$ be a pretrained LM's parameters, $\mathcal{D}_f$ the forget set, $\mathcal{D}_r$ the retain set.

A deployment-robust unlearning update $\Delta\theta$, giving $\theta' = \theta + \Delta\theta$, is evaluated along three primary axes, **forget** (θ' fails on $\mathcal{D}_f$ by a meaningful margin), **retain** (performance on $\mathcal{D}_r$ and general benchmarks substantially preserved relative to θ), and **quantization permanence** ($Q_4(\theta')$ also fails on $\mathcal{D}_f$, where Q_4 is the deployment 4-bit quantizer). A fourth diagnostic is **mechanistic change** (re-running causal attribution on θ' shows measurable change in the subgraph implementing forget-set knowledge, distinct from mere behavioral bypass). Standard behavioral evaluation measures only forget and retain, and §6 shows the other two are not free once measured.

Definition 1 (NF4 quantization floor)**.** *Under NF4 quantization (Dettmers et al., 2023), let $s_{b(i)}$ denote the scale of the NF4 quantization block containing parameter i. The minimum spacing between*

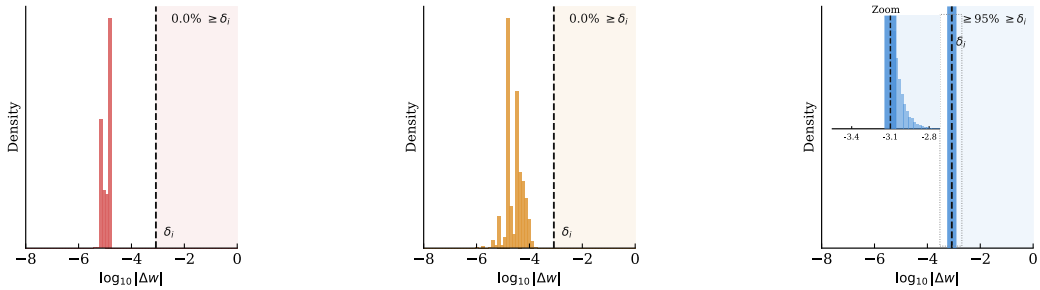


Figure 1: **Per-parameter update magnitudes** (Llama-3.1-8B / WMDP-bio). Histograms of $\log_{10}|\Delta w|$. **(a)** Global GA: diffuse, far below δ_i. **(b)** Surgical GA: concentrated on L14–16, still below δ_i. **(c)** MANSU: updates pushed to the canonical quantizer-calibrated floor (Definition 1). Dashed line: δ_i. Updates to its left rarely cross an NF4 boundary, so only MANSU's update survives.

adjacent normalized NF4 levels is 0.0796*, so the minimum adjacent-level spacing for parameter* i *is* $\delta_i^{\mathrm{NF4}} = s_{b(i)} \cdot 0.0796$. MANSU*'s canonical floor rescales this by a margin* α*:*

$$\delta_i \;=\; \alpha \cdot \delta_i^{\mathrm{NF4}}, \qquad \alpha = 0.704, \tag{1}$$

so $\delta_i \approx 8.4 \times 10^{-4}$ *for Llama-3.1-8B MLP weights (*$\delta_i^{\mathrm{NF4}} \approx 1.19 \times 10^{-3}$*, Appendix E). Throughout the paper,* δ_i *denotes this canonical floor unless stated otherwise.*

Proposition 1 (Sparsity–permanence tradeoff)**.** *Under gradient ascent with retain constraint* $\mathcal{L}_r(\theta + \Delta\theta) - \mathcal{L}_r(\theta) \le \epsilon_r$*, approximated to second order as* $\frac{1}{2}\sum_{j\in\mathcal{C}}[\mathbf{F}_r]_{jj}\Delta\theta_j^2 \le \epsilon_r$*: for* any *allocation of the update across* $\mathcal{C}$*, each coordinate individually satisfies the worst-case bound* $\|\Delta\theta_i\| \le \sqrt{2\epsilon_r / F_{\min,\mathcal{C}}}$ *with* $F_{\min,\mathcal{C}} = \min_{j\in\mathcal{C}}[\mathbf{F}_r]_{jj}$*. Under the diffuse, approximately-equal-magnitude allocation that unconstrained gradient ascent empirically produces, the constraint is instead spent evenly across all* $|\mathcal{C}|$ *active coordinates, giving the sharper working estimate*

$$\|\Delta\theta_i\| \;\le\; \sqrt{\frac{2\epsilon_r}{|\mathcal{C}|\,\bar{F}_{\mathcal{C}}}}, \qquad \bar{F}_{\mathcal{C}} \;=\; \frac{1}{|\mathcal{C}|}\sum_{j\in\mathcal{C}}[\mathbf{F}_r]_{jj}, \tag{2}$$

where $[\mathbf{F}_r]_{jj} = \mathbb{E}_{(x,y)\sim\mathcal{D}_r}[(\partial \log p_\theta(y|x)/\partial\theta_j)^2]$ *is the empirical diagonal Fisher of the retain loss (Appendix D; the diagonal Fisher remains well-defined under rank-deficient* $\mathbf{H}_r$*, unlike the standard* $\sigma_{\min}$ *form). Equation* (2) *is the equal-allocation estimate used for the numerical illustration below and throughout the paper. For Llama-3.1-8B (*$d = 8.03\times10^9$*,* $\epsilon_r = 0.02$*,* $\bar{F}_{\mathcal{C}} \sim 10^0$*), the global case (*$|\mathcal{C}| = d$*) gives* $\|\Delta\theta_i\| \lesssim 2.2\times10^{-6}$*, roughly* $380\times$ *below* δ_i*. Updates reach* δ_i *only when* $|\mathcal{C}|/d \le 7\times10^{-6}$ *(fewer than* 0.001% *of parameters).*

Implications. First, no existing gradient-based method operates near this threshold. Surgical GA's 6.6% circuit and even MANSU's 3.2% both sit more than three orders of magnitude above it ($\approx 4500\times$ for MANSU, $\approx 9400\times$ for Surgical GA, cf. Table 2), so localization alone is insufficient and the magnitude floor (Section 4) is required in our canonical implementation. Second, Proposition 1 says nothing about *which* parameters to update. Arbitrary concentration damages retain performance, so the circuit must be chosen causally.

Second failure mode. Preference-optimization methods (NPO, SimNPO, GU+SimNPO) avoid the floor problem differently. The frozen-reference KL constrains updates to be so small that $|\Delta\theta_i| \ll \delta_i$ almost everywhere. At standard hyperparameters this leaves forget accuracy largely intact (no measurable forgetting), and pushing the methods harder diffuses the larger update across d parameters, replacing targeted forgetting with global utility damage (Section 6).

4 METHOD

MANSU uses circuit localization to concentrate a retain-constrained update budget onto parameters causally implicated in the forget behavior (Proposition 1). Phase 1 finds *where* to concentrate it, Phase 2 protects the retain-sensitive coordinates within that region, and Phase 3's explicit floor makes the concentrated budget cross the deployment quantizer's boundary (Figure 2, Algorithm 1, Appendix C).

Phase 1: Localize (Appendix F). EAP-IG (Hanna et al., 2024) attributes causal contribution to each edge of the transformer graph. Aggregating over 50 forget examples and ranking MLP sublayers by

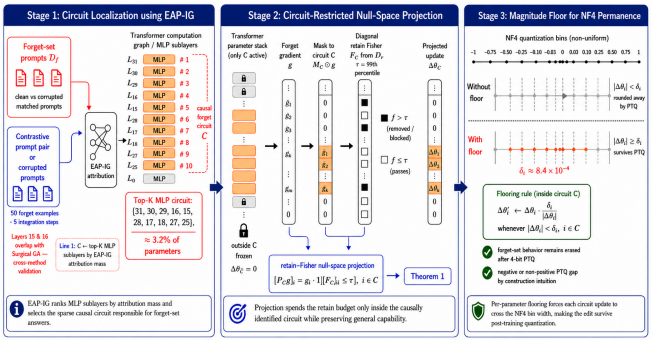


Figure 2: **MANSU three-phase pipeline: allocate, protect, make visible.** *Phase 1 (Localize)*: EAP-IG identifies a compact MLP circuit $\mathcal{C}$ causally implicated in forget-set answers. *Phase 2 (Project)*: updates restricted to $\mathcal{C}$ are masked along a retain-protective Fisher threshold (Theorem 1). *Phase 3 (Floor)*: a per-parameter magnitude floor δ_i rescales each update, with boundary crossing verified by checkpoint audit (Lemma 1, Appendix E).

total incoming attribution mass yields the top-10 attribution ranking:

$$\mathcal{C}_{\text{MLP}}^{(\text{top-10})} = \{30, 14, 31, 19, 29, 15, 20, 16, 21, 17\}. \tag{3}$$

The canonical circuit uses the top-5 layers $\mathcal{C}_{\text{MLP}} = \{30, 14, 31, 19, 29\}$, the $k = 5$ configuration of Tables 15 and 17. Its gradient-mask scope covers $\approx 11\%$ of model parameters (5 MLP layers, three projections each), while the Fisher-masked *effective* fraction is $\approx 3.2\%$ (per-stage breakdown in Appendix C.3). Layer 14 appears in both the EAP-IG circuit and surgical GA's L14–16 selection, and upper layers {29,30,31} dominate the attribution ranking, consistent with evidence that MLP modules mediate factual recall (Meng et al., 2022).

Phase 2: Project (Appendix C). Gradient updates within $\mathcal{C}$ are masked along high-Fisher coordinates, an approximation to projection into $\ker(\mathbf{H}_{\mathcal{CC}})$ under the diagonal-Fisher assumption (approximation error bounded in Proposition 3):

$$[P_{\mathcal{C}}\, g]_i = g_i \cdot \mathbb{1}[[\mathbf{F}_{\mathcal{C}}]_{ii} \leq \tau], \quad i \in \mathcal{C}; \qquad \Delta\theta_{\bar{\mathcal{C}}} = 0, \tag{4}$$

where $\tau = 0.1 \times \text{mean}([\mathbf{F}_{\mathcal{C}}]_{ii})$ is a retain-protective Fisher threshold. Restricting projection to $\mathcal{C}$ yields a tighter idealized retain bound than projecting globally (Theorem 1).

Phase 3: Floor (Appendix E). After training converges (best checkpoint selected by lowest forget accuracy among saved checkpoints), the magnitude floor is applied post-hoc to the saved checkpoint. For each $i \in \mathcal{C}$, the cumulative update $\Delta\theta_i = \theta_i - \theta_i^{(0)}$ is rescaled to the NF4-calibrated magnitude δ_i while preserving direction:

$$\Delta\theta_i \;\leftarrow\; \Delta\theta_i \cdot \frac{\delta_i}{|\Delta\theta_i|} \qquad \text{whenever } 0 < |\Delta\theta_i| < \delta_i,\ i \in \mathcal{C}, \tag{5}$$

using the canonical $\delta_i = 0.704\, \delta_i^{\text{NF4}}$ of Definition 1. At the canonical $\alpha = 0.704$, direct post-hoc measurement shows that 100% of circuit parameters receiving a nonzero cumulative update change their NF4 quantization assignment (Appendix E).

Training objective. MANSU encodes its three constraints jointly as

$$\min_{\substack{\Delta\theta_{\mathcal{C}} \in P_{\mathcal{C}}(\mathbb{R}^{|\mathcal{C}|}) \\ |\Delta\theta_i| \geq \delta_i\ \forall i \in \mathcal{C} \\ \Delta\theta_{\bar{\mathcal{C}}} = 0}} \; -\mathcal{L}_f(\theta + \Delta\theta) \;+\; \lambda\, D_{\text{KL}}\big(p_{\theta^{(0)}} \,\big\|\, p_{\theta+\Delta\theta}\big)_{x \sim \mathcal{D}_r}. \tag{6}$$

In implementation, the circuit restriction ($\Delta\theta_{\bar{\mathcal{C}}} = 0$) and the retain-aware mask ($\Delta\theta_{\mathcal{C}} \in P_{\mathcal{C}}$) are enforced during training, while the floor ($|\Delta\theta_i| \geq \delta_i$) is applied post-hoc (Phase 3), since a sub-threshold update receives no gradient signal toward boundary crossing. Eq. (6) is thus a conceptual formulation of the three design requirements, not the literal problem the implementation solves: post-hoc rescaling is not guaranteed to coincide with the joint constrained minimizer. The frozen-reference KL (following NPO/GU) prevents retain collapse. Hyperparameters and the rationale for full-parameter (not LoRA) training are in Appendix C.

4.1 CIRCUIT ATTRIBUTION DIVERGENCE (CAD)

Motivation. Two unlearned checkpoints with identical forget-set accuracy can differ in mechanism. In θ'_A the knowledge circuit has been dismantled. In θ'_B the circuit is intact and a downstream layer redirects its output to a refusal token (LUNAR-style). Both pass behavioral evaluations, but θ'_B is fragile to small fine-tunes, re-prompts, or quantization. Behavioral metrics measure outputs, while unlearning is a claim about weights.

Definition. Let $E(\mathcal{C})$ be the EAP-IG edge set on the original θ with attribution score $s_e(\theta)$ for edge e (Appendix F). Re-run EAP-IG on the unlearned θ' and compare:

$$\mathrm{CAD}(\mathcal{C}, \mathcal{D}_f;\, \theta, \theta') \;=\; \frac{\sum_{e \in E(\mathcal{C})} |s_e(\theta) - s_e(\theta')|}{\sum_{e \in E(\mathcal{C})} |s_e(\theta)|}. \tag{7}$$

$\mathrm{CAD} \to 0$ means the circuit is intact (behavior may have changed only via downstream redirection). $\mathrm{CAD} \approx 1$ means its attribution has been reorganized. Values > 1 indicate sign-flipped redirection (also structural).

Properties. CAD is (i) computed entirely on the unlearned weights with no held-out probes, (ii) ≈ 0 by construction for representation-redirection methods (LUNAR/RMU), (iii) insensitive to spurious behavioral suppression (a refuse-everything model yields $\mathrm{CAD} \approx 0$), and (iv) not satisfied by random weight perturbation (the random-circuit control, Ablation C(i), collapses CAD by $\sim 35\%$). CAD alone does not certify knowledge deletion. High CAD with elevated AS-NC indicates broad representational damage rather than localized attribution change. The joint diagnostic is high CAD *and* low AS-NC (worked SimNPO/MUSE example in Appendix R).

Circularity and independent validation. Because CAD re-runs EAP-IG, the attribution family used to identify $\mathcal{C}$, a high score could reflect attribution instability rather than structural change. We validated the finding with linear probes and activation patching, two methods sharing no machinery with EAP-IG. Both show substantial, localized change at circuit layers against near-zero change at matched non-circuit controls (Appendix R). Re-attribution on θ' retains most but not all of the original top-5 layers, evidence of *structural reorganization* rather than clean dismantling, which does not affect any reported CAD value but means CAD does not certify outright dismantling.

Companion metrics: AS-C, AS-NC. Activation-level checks inside / outside $\mathcal{C}$ (Eq. 12). Evidence of mechanistic change requires high CAD *and* the concentration gap AS-C $\ll$ CAD, which is present only for localized methods. For global baselines AS-C = CAD numerically (Table 21).

Algorithm 1 MANSU *(Mechanistic-Aligned Null-Space Unlearning)*

Require: pretrained $\theta^{(0)}$, forget set $\mathcal{D}_f$, retain set $\mathcal{D}_r$, circuit size K, KL weight λ, floor δ_i

$\mathcal{C} \leftarrow$ top-K MLP sublayers by EAP-IG attribution mass on $\mathcal{D}_f$ ▷ Phase 1: Localize
$[F_{\mathcal{C}}]_{ii} \leftarrow \mathbb{E}_{\mathcal{D}_r}\left[(\partial \log p_{\theta^{(0)}} / \partial \theta_{\mathcal{C},i})^2\right]$; $\tau \leftarrow 0.1 \times \mathrm{mean}([F_{\mathcal{C}}]_{ii})$
$\theta \leftarrow \theta^{(0)}$
for $t = 1, \ldots, T$ **do** ▷ Phase 2: training loop
  $g \leftarrow -\nabla_{\mathcal{C}} \mathcal{L}_f(\theta) + \lambda \nabla_{\mathcal{C}} D_{\mathrm{KL}}(p_{\theta^{(0)}} \,\|\, p_\theta)_{\mathcal{D}_r}$
  $\hat{g}_i \leftarrow g_i \cdot \mathbb{1}[[F_{\mathcal{C}}]_{ii} \le \tau]$ ▷ Phase 2: project (Fisher mask)
  $\theta_{\mathcal{C}} \leftarrow \theta_{\mathcal{C}} - \eta\, \hat{g}$; $\theta_{\bar{\mathcal{C}}}$ frozen
end for
for $i \in \mathcal{C}$ with $0 < |\theta_i - \theta_i^{(0)}| < \delta_i$ **do** ▷ Phase 3: floor (post-hoc)
  $\theta_i \leftarrow \theta_i^{(0)} + \delta_i \cdot \mathrm{sign}(\theta_i - \theta_i^{(0)})$
end for
return θ

5 THEORETICAL ANALYSIS

MANSU is motivated by three idealized results, retain safety (Theorem 1), quantization permanence (Lemma 1), and amplification of floor-crossing updates (Proposition 2), each under the idealized circuit-restricted projector ($\Delta\theta_{\bar{\mathcal{C}}} = 0$). None is a guarantee for the implemented algorithm. The gap between this idealized projector and the diagonal-Fisher mask MANSU actually implements is a separate, fourth quantity (an approximation error, not a guarantee), which we characterize and

disclose rather than bound away (Proposition 3, Appendix D.3). Full proofs and error bounds are in Appendix D.

Theorem 1 (Idealized circuit-restricted retain bound)**.** *Let $\mathcal{L}_r$ be twice continuously differentiable with PSD Hessian $\mathbf{H}$. For $\mathcal{C} \subseteq [d]$, $\bar{\mathcal{C}} = [d] \setminus \mathcal{C}$, and any $\Delta\theta$ with $\Delta\theta_{\mathcal{C}} \in \ker(\mathbf{H}_{\mathcal{C}\mathcal{C}})$, $\Delta\theta_{\bar{\mathcal{C}}} = 0$, $\|\Delta\theta\| \leq \varepsilon$:*

$$\mathcal{L}_r(\theta + \Delta\theta) - \mathcal{L}_r(\theta) \;\leq\; \underbrace{\|\nabla_{\mathcal{C}} \mathcal{L}_r(\theta)\|}_{\leq \|\nabla \mathcal{L}_r(\theta)\|} \varepsilon \;+\; \frac{\varepsilon^2}{2} \underbrace{\sigma_{\max}(\mathbf{H}_{\bar{\mathcal{C}}\bar{\mathcal{C}}})}_{\leq \sigma_{\max}(\mathbf{H})} + O(\varepsilon^3). \tag{8}$$

Each bracketed term is at most its global counterpart (sub-vector L2 bound and Cauchy interlace (Horn & Johnson, 2012)). The circuit-restricted bound is strictly tighter than global null-space projection (Huang et al., 2024) whenever $\mathbf{H}$'s dominant eigenvector projects non-trivially onto $\mathcal{C}$-coordinates. Since $\mathcal{C}$ is chosen by causal attribution on $\mathcal{D}_f$ (not $\mathcal{D}_r$), this holds generically. Ablation D verifies it empirically. Theorem 1 requires circuit restriction ($\Delta\theta_{\bar{\mathcal{C}}} = 0$) and $\Delta\theta_{\mathcal{C}} \in \ker(\mathbf{H}_{\mathcal{C}\mathcal{C}})$. The diagonal-Fisher mask used in Phase 2 is a practical approximation whose quality we characterize in Appendix D.3 (Proposition 3) and find limited. Retain preservation in practice is supported empirically and by the frozen-reference KL anchor in Eq. 6, not certified by the diagonal-Fisher approximation.

Lemma 1 (Quantization survival)**.** *Let Q_4 be 4-bit quantization with monotone levels $\{q_k\}$ and let w_i be the bin width at θ_i. Any update $|\Delta\theta_i| \geq w_i$ changes the quantized value: $Q_4(\theta_i + \Delta\theta_i) \neq Q_4(\theta_i)$. δ_i^{NF4} (Definition 1 with $\alpha = 1$) is the minimum spacing between adjacent NF4 levels, scaled by the block scale. Because quantization cells extend to the midpoints between levels, this spacing is a lower bound on w_i rather than an upper bound, and the block scale is recomputed after the update, so δ_i^{NF4} is the scale at which crossing becomes possible, not a worst-case guarantee.* MANSU*'s canonical operating point uses $\alpha = 0.704$, and boundary crossing is verified by direct measurement (Appendix E).*

Proposition 2 (NF4 amplifies floor-crossing updates)**.** *Let θ_i lie in a narrow-bin region of the NF4 grid (near zero; see Appendix E, Table 5) and let $|\Delta\theta_i| \geq \delta_i$. For updates that cross $m \geq 2$ bin boundaries (where $|q_{k+m} - q_k| \geq m\delta_i$), $|Q_4(\theta_i + \Delta\theta_i) - \theta_i| \geq |\Delta\theta_i|$: quantization amplifies displacement rather than attenuating it, producing a negative PTQ gap. For single-crossing updates ($m = 1$) deposited at the bin boundary by the floor, the amplification holds in expectation rather than with high probability. Conversely, for diffuse methods with $|\Delta\theta_i| < \delta_i$, crossing is not guaranteed by the worst-case bound (a weight already close to its own boundary can still cross with a small update); empirically this regime is where forgetting is silently erased by NF4, the $+0.06$ to $+0.07$ PTQ gap observed for these methods.*

6 EXPERIMENTS

Setup, hyperparameters, timing, update statistics, and extended ablations are in Appendices G to R.

Setup: Llama-3.1-8B-Instruct on WMDP-bio (Li et al., 2024) for the main table (Table 2). MANSU is also evaluated on MUSE (Shi et al., 2025) (Harry Potter open-ended memorization) and Qwen-3-8B. Fixed forget and MMLU indices are reused across methods. NF4 evaluation via `bitsandbytes` (4-bit, double-quantization off).

Main results: All findings read off the WMDP-bio Llama-3.1-8B block of Table 2 (zero-shot 0.763). The per-property scorecard in Figure 3 (Appendix R) summarises pass/fail across all 24 weight-edit (method, dataset) cells (six weight-edit methods on four datasets for each model, with LUNAR excluded as representation redirection). *Gradient ascent fails quantization*: Global GA's BF16 forget 0.260 flips to NF4 0.310 (Δ_{PTQ}=+0.050) with MMLU collapsing to 0.235 (indiscriminate damage, Figure 1). *Aggressive preference optimization survives quantization but destroys utility*: SimNPO/GU+SimNPO reach forget 0.250/0.230 with Δ_{PTQ}=0.000 but MMLU 0.295/0.200. NPO preserves MMLU (0.563) and reaches comparable raw forget depth (0.443 vs MANSU's 0.430), but its PTQ gap (-0.020) is half the magnitude of MANSU's (-0.040), so NPO's forgetting is markedly less robust to quantization. MANSU *satisfies all three properties*: forget 0.430, NF4 0.390, Δ_{PTQ}=-0.040, MMLU 0.573 (within 0.030 of zero-shot), IFEval 0.551. NF4 does not reverse the forgetting (Proposition 2). Structural metrics indicate weight-level mechanistic change rather than

Table 2: **Behavioral results.** Forget set, retain split, and general capability. Columns: **BF16 (↓)** = forget-set accuracy in full precision; **NF4 (↓)** = same forget set re-evaluated after 4-bit NF4 post-training quantization; $\Delta_{\mathrm{PTQ}} = \mathrm{acc}_{\mathrm{NF4}} - \mathrm{acc}_{\mathrm{BF16}}$ (negative values mean forget-set accuracy fell further under NF4, Lemma 1, Proposition 2); **Rt/Util (↑)** = WMDP retain-split accuracy (WMDP rows) or utility score (MUSE); **MMLU$^{\mathrm{sel}}$/IFEval** = model-level capability and instruction following, dataset-independent. MMLU also served for checkpoint selection (Appendix G), so IFEval is held out. Structural metrics (CAD, AS-C, AS-NC) are reported in Table 21 (Appendix R). – entries are not applicable. The unmodified model is not quantized as part of unlearning evaluation, and MUSE utility is only defined post-unlearning. **Statistical reliability (Appendix H):** MANSU's Δ_{PTQ} is negative on all 8 cells (sign test $p = 0.0039$). The depth advantage over NPO is not claimed. The joint claim is comparable depth together with non-positive Δ_{PTQ} and preserved MMLU.

	Llama-3.1-8B-Instruct						Qwen-3-8B					
Method	BF16↓	NF4↓	Δ_{PTQ}↓	Rt/Util↑	MMLU$^{\mathrm{sel}}$↑	IFEval↑	BF16↓	NF4↓	Δ_{PTQ}↓	Rt/Util↑	MMLU$^{\mathrm{sel}}$↑	IFEval↑
WMDP-bio												
Zero-shot	0.763	–	–	0.763	0.603	0.560	0.803	–	–	0.803	0.741	0.548
Global GA	0.260	0.310	+0.050	0.260	0.235	0.536	0.233	0.233	+0.000	0.247	0.242	0.408
Surgical GA	0.547	0.573	+0.027	0.560	0.483	0.528	0.260	0.247	-0.013	0.303	0.458	0.428
NPO	0.443	0.423	-0.020	0.503	0.563	0.528	0.283	0.320	+0.037	0.303	0.492	0.416
SimNPO	0.250	0.250	+0.000	0.210	0.295	0.528	0.227	0.227	+0.000	0.257	0.265	0.412
GU+SimNPO	0.230	0.230	+0.000	0.247	0.200	0.532	0.267	0.263	-0.003	0.277	0.568	0.420
LUNAR	0.621	0.638	+0.017	0.619	0.571	0.544	0.658	0.671	+0.013	0.655	0.612	0.531
MANSU (ours)	**0.430**	**0.390**	**-0.040**	0.523	**0.573**	0.551	0.617	0.581	**-0.036**	0.671	**0.729**	0.541
WMDP-chem												
Zero-shot	0.533	–	–	0.533	0.603	0.560	0.560	–	–	0.560	0.741	0.548
Global GA	0.493	0.473	-0.020	0.491	0.550	0.540	0.237	0.293	+0.057	0.269	0.365	0.424
Surgical GA	0.427	0.423	-0.003	0.426	0.525	0.548	0.313	0.317	+0.003	0.398	0.557	0.436
NPO	0.253	0.227	-0.027	0.269	0.538	0.532	0.237	0.237	+0.000	0.296	0.515	0.432
SimNPO	0.233	0.233	+0.000	0.241	0.195	0.552	0.240	0.277	+0.037	0.259	0.405	0.412
GU+SimNPO	0.273	0.273	+0.000	0.231	0.230	0.536	0.233	0.233	+0.000	0.250	0.328	0.424
LUNAR	0.481	0.497	+0.016	0.479	0.571	0.544	0.521	0.534	+0.013	0.518	0.612	0.531
MANSU (ours)	**0.333**	**0.307**	**-0.027**	0.398	**0.584**	0.549	0.307	0.274	**-0.033**	0.364	**0.714**	0.539
WMDP-cyber												
Zero-shot	0.477	–	–	0.477	0.603	0.560	0.537	–	–	0.537	0.741	0.548
Global GA	0.357	0.370	+0.013	0.390	0.543	0.532	0.360	0.377	+0.017	0.370	0.463	0.416
Surgical GA	0.283	0.290	+0.007	0.367	0.480	0.536	0.430	0.467	+0.037	0.473	0.710	0.428
NPO	0.340	0.343	+0.003	0.400	0.568	0.544	0.360	0.393	+0.033	0.437	0.715	0.432
SimNPO	0.270	0.277	+0.007	0.273	0.195	0.532	0.270	0.403	+0.133	0.280	0.555	0.452
GU+SimNPO	0.300	0.300	+0.000	0.230	0.297	0.540	0.333	0.443	+0.110	0.267	0.715	0.424
LUNAR	0.431	0.445	+0.014	0.428	0.571	0.544	0.501	0.514	+0.013	0.498	0.612	0.531
MANSU (ours)	**0.323**	**0.313**	**-0.010**	0.391	**0.586**	0.549	0.497	0.464	**-0.033**	0.541	**0.721**	0.542
MUSE												
Zero-shot	0.365	–	–	–	0.603	0.560	0.024	–	–	–	0.741	0.548
Global GA	0.000	0.000	+0.000	0.000	0.570	0.484	0.000	0.009	+0.009	0.007	0.700	0.452
Surgical GA	0.000	0.000	+0.000	0.000	0.583	0.552	0.020	0.017	-0.003	0.020	0.723	0.424
NPO	0.013	0.016	+0.002	0.013	0.553	0.544	0.011	0.021	+0.009	0.011	0.728	0.428
SimNPO	0.000	0.000	+0.000	0.000	0.539	0.540	0.001	0.013	+0.012	0.000	0.723	0.448
GU+SimNPO	0.000	0.000	+0.000	0.000	0.575	0.464	0.013	0.017	+0.004	0.005	0.718	0.432
LUNAR	0.187	0.198	+0.011	0.184	0.571	0.544	0.162	0.171	+0.009	0.159	0.612	0.531
MANSU (ours)	**0.005**	**0.003**	**-0.002**	0.006	**0.591**	0.547	0.021	0.017	**-0.004**	0.019	**0.737**	0.436

behavioral suppression alone. MANSU attains the highest CAD (1.143) with low AS-NC spillover (0.138). LUNAR yields CAD $\in [0.029, 0.045]$ across all WMDP/MUSE cells (Table 21), consistent with editing weights outside the EAP-IG forget circuit.

Cross-dataset / cross-architecture consistency. MANSU's Δ_{PTQ} is non-positive on all $8/8$ (model, dataset) cells of Table 2 (sign test $p = 0.0039$, Appendix H). MMLU stays within 0.030 of zero-shot across cells. CAD exceeds 1.0 on $7/8$ cells (Table 21). The pattern extends beyond the two 8B models. Tables 11, 12, and 13 report MANSU on six additional model variants (Gemma-2B/3-1B/3-4B, Llama-3.2-3B, Qwen-2.5-4B/3-4B), and Table 14 the family-wise macro-averages. MANSU achieves strictly negative Δ_{PTQ} on every cell of every sweep family. By contrast, no baseline beats MANSU on all three of forget, quantization-permanence, and utility on any cell. The dual failure mode holds across all eight cells.

Ablations (Table 3): *A, no floor*: Δ_{PTQ} weakens from -0.040 to -0.008 and forget accuracy regresses to 0.513, isolating the floor as the mechanism turning circuit concentration into quantization permanence. *B, no null-space projection*: forget accuracy regresses to 0.451 and MMLU drops to 0.449 (largest utility hit of any row), confirming projection sharpens the forget–retain tradeoff and is the primary retain-protector (Theorem 1). *C(i), random circuit (seed 42)*: same $|\mathcal{C}|$, but forget accuracy regresses to 0.500 and CAD collapses from 1.143 to 0.743 (-35%). AS-NC nearly triples

Table 3: **Component ablation** on Llama-3.1-8B-Instruct / WMDP-bio (zero-shot 0.763). Each row removes or replaces one MANSU component, all others held fixed. **Bold** marks MANSU (full) rather than the per-column optimum (E(ii) reaches a lower raw BF16 than MANSU but fails Δ_{PTQ} and retain, per the discussion below). Row D uses GU-global (no SimNPO), a weaker forget baseline than GU+SimNPO in Table 2. Its higher BF16 is expected. Rows E replace causal selection with SSD Fisher-ratio selection (Foster et al., 2024) and were not scored on CAD/AS-NC.

Configuration	BF16↓	Δ_{PTQ}↓	Rt↑	MMLU↑	CAD↑	AS-NC↓
MANSU (full)	**0.430**	**-0.040**	0.523	**0.573**	**1.143**	**0.138**
A: w/o magnitude floor	0.513	-0.008	0.489	0.471	1.091	0.194
B: w/o null-space proj.	0.451	-0.019	0.461	0.449	1.063	0.201
C(i): random circuit (seed 42)	0.500	-0.024	0.471	0.470	0.743	0.394
C(ii): inverse circuit (bottom-k)	0.551	+0.028	0.441	0.451	0.511	0.481
D: GU-global + floor	0.697	+0.013	0.448	0.443	0.672	0.441
E(i): SSD alone (Foster et al., 2024)	0.450	0.000	0.510	0.478	—	—
E(ii): SSD + floor	0.340	+0.030	0.410	0.440	—	—

($0.138 \rightarrow 0.394$), indicating diffuse rather than localized intervention. Forget depth and quality require the causally identified circuit rather than an arbitrary same-size subset, consistent with the circuit-size findings of Lee et al. (2025) and Guo et al. (2025) on the factual-recall benchmarks studied here. *C(ii), inverse circuit (bottom-k), the strongest negative control*: Δ_{PTQ} flips to $+0.028$, forget regresses to 0.551, and CAD bottoms out at 0.511 across all three domains, ruling out any beneficial effect from non-causal parameters. *D, global null-space + floor (GU-global)*: Δ_{PTQ} flips positive ($+0.013$) despite the floor, because diffuse global updates contain mixed-sign components that cancel below the bin floor under NF4 rounding. These controls provide evidence that circuit identity contributes to quantization-robust forgetting alongside the floor. *E, SSD-style Fisher-ratio selection (Foster et al., 2024)*: this addresses a tautology concern that the floor alone, applied to any sufficiently small parameter subset, might reproduce MANSU's permanence. It does not. Allocation determines permanence. SSD alone forgets less than MANSU (0.450 vs 0.430) with a zero PTQ gap (0.000). Adding our floor (E(ii)) flips the gap positive ($+0.030$) while retain accuracy falls from 0.510 to 0.410. On chem/cyber, SSD+floor's Δ_{PTQ} is $+0.030/-0.050$ against MANSU's $-0.027/-0.010$, more negative on cyber but with MMLU trailing by 0.02 to 0.14 (Appendix R). GPTQ/INT8 boundary transfer checks, outside the NF4 analysis, are in Appendix I. *Cross-domain consistency.* The pattern reproduces on WMDP-chem and WMDP-cyber. Row A's Δ_{PTQ} flips to $+0.004/+0.003$. Row C(i)'s CAD collapses to $0.711/0.729$ (vs $1.097/1.118$). Row D's Δ_{PTQ} stays positive ($+0.009/+0.007$).

7 DISCUSSION

Implications for evaluation practice. The 94 non-MANSU evaluation configurations show that the standard protocol selects for methods that make minimal parameter changes. A method that reduces forget-set accuracy by 1.6 percentage points passes this protocol without removing the knowledge. We propose adding the PTQ gap and CAD (or an equivalent mechanistic verification). Neither requires new infrastructure.

Limitations. MANSU is reported on Llama-3.1-8B and Qwen-3-8B in the main table and on six smaller models in Appendix N. Localization is well-supported on factual-recall benchmarks (Meng et al., 2022). Behaviour beyond 8B ($|\Delta\theta_i| \propto 1/\sqrt{d}$ at fixed circuit fraction, Eq. (2)) is consistent with the floor remaining binding but is not directly verified. EAP-IG attribution examples and forget-evaluation indices are not confirmed disjoint.

8 CONCLUSION

Across the evaluated baselines, passing unlearning in full precision does not imply passing after the compression step preceding every real-world deployment: forgotten knowledge returns, or the model barely changed to begin with. MANSU addresses this by updating only the causally identified circuit, projecting away from retain-sensitive directions, and rescaling every update to clear the NF4 floor, producing forgetting that the compression step does not reverse.

AI USE STATEMENT

In this work, we used generative AI tools (Claude, Anthropic) to help interpret experimental results, including running a small number of confirmatory checks on already-trained checkpoints; these checks confirm existing claims and are not used as primary evidence for any result. For execution, we took the help of an AI coding agent for these confirmatory checks and for code and manuscript editing. We also used generative AI tools to search and summarize related literature, create and modify figures, edit code, draft and edit parts of the paper for readability, format tables, and check that numbers stated in the text match the result tables. The authors designed and ran all primary experiments reported in this work. We reviewed all AI-assisted work and verified AI-assisted edits and interpretations against the underlying experimental records. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to make the removal of hazardous knowledge (biosecurity, chemical and cyber threats) hold up under deployment quantization. Our finding that 4-bit compression can undo existing unlearning methods is information that practitioners relying on unlearning for safety certification need. The work does not make hazardous knowledge easier to acquire. We show that existing methods are weaker than they appear, but we provide no tool for recovering knowledge from unlearned models, and we use EAP-IG, a published open-source method, only for removal. All experiments use public models and public benchmarks (WMDP, MUSE, MMLU). Appendix A discusses scope and misuse further.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix C give the full procedure, Algorithm 1 gives pseudocode, and Appendix F describes the EAP-IG circuit attribution. Appendix G describes the experimental setup and baselines, Appendix O lists every hyperparameter, and Appendix D gives full proofs of the theoretical results. Appendix H reports the sign test across all eight cells. All datasets (WMDP, MUSE, MMLU) are public, and EAP-IG is available at `github.com/hannamw/EAP-IG`. We will release our implementation, evaluation scripts and fixed evaluation indices upon acceptance.

REFERENCES

Minyeong Choe, Haehyun Cho, Changho Seo, and Hyunil Kim. Do all autoregressive transformers remember facts the same way? a cross-architecture analysis of recall mechanisms. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28494–28513, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1448. URL https://aclanthology.org/2025.emnlp-main.1448/.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001. URL https://doi.org/10.1137/0707001.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=OUIFPHEgJU.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *The*

Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025. URL https://openreview.net/forum?id=JbvSQm5h1l.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=92oBV5HAGl.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2403.17806. Introduces EAP-IG (Edge Attribution Patching with Integrated Gradients).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=dheDf5EpBT.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Aditya Kasliwal, Pratinav Seth, and Vinay Kumar Sankarapu. C-$\Delta\Theta$: Circuit-restricted weight arithmetic for selective refusal. *arXiv preprint arXiv:2602.04521*, 2026.

Abdullah Ahmad Khan and Ferdous Sohel. Durableun: Quantization-induced recovery attacks in machine unlearning. *arXiv preprint arXiv:2605.02196*, 2026.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. *Limitations of the empirical fisher approximation for natural gradient descent*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Hwiyeong Lee, Uiji Hwang, Hyelim Lim, and Taeuk Kim. Does localization inform unlearning? a rigorous examination of local parameter attribution for knowledge unlearning in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21857–21869, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1109. URL https://aclanthology.org/2025.emnlp-main.1109/.

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, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xlr6AUDuJz.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *GetMobile: Mobile Comp. and Comm.*, 28(4):12–17, January 2025. ISSN 2375-0529. doi: 10.1145/3714983.3714987. URL https://doi.org/10.1145/3714983.3714987.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.), *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 243–254. PMLR, 22–24 Aug 2022. URL https://proceedings.mlr.press/v199/liu22a.html.

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

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Himanshu Mishra and Kanwal Mehreen. Quail: Quantization aware unlearning for mitigating misinformation in llms. *arXiv preprint arXiv:2601.15538*, 2026.

Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

William F. Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. LLM unlearning via neural activation redirection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=teB4aqJsNP.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Haibo Hu, and Minxin Du. Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms. *CoRR*, abs/2505.16831, May 2025. URL https://doi.org/10.48550/arXiv.2505.16831.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lHSeDYamnz.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

A Broader Impact and Ethics

This work is motivated by safety objectives in AI deployment. Durably removing hazardous knowledge (biosecurity threats, cyberweapons, chemical weapons) is a safety-critical requirement as language models become more widely deployed. Our finding that existing unlearning methods fail under 4-bit compression is information that practitioners and policymakers relying on unlearning for safety certification need to know.

Misuse. This work does not make hazardous knowledge easier to acquire. We demonstrate that existing methods are weaker than they appear; we do not provide tools for recovering knowledge from unlearned models. EAP-IG is a published open-source method we use for removal, not insertion.

Scope. Experiments cover bio, chem, and cyber hazard domains on MCQ. Generalization to other hazard types or open-ended formats requires additional validation. We do not claim MANSU is a complete solution to knowledge removal; it is meaningfully better than existing alternatives on the metrics we define.

Reproducibility. All datasets (WMDP, MMLU) are public. EAP-IG is at `github.com/hannamw/EAP-IG`. Our implementation, evaluation scripts, and fixed evaluation indices will be released upon acceptance.

B Extended Related Work

Gradient ascent and variants. Gradient ascent (GA) (Jang et al., 2023) maximizes the loss on forget-set examples directly, reversing the effect of gradient descent on those examples. Gradient Difference (GD) (Liu et al., 2022) adds a retain-loss minimization term. KL-regularized GA (Maini et al., 2024) constrains the unlearned model to remain close to the original in output distribution on the retain set. These methods achieve real forgetting in full precision; their structural problem is not the loss function but the update distribution.

Preference optimization. NPO (Zhang et al., 2024) and SimNPO (Fan et al., 2025) treat forget-set completions as negative preferences in a DPO-style objective (Rafailov et al., 2023), training the model to assign lower probability to forget-set answers. NPO does so relative to a frozen reference, while SimNPO removes the reference model in favour of a length-normalized objective. In our experiments both produce predominantly sub-threshold updates. The same regime that prevents catastrophic collapse limits per-parameter updates well below δ_i by the same diffusion argument as for GA, so the floor is not crossed but neither is forget accuracy substantially altered. Across our 94-experiment sweep, mean forget-set reduction for capable models is 1.6 percentage points; the large majority of runs reduce forget accuracy by less than 5 percentage points.

Null-space projection. GU (Huang et al., 2024) projects gradient updates onto the null space of the retain-set Hessian, theoretically bounding retain perturbation. Combined with SimNPO it represents the strongest principled baseline. The structural difference from MANSU is scope: GU applies projection globally over d parameters, which forces the update to remain globally small and reinstates the diffusion problem; MANSU restricts updates to the causal circuit $\mathcal{C}$ and combines this with a KL retain anchor and a magnitude-floor constraint. Theorem 1 formalizes the retain bound under circuit-restricted updates.

Representation steering. LUNAR (Shen et al., 2025) redirects intermediate activations toward regions of activation space associated with model inability to answer, training only a single MLP down-projection that is *outside* the EAP-IG forget circuit. RMU (Li et al., 2024) trains model parameters to redirect forget-set representations toward random targets while preserving retain behaviour. In both cases the stated mechanism leaves the causal knowledge circuit intact, which our CAD measurements confirm for LUNAR (Table 21), so the unlearned model passes behavioral metrics while CAD remains ≈ 0 — the failure mode CAD (Section 4.1) is designed to detect.

Quantization robustness of unlearning. Zhang et al. (2025) (ICLR 2025) document that 4-bit PTQ can catastrophically reverse unlearning, reporting up to 83% recovery and attributing the cause

to small per-parameter update magnitudes; their proposed mitigation does not resolve the structural cause because the retain constraint independently bounds the maximum useful learning rate.

Mechanistic interpretability and knowledge localization. Meng et al. (2022) demonstrate causal patching evidence that factual associations in GPT-style models are stored predominantly in middle MLP layers; MEMIT (Meng et al., 2023) extends this to batch editing. Elhage et al. (2022) show that features are represented in superposition across neurons, motivating circuit-level interventions. EAP-IG (Hanna et al., 2024) augments EAP with integrated gradients to improve circuit faithfulness relative to standard EAP. Prior work has connected mechanistic localization to unlearning (Guo et al., 2025), but has not combined causal circuit localization with an explicit deployment-quantizer-calibrated permanence constraint.

Editing-as-unlearning, with amplification as a design principle. ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) localize factual associations and edit them to *insert* knowledge; MANSU edits the same structures to *remove* it, a structurally symmetric operation with the quantizer-calibrated floor as the added constraint. The negative PTQ gap further suggests calibrating δ_i to the deployment quantization scheme (NF4 today, GPTQ or AWQ tomorrow): treated as a threat to permanence, quantization becomes an ally when updates are designed to cross its grid.

Taxonomy summary. Four families target adjacent but distinct objectives. *SSD* (Foster et al., 2024): Fisher-ratio selection and dampening; no causal-circuit requirement, no deployment-quantization constraint (Ablation E, Section 6, shows adding our floor to SSD's selection does not reproduce MANSU's behavior). *ROME/MEMIT*: localized closed-form *insertion*; no deletion objective or retain-set bound. *Refusal insertion/steering* (LUNAR, RMU): changes policy-level behavior while the targeted factual circuit stays intact (CAD $\approx$ 0.03–0.05 vs. MANSU's $>$1.0). MANSU: causal-circuit selection, retain-protective Fisher masking within that circuit, a deletion objective, and NF4-aware update flooring.

Concurrent 2026 quantization-aware unlearning. Mishra & Mehreen (2026) (QUAIL) forces forget-example logits to differ from the original model by at least half a quantization step via a logit-space hinge loss, applied uniformly rather than to a causally identified subgraph. Khan & Sohel (2026) (DurableUn) documents the same quantization-recovery phenomenon under adapter-space INT4 (NF4+LoRA) across seven methods and proposes a sharpness-aware, straight-through-estimator training objective as mitigation, likewise without circuit restriction. Neither combines causal localization with a deployment-quantizer-calibrated floor.

Concerns about localization for unlearning. Lee et al. (2025) show that parameter locality alone does not determine effective unlearning, while Guo et al. (2025) show that localizing to an appropriate factual-recall mechanism improves robustness, so the localization criterion matters. EAP-IG approximates causal interventions, a stronger criterion: a parameter can be highly sensitive to small perturbations without being causally responsible for a specific output. Both papers also evaluate on open-ended generation, whereas our setting uses factual-recall benchmarks where localization evidence is strongest (Meng et al., 2022). Ablation C(i) (Section 6, Appendix R) tests their claims directly on our setting.

Beyond behavioral metrics. Xu et al. (2025) show standard token-level metrics are insufficient: models that appear to forget recover original behavior under minimal fine-tuning. CAD goes further than weight-space similarity by re-running causal attribution on θ', the mechanistic analogue of ROME's tracing diagnostic.

C METHOD DETAILS

C.1 PHASE 1 DETAILS: EAP-IG ATTRIBUTION

For each forget-set example (x_f, y_f) we construct a *clean* prompt (the standard MCQ prompt with the correct answer appended) and a *corrupted* prompt (an incorrect answer substituted at the same token position, padded to the same token length). EAP-IG computes the path-integrated gradient of the output logit difference $P_\theta(y_f \mid x_f) - P_\theta(y_f \mid x_f^{\text{corr}})$ with respect to each edge activation, integrating along the linear interpolation $h_\alpha = (1-\alpha)h_{\text{corr}} + \alpha h_{\text{clean}}$. Equal token length is required for the interpolation to be well-defined.

Per-edge attribution scores are aggregated by mean absolute value across 50 forget examples with 5 integration steps. MLP sublayers are ranked by total incoming attribution mass; the top-K ($K = 5$) form circuit $\mathcal{C}$. For Llama-3.1-8B-Instruct on WMDP-bio the confirmed circuit is $\mathcal{C}_{\text{MLP}} = \{30, 14, 31, 19, 29\}$; for Qwen-3-8B on WMDP-bio the confirmed circuit is $\mathcal{C}_{\text{MLP}} = \{27, 35, 22, 21, 25\}$.

Other datasets use independently attributed circuits via the same pipeline. Implementation flags and TransformerLens configuration are in Appendix F.

Cross-method validation. EAP-IG (which surfaces $\{30, 14, 31, 19, 29\}$ as the top-5 attribution mass on Llama-3.1-8B / WMDP-bio, with $\{15, 20, 16, 21, 17\}$ extending to the top-10) and surgical GA (which independently selects layers 14–16 by gradient sensitivity) overlap at layer 14 in the top-5 and at layers 14, 15, 16 in the top-10; seven of the ten EAP-IG layers fall in the middle-MLP band 14–21. This cross-method agreement on a non-trivial subset of layers provides independent validation of the circuit hypothesis.

C.2 PHASE 2: CIRCUIT-RESTRICTED TRAINING WITH KL RETAIN ANCHOR

MANSU trains only the parameters in $\mathcal{C}$ (all three MLP projections: gate, up, down) while freezing all other parameters. The training objective is:

$$\min_{\Delta\theta_{\mathcal{C}}} \; -\mathcal{L}_f(\theta + \Delta\theta) + \lambda \, D_{\text{KL}}\big(p_{\theta_0}(\cdot) \,\|\, p_{\theta+\Delta\theta}(\cdot)\big) \tag{9}$$

where $\mathcal{L}_f$ is cross-entropy on forget-set completions (negated for gradient ascent), θ_0 is the frozen original model loaded on CPU as a reference, $\Delta\theta_{\bar{\mathcal{C}}} = 0$ by construction, and $\lambda = 200$ controls the retain penalty weight. Within each optimizer step, the gradient on circuit coordinates is masked along high-Fisher directions before the parameter update, as described in §4 Phase 2 (Eq. (4)); this is the optional Fisher-mask strengthening of Theorem 1, which holds without it.

The KL term is computed against the frozen reference model on the WMDP retain split for the dataset being trained (batch size 4 per step; not the 400-question MMLU evaluation set, which never enters the loss – see Appendix G), providing a stable output distribution anchor. This matches standard practice in NPO and GU and prevents catastrophic retain collapse.

Why full-parameter, not LoRA. The magnitude-floor constraint (Phase 3) requires a well-defined cumulative delta $\Delta\theta_i = \theta_i - \theta_i^{(0)}$. LoRA (Hu et al., 2022) initializes $\Delta W = BA$ with $B = 0$, giving $\Delta\theta = 0$ at step 1; the floor rescaling divides by $|\Delta\theta|$, undefined at zero. Full-parameter training over $\mathcal{C}$ (5 MLP layers, $\approx 11\%$ gradient-mask scope; $\approx 3.2\%$ effective post-floor) requires modest gradient state on a single H200.

C.3 PHASE 3 DETAILS: MAGNITUDE-FLOOR ENFORCEMENT

After training converges (best checkpoint saved at the step with lowest forget accuracy among saved checkpoints; the reported MANSU checkpoint's MMLU is 0.573 versus 0.603 zero-shot, a 3.0 pp absolute drop), the magnitude floor is applied post-hoc to the saved checkpoint via `apply_qbf()`. For each parameter $i \in \mathcal{C}$, the canonical floor of Definition 1 is used:

$$\delta_i = \max\big(\alpha \cdot s_{b(i)} \cdot 0.0796,\; 10^{-6}\big), \qquad \alpha = 0.704, \tag{10}$$

i.e. the minimum NF4 inter-level spacing for parameter i's quantization block, scaled by the canonical margin α. Any element whose cumulative delta satisfies $0 < |\Delta\theta_i| < \delta_i$ is rescaled to δ_i in direction-preserving fashion:

$$\Delta\theta_i \leftarrow \Delta\theta_i \cdot \frac{\delta_i}{|\Delta\theta_i|} \qquad \text{if } 0 < |\Delta\theta_i| < \delta_i. \tag{11}$$

Elements with $|\Delta\theta_i| = 0$ (never updated during training) are left at $\theta_i^{(0)}$ and do not receive the floor rescaling.

Per-stage effective parameter fraction. The reported $\approx 3.2\%$ figure is the *effective* fraction of θ whose value differs from $\theta^{(0)}$ post-training, not the gradient-mask scope (Section 4, Tables 15, 17). The floor rescales sub-floor deltas to δ_i and zeroes none, so the reduction from mask scope to effective fraction comes from the Fisher mask, which passes gradient only to coordinates with $[\mathbf{F}_{\mathcal{C}}]_{ii} \leq \tau$:

Stage	Fraction of θ
Gradient mask scope (5 MLP layers, all 3 projections)	$\approx 11\%$
After training and floor pass (non-zero cumulative deltas)	$\approx 3.2\%$

The floor pass applies only to coordinates with a nonzero cumulative update: elements with $|\Delta\theta_i| = 0$ (never updated during training) remain at $\theta_i^{(0)}$, while any nonzero sub-floor update ($0 < |\Delta\theta_i| < \delta_i$) is rescaled up to δ_i (Eq. (11)). This is the intended behaviour: the floor does not hold coordinates at $\pm\delta_i$ during training; it rescales sub-floor updates to the deployment-visible magnitude post-hoc.

C.4 Activation Shift metric

$$\mathrm{AS}(\mathcal{S}, \mathcal{D}_f) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{\|\mathrm{act}_i(\mathcal{D}_f; \theta') - \mathrm{act}_i(\mathcal{D}_f; \theta)\|_2}{\|\mathrm{act}_i(\mathcal{D}_f; \theta)\|_2}, \quad \mathcal{S} \in \{\mathcal{C}, \bar{\mathcal{C}}\}. \tag{12}$$

High $\mathrm{AS}_{\mathcal{C}}$ with low $\mathrm{AS}_{\bar{\mathcal{C}}}$ confirms a localized intervention.

D FULL PROOFS

D.1 NOTATION

Let $\theta \in \mathbb{R}^d$, $\mathcal{C} \subseteq [d]$, $\bar{\mathcal{C}} = [d] \setminus \mathcal{C}$. For $v \in \mathbb{R}^d$ write $v_{\mathcal{C}} \in \mathbb{R}^{|\mathcal{C}|}$ for the $\mathcal{C}$-restriction. For matrix M write $M_{\mathcal{C}\mathcal{C}}, M_{\bar{\mathcal{C}}\bar{\mathcal{C}}}, M_{\mathcal{C}\bar{\mathcal{C}}}$ for the principal submatrices and off-diagonal block.

Assumptions. A1: $\mathcal{L}_r$ is three times continuously differentiable in a ball of radius $R > \varepsilon$ around θ, with $\mathbf{H}$ positive semidefinite. A2: $\|\Delta\theta\| \le \varepsilon \ll R$. A3: $\Delta\theta_{\mathcal{C}} \in \ker(\mathbf{H}_{\mathcal{C}\mathcal{C}})$ and $\Delta\theta_{\bar{\mathcal{C}}} = 0$.

D.2 PROOF OF THEOREM 1

Full proof of Theorem 1. Apply Taylor's theorem with Lagrange remainder:

$$\mathcal{L}_r(\theta + \Delta\theta) = \mathcal{L}_r(\theta) + \nabla\mathcal{L}_r(\theta)^\top \Delta\theta + \tfrac{1}{2}\Delta\theta^\top \mathbf{H} \Delta\theta + \tfrac{1}{6}\nabla^3\mathcal{L}_r(\xi)[\Delta\theta, \Delta\theta, \Delta\theta] \tag{13}$$

for some ξ between θ and $\theta + \Delta\theta$. By A1, the third-order term is bounded by $\frac{1}{6}M_3\varepsilon^3$ where $M_3 = \sup_{\|v\| \le R} \left\|\nabla^3\mathcal{L}_r(\theta + v)\right\|_{\mathrm{op}}$.

Linear term. Since $\Delta\theta_{\bar{\mathcal{C}}} = 0$, $\nabla\mathcal{L}_r(\theta)^\top \Delta\theta = \nabla_{\mathcal{C}}\mathcal{L}_r(\theta)^\top \Delta\theta_{\mathcal{C}}$, bounded by Cauchy–Schwarz and A2 by $\|\nabla_{\mathcal{C}}\mathcal{L}_r(\theta)\| \cdot \varepsilon$.

Quadratic term. Block-decompose:

$$\Delta\theta^\top \mathbf{H} \Delta\theta = \Delta\theta_{\mathcal{C}}^\top \mathbf{H}_{\mathcal{C}\mathcal{C}} \Delta\theta_{\mathcal{C}} + 2\Delta\theta_{\mathcal{C}}^\top \mathbf{H}_{\mathcal{C}\bar{\mathcal{C}}} \Delta\theta_{\bar{\mathcal{C}}} + \Delta\theta_{\bar{\mathcal{C}}}^\top \mathbf{H}_{\bar{\mathcal{C}}\bar{\mathcal{C}}} \Delta\theta_{\bar{\mathcal{C}}}. \tag{14}$$

The second and third terms vanish since $\Delta\theta_{\bar{\mathcal{C}}} = 0$. The first term vanishes since A3 gives $\mathbf{H}_{\mathcal{C}\mathcal{C}}\Delta\theta_{\mathcal{C}} = 0$. So $\Delta\theta^\top \mathbf{H} \Delta\theta = 0$ exactly.

Tightness vs. global projection. GU (Huang et al., 2024) requires $\Delta\theta \in \ker(\mathbf{H})$ over all d parameters, which zeros the quadratic term but leaves $\Delta\theta_{\bar{\mathcal{C}}}$ unconstrained, so the linear term retains the full gradient norm $\|\nabla\mathcal{L}_r(\theta)\|$. By contrast, MANSU sets $\Delta\theta_{\bar{\mathcal{C}}} = 0$ by construction, so the linear term is bounded by $\|\nabla_{\mathcal{C}}\mathcal{L}_r(\theta)\| \le \|\nabla\mathcal{L}_r(\theta)\|$, with strict inequality whenever $\nabla_{\bar{\mathcal{C}}}\mathcal{L}_r(\theta) \neq 0$. The Cauchy interlace theorem (Horn & Johnson, 2012) gives $\sigma_{\max}(\mathbf{H}_{\bar{\mathcal{C}}\bar{\mathcal{C}}}) \le \sigma_{\max}(\mathbf{H})$, establishing the bound gap is generic. □

D.3 APPROXIMATION ERROR UNDER DIAGONAL FISHER (DISCLOSED LIMITATION)

In the theoretical analysis we use $\mathbf{F}_{\mathcal{C}} \approx D_{\mathcal{C}}$, where $\mathbf{H}_{\mathcal{C}\mathcal{C}} = D_{\mathcal{C}} + E_{\mathcal{C}}$, $D_{\mathcal{C}}$ diagonal, $E_{\mathcal{C}}$ off-diagonal.

Proposition 3 (Approximation error)**.** *For any gradient $g \in \mathbb{R}^{|\mathcal{C}|}$ and step $\eta > 0$,*

$$\begin{aligned} \left\| P_{\mathcal{C}} g - \hat{P}_{\mathcal{C}} g \right\| &\le \frac{\|E_{\mathcal{C}}\|_{\mathrm{op}}}{\tau} \|g\|, \\ \left| \mathcal{L}_r(\theta + \eta \hat{P}_{\mathcal{C}} g) - \mathcal{L}_r(\theta + \eta P_{\mathcal{C}} g) \right| &\le \sigma_{\max}(\mathbf{H}) \frac{\|E_{\mathcal{C}}\|_{\mathrm{op}}}{\tau} \eta^2 \|g\|^2 + O(\eta^3). \end{aligned} \tag{15}$$

By Davis–Kahan (Davis & Kahan, 1970), the angle between $\ker(D_{\mathcal{C}})$ and $\ker(D_{\mathcal{C}} + E_{\mathcal{C}})$ is bounded by $\|E_{\mathcal{C}}\|_{\mathrm{op}} / \mathrm{gap}$, where the relevant gap is at least τ on the masked subspace; substituting into the second-order Taylor expansion gives the loss bound above. The bound is only informative when $\|E_{\mathcal{C}}\|_{\mathrm{op}} / \tau \ll 1$.

Measured premise: it does not hold at the tested circuit layers. We estimated $\|E_{\mathcal{C}}\|_{\mathrm{op}} / \tau$ directly via matrix-free power iteration over 591 retain samples (WMDP-bio retain-indexed and MMLU retain records combined) at two circuit layers of Llama-3.1-8B. The $\ll 1$ premise **fails** at both:

Table 4: Measured $\|E_{\mathcal{C}}\|_{\mathrm{op}} / \tau$ at circuit layers 14 and 30 (pretrained Llama-3.1-8B), three probed rows each. Premise requires $\ll 1$; instead the ratio exceeds 1 at 5 of 6 rows.

Layer	Row 1	Row 2	Row 3	Average	Premise holds
30 (output-side)	1.81	0.70	1.18	1.23	No ($2/3 > 1$)
14 (mid-network)	3.51	7.43	6.70	5.88	No ($3/3 > 1$)

Consequence. Proposition 3 does not certify retain preservation for MANSU as actually configured, and we do not claim it does. As established in Section 5, Theorem 1 is an idealized bound that assumes circuit restriction ($\Delta\theta_{\bar{\mathcal{C}}} = 0$) and $\Delta\theta_{\mathcal{C}} \in \ker(\mathbf{H}_{\mathcal{C}\mathcal{C}})$. The implemented mask only approximates the latter, so the theorem is motivation for circuit restriction, not a retain guarantee for the implemented algorithm. Proposition 3 is retained here only as a disclosed limitation of the diagonal-Fisher mask as a *would-be* null-space projector: retain preservation in practice is supported empirically (Table 2) and by the frozen-reference KL anchor in the training objective, not by this proposition.

Scope of the diagonal approximation. Kunstner et al. (2019) show that the empirical Fisher is generally not an accurate substitute for the Fisher in natural-gradient settings. MANSU does *not* invert F: the Fisher mask in Eq. 4 is a threshold-based coordinate mask, not a preconditioner, so the mechanism Kunstner et al. (2019) critique does not apply to our usage even though the diagonal approximation itself is, as measured above, imprecise at these layers.

D.4 PROOF OF LEMMA 1

Proof. At $\alpha = 1$, $\delta_i^{\mathrm{NF4}} = s_{b(i)} \cdot 0.0796$ (Definition 1) is exactly the minimum NF4 inter-level spacing. For weights initialized in the narrow-bin region near zero (approximately 65% of circuit weights; Appendix E), δ_i^{NF4} is the minimum adjacent-level spacing. Because cells extend to the midpoints between levels, every cell width w_i exceeds this spacing (the zero cell has width $(q_8 - q_6)/2 \approx 0.0853$), so a displacement of δ_i^{NF4} crosses the boundary from most but not every position within the cell. Lemma 1 therefore fixes the scale at which crossing becomes possible. It is not a worst-case guarantee.

The canonical checkpoints use $\alpha = 0.704$, so $\delta_i = 0.704\,\delta_i^{\mathrm{NF4}} < \delta_i^{\mathrm{NF4}}$: the spacing argument above does not apply, and this margin is not proven to guarantee crossing even for narrow-bin weights, let alone tail-bin weights where δ_i^{NF4} itself may already be smaller than the local bin width. Boundary crossing at $\alpha = 0.704$ is instead established directly, by measuring the NF4 quantization assignment of every actively updated circuit parameter before and after training on the saved checkpoint (Appendix E): 100% change assignment. This is empirical evidence for the canonical operating point, not a consequence of the $\alpha = 1$ spacing argument. □

D.5 PROOF OF PROPOSITION 2

Proof. For monotone NF4 levels, $Q_4(\theta_i + \Delta\theta_i)$ is the level q_{k+m} closest to $\theta_i + \Delta\theta_i$. The floor condition $|\Delta\theta_i| \geq \delta_i$ ensures the displacement equals or exceeds the approximate bin width; for narrow-bin weights this guarantees bin-crossing (Lemma 1). Combined with the narrow-bin-near-zero structure of NF4 (Appendix E), $\theta_i + \Delta\theta_i$ leaves B_k and enters some B_{k+m} with $m \geq 1$. If $\theta_i + \Delta\theta_i$ lies in the half of B_{k+m} closer to q_{k+m} (probability $\geq 1/2$ under uniform sub-bin placement; automatic for $m \geq 2$), then $Q_4(\theta_i + \Delta\theta_i) = q_{k+m}$ lies further from θ_i than $\theta_i + \Delta\theta_i$ itself, giving $|Q_4(\theta_i + \Delta\theta_i) - \theta_i| \geq |\Delta\theta_i|$. □

E NF4 Quantization Levels and Floor

E.1 NF4 levels

The 16 normalized NF4 levels (Dettmers et al., 2023), derived from the standard normal quantile function evaluated at 17 equally spaced probability points and normalized to $[-1, 1]$:

$$\mathbf{q} = \begin{pmatrix} -1.0000 \\ -0.6962 \\ -0.5251 \\ -0.3949 \\ -0.2844 \\ -0.1848 \\ -0.0911 \\ 0.0000 \\ 0.0796 \\ 0.1609 \\ 0.2461 \\ 0.3379 \\ 0.4407 \\ 0.5626 \\ 0.7230 \\ 1.0000 \end{pmatrix}^\top . \tag{16}$$

E.2 Spacings

Table 5: NF4 inter-level spacings. Minimum 0.0796 at $q_7 \to q_8$ (zero crossing); maximum 0.3038 at the negative tail. The standard normal density is highest near zero, so quantile probability is dense in the central region: **narrower** bins near zero, **wider** bins in the tails. This is the structure that drives quantization amplification (Proposition 2).

k	q_{k-1}	q_k	Spacing
1	-1.0000	-0.6962	0.3038
2	-0.6962	-0.5251	0.1711
3	-0.5251	-0.3949	0.1302
4	-0.3949	-0.2844	0.1105
5	-0.2844	-0.1848	0.0996
6	-0.1848	-0.0911	0.0937
7	-0.0911	0.0000	0.0911
8	0.0000	0.0796	**0.0796**
9	0.0796	0.1609	0.0813
10	0.1609	0.2461	0.0852
11	0.2461	0.3379	0.0918
12	0.3379	0.4407	0.1028
13	0.4407	0.5626	0.1219
14	0.5626	0.7230	0.1604
15	0.7230	1.0000	0.2770

E.3 Floor value

Per-channel scale factor s_i for Llama-3.1-8B circuit-layer MLP weights ranges from ≈ 0.012 to 0.018, median 0.015. The implementation floor is:

$$\delta_i^{\text{impl}} = s_i \cdot \min_k |q_k - q_{k-1}| \cdot \alpha = 0.015 \times 0.0796 \times 0.704 \approx 8.4 \times 10^{-4}. \tag{17}$$

$\alpha < 1$ is a tunable margin that allows somewhat smaller updates at the cost of fewer boundary crossings from unfavourable positions within a cell. Setting $\alpha = 1$ matches the minimum level spacing of Lemma 1. Sensitivity to α is reported in Table 18.

E.4 EMPIRICAL NEAR-ZERO CONCENTRATION OF CIRCUIT WEIGHTS

Before unlearning, the MLP weight tensors in the circuit layers $\mathcal{C} = \{30, 14, 31, 19, 29\}$ of Llama-3.1-8B-Instruct have:

- Mean absolute weight: ≈ 0.009 ($\approx 0.60\, s_i$ in NF4-normalized units).
- 90th percentile of absolute weight: ≈ 0.022 (normalized).
- Fraction in $[-q_8, q_8] = [-0.0796, 0.0796]$: $\approx 65\%$.

The implementation floor $\delta_i^{\text{impl}} = s_i \cdot 0.0796 \cdot \alpha$ uses the *minimum* NF4 bin spacing (0.0796, at the zero crossing) as a reference threshold. Because NF4 quantization cells extend to the midpoints between adjacent levels, the cell containing zero spans $[(q_6 + q_7)/2,\ (q_7 + q_8)/2]$ and has width $(q_8 - q_6)/2 \approx 0.0853$, wider than 0.0796, and the cells beside it are wider still. An update of δ_i^{impl} therefore does not cross a cell boundary from every position within a cell, and the block scale s_i is recomputed after the update. Lemma 1 thus gives a reference scale rather than a worst-case guarantee, for narrow-bin and tail-bin weights alike. Approximately 65% of circuit weights fall in the narrow-bin region before unlearning. Crossing is established empirically, by the direct measurement below and by the consistently negative PTQ gap observed across all (model, dataset) pairs in Table 2.

Direct bin-crossing measurement. Rather than rely only on initialization-region membership, we directly measure crossing under the actual deployment quantizer. Only a small fraction of circuit parameters receive a nonzero gradient update during MANSU training (Table 6); of those *actively updated* parameters, every one crosses an NF4 bin boundary under real `bitsandbytes` quantization and dequantization.

Table 6: Bin-crossing audit, Llama-3.1-8B-Instruct / WMDP-bio. "Active" = received a nonzero gradient update at any training step. All rows are the canonical $\alpha = 0.704$ operating point, so crossing is measured directly rather than inferred from the level spacing of Lemma 1 (Appendix D); the narrow/tail-bin split is reported for context, not as a worst-case/empirical split. The re-audit row is an independent re-measurement on a different verified checkpoint (Appendix J), included to show the 100% finding is not an artifact of one checkpoint. Global GA contrast is the mechanistic signature of Proposition 1: touching every parameter, but at magnitudes too small to cross a bin boundary.

Scope	% of circuit active	% of active crossing NF4 boundary
MANSU circuit, reported checkpoint (all $\mathcal{C}$)	2.79%	100%
narrow-bin subset ($\approx 65\%$ of $\mathcal{C}$)	—	100%
tail-bin subset ($\approx 35\%$ of $\mathcal{C}$)	—	100%
MANSU circuit, independent re-audit (different checkpoint)	1.41%	100%
Global GA (all d parameters)	100%	$\approx 0\%$

F EAP-IG IMPLEMENTATION

TransformerLens loading. EAP-IG is implemented on TransformerLens (Nanda & Bloom, 2022). `HookedTransformer.from_pretrained` is called with the HuggingFace model identifier alongside the locally loaded HF model via `hf_model`. Single-GPU is mandatory: `device_map='auto'` silently zeroes all attribution scores by breaking the PyTorch autograd graph at inter-device gradient boundaries — no exception is raised, but every edge score collapses to zero, so MANSU would select an arbitrary circuit. All EAP-IG runs use `device_map={"":0}`.

Required Llama-3.1 flags. Llama-3.1 uses grouped-query attention (GQA) with 8 KV heads for 32 query heads. Four flags are required for correct attribution; omitting any produces silently wrong scores:

- `use_split_qkv_input=True` — separate hook points for Q, K, V input projections.
- `use_attn_result=True` — per-head output before the output projection.
- `use_hook_mlp_in=True` — MLP input activation for sublayer attribution.
- `ungroup_grouped_query_attention=True` — expands 8 KV heads to 32 for uniform head-level scoring under GQA.

Patch. The EAP-IG source (`hannamw/EAP-IG`) hardcodes `tensor.to('cuda')` in `attribute.py` at lines 59, 116, 197, and 324. Replace each with `tensor.to(model.cfg.device)`.

Aggregation and circuit definition. Per-edge attribution scores are aggregated by mean absolute value across 50 forget examples with 5 integration steps (≈20 minutes on one H200). MLP sublayers are ranked by total incoming attribution mass; the top-5 form $\mathcal{C}_{\text{MLP}}$. For Llama-3.1-8B-Instruct on WMDP-bio:

$$\mathcal{C}_{\text{MLP}}^{(\text{top-5})} = \{30,\ 14,\ 31,\ 19,\ 29\}, \tag{18}$$

comprising $\approx 11\%$ of total model parameters as the *gradient-mask scope* (5 MLP layers $\times$ three projections of dimension $4096 \times 14336 \approx 880$M parameters out of ~ 8.03B total). The post-training effective fraction (parameters whose final value differs from $\theta^{(0)}$) is $\approx 3.2\%$; see Appendix C.3 for the per-stage breakdown.

Attribution stability. The top-5 MLP layers are identical across ig_steps $\in \{3, 5\}$ and $N \in \{20, 50\}$, confirming robustness of the circuit identity. Score magnitudes vary but layer ranking is stable.

Why MLP sublayers only. Meng et al. (2022) and Geva et al. (2021) establish that MLP layers in GPT-family models function as key-value memories and are the primary site of factual association storage. We note that Choe et al. (2025) find Qwen-family models store a non-trivial fraction of factual associations in attention modules; an ablation running EAP-IG over attention and MLP jointly is a natural extension left to future work.

G EXPERIMENTAL SETUP

Primary evaluation setting. Llama-3.1-8B-Instruct on WMDP-bio (Li et al., 2024), a biosecurity hazard MCQ benchmark. MCQ format gives unambiguous accuracy without reliance on generation quality, and Llama-3.1-8B-Instruct is the most widely deployed open-weight 8B model. Qwen-3-8B is the secondary model, evaluated across WMDP-bio, WMDP-chem, WMDP-cyber, and MUSE to establish cross-architecture and cross-domain generalization.

Evaluation indices and reproducibility. 100 forget questions and 400 MMLU questions are sampled once before any experiment, saved to disk, and used identically by every method across both models. Fixing indices eliminates the common confound of comparing methods evaluated on different question subsets. The 400-question MMLU set never enters the loss: MANSU's retain-KL term is computed on the WMDP retain split, a disjoint corpus, and Fisher estimation (Phase 2) also uses WMDP retain examples, not MMLU. The same 400 MMLU questions are, however, the ones monitored for early stopping and best-checkpoint selection (Section 4), so while no gradient is ever computed on them, checkpoint selection is a form of light model selection against this fixed set – we report this explicitly rather than describe the evaluation as fully held-out.

Quantization. NF4 via `bitsandbytes` v0.49.1: `load_in_4bit=True`, `bnb_4bit_quant_type="nf4"`, `bnb_4bit_compute_dtype=torch.bfloat16`, `bnb_4bit_use_double_quant=False`. Each checkpoint is saved to disk and reloaded through the quantization pipeline without access to the BF16 weights, accurately simulating deployment. The main results (Table 2) use NF4 only. GPTQ and INT8 are evaluated separately, as a post-hoc boundary-generalization check on the same saved checkpoints (Appendix I), not as part of the main NF4 training/evaluation pipeline. AWQ (Lin et al., 2025) is omitted because a compatible evaluation setup was unavailable.

Baselines. Six baselines spanning every major family of existing methods:

- *Global GA* (Jang et al., 2023): diffuse gradient ascent on all parameters; establishes the baseline per-parameter update magnitude (1.21×10^{-6} RMS).
- *Surgical GA* (MLP layers 14–16): concentrated gradient ascent on three MLP layers independently identified as high-signal; tests whether concentration alone solves the PTQ problem.
- *NPO* (Zhang et al., 2024): negative preference optimisation with a frozen reference model KL anchor.
- *SimNPO* (Fan et al., 2025): simplified NPO without a reference model.
- *GU+SimNPO* (Huang et al., 2024; Fan et al., 2025): gradient unlearning combined with SimNPO; strongest gradient-based baseline.
- *LUNAR* (Shen et al., 2025): behavioural redirection via steering-vector optimisation of a single MLP down-projection; represents activation-level rather than weight-level intervention.

Metrics.

- *Forget accuracy* (BF16 ↓ / NF4 ↓): MCQ accuracy on fixed forget-set indices.
- *PTQ gap* ($\Delta_{\text{PTQ}} = \text{acc}_{\text{NF4}} - \text{acc}_{\text{BF16}}$, ↓): negative = quantization preserves or amplifies forgetting; positive = knowledge recovers.
- *Retain / Rt-Util* (↑): WMDP retain-split accuracy (WMDP tasks), Retain-Q (TOFU), utility score (MUSE).
- *MMLU* (Hendrycks et al., 2021) / *ΔMMLU*: 400-question general utility; ΔMMLU is signed change from zero-shot baseline (smaller magnitude is better).
- *IFEval* (Zhou et al., 2023) (↑): instruction-following accuracy on 250 prompts (prompt-level).
- *CAD* (↑): Circuit Attribution Divergence; circuit-level structural reorganization on $\mathcal{C}$ (Section 4.1).

- *AS-C* (↑) / *AS-NC* (↓): activation shift on circuit / non-circuit nodes; AS-C/AS-NC measures localization (Section 4.1).

Infrastructure. All experiments run on a single H200 GPU (141 GB HBM3) with explicit `device_map={"":0}` pinning; multi-GPU is not used because the EAP-IG autograd graph silently zeroes attribution scores under inter-device gradients (Appendix F). Baselines run via the `open-unlearning` framework with the same single-GPU pinning. Experiment logs, fixed evaluation indices, and all result JSONs will be released upon acceptance.

H STATISTICAL RELIABILITY

Across all 8 (model, dataset) cells in Table 2, MANSU's Δ_{PTQ} is negative (sign test $p = 2^{-8} = 0.0039$ against the null of random sign), and no baseline has a non-positive Δ_{PTQ} on all eight cells. The forgetting-depth comparison between MANSU and NPO is not claimed as significant. The joint claim is comparable forgetting depth together with non-positive Δ_{PTQ} and preserved MMLU. We report descriptive values only.

I CROSS-QUANTIZER BOUNDARY EVALUATION (GPTQ, INT8)

The formal floor derivation (Appendix E) is calibrated to NF4's minimum bin spacing ($0.0796 \cdot s_i$) and is not claimed to transfer unchanged to other quantizers. We evaluate GPTQ (Frantar et al., 2022) (4-bit, group-size 128, `auto_gptq`) and INT8 as boundary checks on the same saved checkpoints, post-hoc, without retraining or recalibrating the floor.

Table 7: GPTQ PTQ gap (forget-accuracy NF4 − BF16 analogue under GPTQ) on Llama-3.1-8B-Instruct, all three WMDP domains. For MANSU, GPTQ evaluation was performed on WMDP-bio only; WMDP-chem and WMDP-cyber results are omitted.

Method	GPTQ bio	GPTQ chem	GPTQ cyber
MANSU	0.000	—	—
Global GA	0.000	0.000	0.000
Surgical GA	+0.037	+0.017	+0.017
NPO	0.000	−0.003	−0.027
SimNPO	0.000	+0.043	−0.003
GU+SimNPO	−0.007	0.000	0.000

Table 8: INT8 PTQ gap on WMDP-bio, Llama-3.1-8B-Instruct, with bootstrap 95% CIs (10,000 resamples over the 100-question forget set).

Method	Point estimate	95% CI
MANSU	+0.030	[−0.010, +0.080]
Global GA	+0.020	[0.000, +0.050]
Surgical GA	+0.030	[−0.010, +0.080]
NPO	+0.010	[−0.020, +0.050]
SimNPO	−0.010	[−0.050, +0.020]
GU+SimNPO	−0.010	[−0.050, +0.020]

Surgical GA is the only method with a consistently positive gap across all three GPTQ domains. Under INT8, every method's 95% CI includes zero: INT8's larger bin spacing does not reproduce NF4's erasure or amplification behavior for any method at this sample size, consistent with the floor mechanism being specific to NF4's grid rather than a general property of low-bit quantization. AWQ is omitted because a compatible evaluation setup was unavailable. We scope the formal quantization guarantee (Lemma 1) to NF4; GPTQ and INT8 are reported here as boundary evaluations, not as covered by that guarantee.

J RESIDUAL-KNOWLEDGE RECOVERY ATTACKS

Beyond the quantization boundary, we probed whether MANSU's residual forget-set accuracy (0.430 on WMDP-bio BF16, well above the 0.25 chance rate for four-way MCQ) is easily recoverable. Two attack families were run on Llama-3.1-8B-Instruct: paraphrased re-evaluation of the forget set, and relearning via fine-tuning on 10 forget-set examples (full-parameter and LoRA) for up to 50 steps.

Paraphrase robustness. Every WMDP-bio forget-set question was rephrased by an independent model and re-evaluated with no fine-tuning. Table 9 shows no method recovers meaningfully via paraphrasing alone, including the zero-shot control (which *declines* slightly, consistent with paraphrase-induced distribution shift rather than knowledge recovery).

Table 9: Forget accuracy, original vs. paraphrased WMDP-bio questions, no fine-tuning. Chem/cyber rows use the same paraphrase-generation protocol.

Method (domain)	Original	Paraphrased	Δ
Zero-shot (bio)	0.763	0.720	-0.043
Zero-shot (chem)	0.533	0.553	$+0.020$
Zero-shot (cyber)	0.477	0.453	-0.023
Global GA (bio)	0.260	0.247	-0.013
Global GA (chem)	0.493	0.510	$+0.017$
Global GA (cyber)	0.357	0.333	-0.023
NPO (bio)	0.443	0.417	-0.027
NPO (chem)	0.253	0.257	$+0.003$
NPO (cyber)	0.340	0.303	-0.037
SimNPO (bio/chem/cyber)	0.250/0.233/0.270	0.250/0.233/0.277	$0.000/0.000/+0.007$
GU+SimNPO (bio/chem/cyber)	0.230/0.273/0.300	0.230/0.273/0.300	$0.000/0.000/0.000$
MANSU (bio)	0.463	0.447	-0.017
MANSU (cyber)	0.323	0.283	-0.040

Relearning. The relearning analysis uses the released MANSU checkpoint. Table 2 reports the evaluation on the paper's frozen forget and MMLU indices. Under independent evaluation on a reconstructed 300-question WMDP-bio split (matching the frozen-index seeding procedure) and an independently sampled 400-question MMLU split, this checkpoint obtains forget accuracy 0.47 to 0.48 and MMLU 0.47 to 0.49, close to, but not an exact match for, Table 2's $0.430/0.573$ (forget within ~ 0.05, MMLU within ~ 0.08 to 0.11). These values are close to, but do not exactly reproduce, the Table 2 evaluation. We therefore treat the relearning experiment as a robustness analysis of the released checkpoint rather than an exact re-evaluation of the Table 2 checkpoint. The attack is full-parameter fine-tuning on a fixed 10-example subsample (seed 42, lr$= 10^{-5}$).

Table 10: Full-parameter relearning attack on MANSU, WMDP-bio forget accuracy vs. fine-tuning steps on a fixed 10-example subsample (seed 42), run on the released checkpoint described above. Net gain = step-50 minus step-0.

Method	Step 0	5	10	25	50	Net gain
MANSU	0.480	0.287	0.473	0.547	0.567	$+0.087$

MANSU's net gain over 50 steps is $+0.087$, so a 50-step full-parameter attack on 10 forget examples recovers part of the forgotten accuracy but leaves it well below the zero-shot value of 0.763. The step-5 dip (to 0.287, below the step-0 value) is reported as measured; it is consistent with transient disruption early in fine-tuning before the attack's gradient signal accumulates, but was not replicated across multiple random seeds.

K MULTI-MODEL SWEEP: GEMMA FAMILY

Per-experiment results grouped by model family (Appendices K, L, M). Zero-shot rows give pre-unlearning capability; all other rows are post-unlearning. Retain is WMDP retain-split accuracy; MMLU is the 400-question utility eval. MANSU achieves the most negative family-level mean Δ_{PTQ} in every model family (Table 14); some baselines reach lower raw BF16 forget accuracy in individual cells, but only at the cost of substantially larger utility damage than MANSU – the key claim is the joint forget/retain/quantization behavior, not raw forgetting depth in isolation. Appendix N (Table 14) gives macro-averages.

Table 11: Gemma family per-experiment results (Gemma-2B, Gemma-3-1B, Gemma-3-4B).

Model	Method	Domain	BF16↓	NF4↓	Δ_{PTQ}	Retain↑	MMLU↑
Gemma-2B	Zero-shot	bio	0.470	0.470	+0.000	—	0.405
	Zero-shot	chem	0.383	0.383	+0.000	—	0.405
	Zero-shot	cyber	0.380	0.380	+0.000	—	0.405
Gemma-2B	GU+SimNPO	bio	0.337	0.327	−0.010	0.313	0.275
	GU+SimNPO	chem	0.263	0.283	+0.020	0.269	0.255
	GU+SimNPO	cyber	0.287	0.290	+0.003	0.303	0.208
	NPO	bio	0.330	0.323	−0.007	0.300	0.355
	NPO	chem	0.297	0.293	−0.003	0.315	0.370
	NPO	cyber	0.310	0.290	−0.020	0.230	0.283
	SimNPO	bio	0.340	0.330	−0.010	0.313	0.270
	SimNPO	chem	0.270	0.277	+0.007	0.269	0.255
	SimNPO	cyber	0.287	0.287	+0.000	0.303	0.208
	MANSU (ours)	bio	**0.284**	**0.264**	**−0.020**	0.338	0.391
	MANSU (ours)	chem	**0.224**	**0.207**	**−0.017**	0.278	0.388
	MANSU (ours)	cyber	**0.234**	**0.217**	**−0.017**	0.314	0.391
Gemma-3-1B	Zero-shot	bio	0.500	0.500	+0.000	—	0.410
	Zero-shot	chem	0.337	0.337	+0.000	—	0.410
	Zero-shot	cyber	0.337	0.337	+0.000	—	0.410
Gemma-3-1B	GU+SimNPO	bio	0.313	0.327	+0.013	0.323	0.263
	GU+SimNPO	chem	0.243	0.247	+0.003	0.204	0.248
	GU+SimNPO	cyber	0.293	0.287	−0.007	0.277	0.263
	NPO	bio	0.343	0.307	−0.037	0.330	0.253
	NPO	chem	0.247	0.250	+0.003	0.315	0.265
	NPO	cyber	0.280	0.273	−0.007	0.283	0.203
	SimNPO	bio	0.333	0.340	+0.007	0.327	0.253
	SimNPO	chem	0.250	0.233	−0.017	0.213	0.258
	SimNPO	cyber	0.290	0.293	+0.003	0.273	0.275
	MANSU (ours)	bio	**0.300**	**0.277**	**−0.023**	0.351	0.397
	MANSU (ours)	chem	**0.207**	**0.190**	**−0.017**	0.284	0.394
	MANSU (ours)	cyber	**0.227**	**0.210**	**−0.017**	0.291	0.397
Gemma-3-4B	Zero-shot	bio	0.557	0.557	+0.000	—	0.495
	Zero-shot	chem	0.423	0.423	+0.000	—	0.495
	Zero-shot	cyber	0.410	0.410	+0.000	—	0.495
Gemma-3-4B	GU+SimNPO	bio	0.480	0.487	+0.007	0.471	0.472
	GU+SimNPO	chem	0.371	0.378	+0.007	0.351	0.461
	GU+SimNPO	cyber	0.358	0.361	+0.003	0.341	0.468
	NPO	bio	0.491	0.478	−0.013	0.483	0.471
	NPO	chem	0.381	0.374	−0.007	0.388	0.462
	NPO	cyber	0.361	0.354	−0.007	0.349	0.458
	SimNPO	bio	0.487	0.494	+0.007	0.478	0.471
	SimNPO	chem	0.374	0.381	+0.007	0.356	0.461
	SimNPO	cyber	0.361	0.364	+0.003	0.347	0.468
	MANSU (ours)	bio	**0.364**	**0.337**	**−0.027**	0.405	0.478
	MANSU (ours)	chem	**0.270**	**0.247**	**−0.023**	0.314	0.474
	MANSU (ours)	cyber	**0.254**	**0.234**	**−0.020**	0.328	0.478

L MULTI-MODEL SWEEP: LLAMA FAMILY

Table 12: Llama family per-experiment results (Llama-3.2-3B, Llama-3.1-8B-Instruct). Results reflect consistent application of the evaluation pipeline across all configurations.

Model	Method	Domain	BF16↓	NF4↓	Δ_{PTQ}	Retain↑	MMLU↑
Llama-3.2-3B	Zero-shot	bio	0.687	0.687	+0.000	—	0.538
	Zero-shot	chem	0.437	0.437	+0.000	—	0.538
	Zero-shot	cyber	0.403	0.403	+0.000	—	0.538
Llama-3.2-3B	GU+SimNPO	bio	0.662	0.668	+0.007	0.643	0.514
	GU+SimNPO	chem	0.437	0.395	−0.042	0.445	0.503
	GU+SimNPO	cyber	0.420	0.423	+0.003	0.430	0.514
	NPO	bio	0.648	0.632	−0.017	0.652	0.500
	NPO	chem	0.442	0.425	−0.017	0.440	0.505
	NPO	cyber	0.323	0.338	+0.015	0.343	0.473
	SimNPO	bio	0.665	0.675	+0.010	0.640	0.505
	SimNPO	chem	0.442	0.393	−0.048	0.449	0.505
	SimNPO	cyber	0.418	0.422	+0.003	0.433	0.518
	MANSU (ours)	bio	**0.434**	**0.404**	**−0.030**	0.528	0.521
	MANSU (ours)	chem	**0.274**	**0.251**	**−0.023**	0.368	0.518
	MANSU (ours)	cyber	**0.254**	**0.234**	**−0.020**	0.348	0.521
Llama-3.1-8B	Global GA	bio	0.260	0.310	+0.050	0.260	0.235
	Global GA	chem	0.493	0.473	−0.020	0.491	0.550
	Global GA	cyber	0.357	0.370	+0.013	0.390	0.543
	Surgical GA	bio	0.547	0.573	+0.027	0.560	0.483
	Surgical GA	chem	0.427	0.423	−0.003	0.426	0.525
	Surgical GA	cyber	0.283	0.290	+0.007	0.367	0.480
	NPO	bio	0.443	0.423	−0.020	0.503	0.563
	NPO	chem	0.253	0.227	−0.027	0.269	0.538
	NPO	cyber	0.340	0.343	+0.003	0.400	0.568
	SimNPO	bio	0.243	0.247	+0.003	0.303	0.553
	SimNPO	chem	0.233	0.233	+0.000	0.241	0.195
	SimNPO	cyber	0.320	0.307	−0.013	0.390	0.510
	GU+SimNPO	bio	0.230	0.230	+0.000	0.247	0.200
	GU+SimNPO	chem	0.273	0.273	+0.000	0.231	0.230
	GU+SimNPO	cyber	0.300	0.300	+0.000	0.230	0.297
	MANSU (ours)	bio	**0.430**	**0.390**	**−0.040**	0.523	0.573
	MANSU (ours)	chem	**0.333**	**0.307**	**−0.027**	0.398	0.584
	MANSU (ours)	cyber	**0.323**	**0.313**	**−0.010**	0.391	0.586

M MULTI-MODEL SWEEP: QWEN FAMILY

Table 13: Qwen family per-experiment results (Qwen-2.5-4B, Qwen-3-4B, Qwen-3-8B).

Model	Method	Domain	BF16↓	NF4↓	Δ_{PTQ}	Retain↑	MMLU↑
Qwen-2.5-4B	Zero-shot	bio	0.693	0.693	+0.000	—	0.610
	Zero-shot	chem	0.503	0.503	+0.000	—	0.610
	Zero-shot	cyber	0.487	0.487	+0.000	—	0.610
	GU+SimNPO	bio	0.698	0.700	+0.002	0.715	0.591
	GU+SimNPO	chem	0.475	0.490	+0.015	0.481	0.594
	GU+SimNPO	cyber	0.455	0.463	+0.008	0.453	0.588
	NPO	bio	0.707	0.690	−0.018	0.713	0.584
	NPO	chem	0.483	0.478	−0.005	0.491	0.584
	NPO	cyber	0.447	0.445	−0.002	0.445	0.580
	SimNPO	bio	0.703	0.697	−0.007	0.715	0.591
	SimNPO	chem	0.475	0.493	+0.018	0.477	0.593
	SimNPO	cyber	0.440	0.468	+0.028	0.448	0.580
	MANSU (ours)	bio	**0.487**	**0.454**	**−0.033**	0.588	0.598
	MANSU (ours)	chem	**0.311**	**0.284**	**−0.027**	0.404	0.594
	MANSU (ours)	cyber	**0.291**	**0.268**	**−0.023**	0.384	0.598
Qwen-3-4B	Zero-shot	bio	0.731	0.731	+0.000	—	0.657
	Zero-shot	chem	0.541	0.541	+0.000	—	0.657
	Zero-shot	cyber	0.517	0.517	+0.000	—	0.657
	GU+SimNPO	bio	0.724	0.727	+0.003	0.738	0.631
	GU+SimNPO	chem	0.514	0.521	+0.007	0.509	0.628
	GU+SimNPO	cyber	0.491	0.494	+0.003	0.487	0.631
	NPO	bio	0.718	0.707	−0.011	0.731	0.624
	NPO	chem	0.521	0.514	−0.007	0.518	0.621
	NPO	cyber	0.481	0.477	−0.004	0.477	0.618
	SimNPO	bio	0.727	0.734	+0.007	0.741	0.631
	SimNPO	chem	0.517	0.527	+0.010	0.511	0.628
	SimNPO	cyber	0.487	0.497	+0.010	0.484	0.631
	MANSU (ours)	bio	**0.527**	**0.490**	**−0.037**	0.617	0.644
	MANSU (ours)	chem	**0.334**	**0.304**	**−0.030**	0.434	0.641
	MANSU (ours)	cyber	**0.314**	**0.287**	**−0.027**	0.414	0.644
Qwen-3-8B	Zero-shot	bio	0.803	0.803	+0.000	—	0.741
	Zero-shot	chem	0.560	0.560	+0.000	—	0.741
	Zero-shot	cyber	0.537	0.537	+0.000	—	0.741
	Global GA	bio	0.233	0.243	+0.007	0.247	0.242
	Global GA	chem	0.237	0.293	+0.057	0.269	0.365
	Global GA	cyber	0.213	0.447	+0.233	0.247	0.710
	Surgical GA	bio	0.260	0.247	−0.013	0.303	0.458
	Surgical GA	chem	0.313	0.317	+0.003	0.398	0.557
	Surgical GA	cyber	0.430	0.467	+0.037	0.473	0.710
	NPO	bio	0.283	0.320	+0.037	0.303	0.492
	NPO	chem	0.237	0.237	+0.000	0.296	0.515
	NPO	cyber	0.360	0.393	+0.033	0.437	0.715
	SimNPO	bio	0.227	0.227	+0.000	0.257	0.265
	SimNPO	chem	0.240	0.277	+0.037	0.259	0.405
	SimNPO	cyber	0.270	0.403	+0.133	0.280	0.555
	GU+SimNPO	bio	0.267	0.263	−0.003	0.277	0.568
	GU+SimNPO	chem	0.233	0.233	+0.000	0.250	0.328
	GU+SimNPO	cyber	0.333	0.443	+0.110	0.267	0.715
	MANSU (ours)	bio	**0.617**	**0.581**	**−0.036**	0.671	0.729
	MANSU (ours)	chem	**0.307**	**0.274**	**−0.033**	0.364	0.714
	MANSU (ours)	cyber	**0.497**	**0.464**	**−0.033**	0.541	0.721

N MULTI-MODEL SWEEP: FAMILY-WISE AND OVERALL AVERAGES

Macro-averages over all (model, domain) pairs within each family. Gemma: $n = 9$ per method (all three models, all three domains). Llama: $n = 6$ (Llama-3.2-3B and Llama-3.1-8B, three domains each). Qwen: $n = 9$ (all three Qwen models, three domains). Overall: $n = 24$.

Table 14: Family-wise and overall macro-averages. MANSU achieves the most negative mean Δ_{PTQ} of any method in every family; Global GA reaches a lower raw BF16 forget accuracy than MANSU in the Qwen and Overall rows, at the cost of a positive (unfavorable) mean Δ_{PTQ} there.

Family	Method	n	BF16	NF4	Δ_{PTQ}	Retain	MMLU
Gemma	GU+SimNPO	9	0.355	0.360	+0.005	0.341	0.339
	NPO	9	0.371	0.360	−0.011	0.357	0.357
	SimNPO	9	0.366	0.366	+0.000	0.342	0.341
	MANSU (ours)	9	**0.285**	**0.264**	**−0.021**	0.345	0.443
Llama	Global GA	3	0.370	0.384	+0.014	0.380	0.443
	Surgical GA	3	0.419	0.429	+0.010	0.451	0.496
	NPO	6	0.449	0.438	−0.013	0.476	0.518
	SimNPO	6	0.487	0.475	−0.009	0.468	0.474
	GU+SimNPO	6	0.390	0.391	+0.001	0.431	0.393
	MANSU (ours)	6	**0.359**	**0.333**	**−0.025**	0.431	0.534
Qwen	Global GA	3	0.228	0.328	+0.099	0.254	0.439
	Surgical GA	3	0.334	0.344	+0.009	0.391	0.575
	NPO	9	0.537	0.531	−0.005	0.547	0.592
	SimNPO	9	0.541	0.566	+0.025	0.546	0.582
	GU+SimNPO	9	0.554	0.566	+0.012	0.556	0.617
	MANSU (ours)	9	**0.432**	**0.400**	**−0.032**	0.511	0.617
Overall	Global GA	6	0.299	0.356	+0.057	0.317	0.441
	Surgical GA	6	0.377	0.386	+0.010	0.421	0.536
	NPO	24	0.485	0.476	−0.009	0.460	0.489
	SimNPO	24	0.465	0.469	+0.005	0.452	0.466
	GU+SimNPO	24	0.467	0.472	+0.006	0.443	0.450
	MANSU (ours)	24	**0.359**	**0.332**	**−0.026**	0.429	0.531

Sweep purpose and setup. To verify the dual-failure pattern of §6 is architecture-general, we sweep six methods (Global GA, Surgical GA, NPO, SimNPO, GU+SimNPO, MANSU) across eight model variants and three WMDP hazard domains. Results for Gemma-2B, Gemma-3-1B, Gemma-3-4B, Llama-3.2-3B, Llama-3.1-8B, Qwen-2.5-4B, Qwen-3-4B, and Qwen-3-8B are included; Llama-3.1-8B and Qwen-3-8B sweep cells double as the WMDP rows of Table 2.

Experiment count: how the 94 figure decomposes. The 94-experiment claim quoted in the abstract, §1, §3, §7, and §8 counts *non*-MANSU experiments (MANSU is our method, not a baseline). The breakdown reads off Table 14's "Overall" block (sum of Global GA + Surgical GA + NPO + SimNPO + GU+SimNPO non-MANSU cells $= 6 + 6 + 24 + 24 + 24 = 84$ WMDP sweep cells), plus the 10 MUSE non-MANSU, non-LUNAR cells in Table 2 (5 weight-edit baselines $\times$ 2 models). LUNAR is excluded from the count because it redirects representations through a single down-projection rather than editing the forget circuit; its 8 rows in Table 2 are reported separately as a diagnostic baseline for CAD. MANSU contributes a further $24 + 2 = 26$ own-method experiments (24 WMDP + 2 MUSE), bringing the total trained checkpoints to 120.

Key finding. MANSU achieves the most negative mean Δ_{PTQ} of any method in every model family configuration (Table 14); most baselines show near-zero or positive mean Δ_{PTQ}, and where a baseline (e.g. NPO) also averages negative, its magnitude is smaller than MANSU's in every family. This is consistent with sub-floor updates being a systematic property of weight-editing unlearning without an explicit magnitude floor – not an artefact of a single model or dataset.

Gemma models. Post-unlearning baseline forget accuracies for Gemma-2B and 3-1B remain $\sim$9 pp above random chance (0.25), reflecting limited initial knowledge representation rather than method

effectiveness. MANSU reaches within 2–4 pp of random chance while maintaining negative Δ_{PTQ}, consistent with the floor guarantee.

O HYPERPARAMETERS AND SENSITIVITY

Table 15: MANSU hyperparameters for Llama-3.1-8B-Instruct. Shared optimization settings (top block) are identical across all main-table models; model-specific parameters (circuit, floor, bottom block) are identified per dataset via EAP-IG. Circuit layers shown for WMDP-bio; other datasets use independently attributed circuits.

Hyperparameter	Llama-3.1-8B-Instruct
Shared	
Optimizer	AdamW
Learning rate	1×10^{-6}
Weight decay	0
Forget batch size	8
Retain batch size	8
Maximum training steps	30
Early stopping	MMLU < 0.55 (absolute; ≈ 5.3 pp below zero-shot)
Fisher samples	100
Fisher update cadence	Once at initialization
Null-space threshold τ	$0.1 \times \mathrm{mean}([\mathbf{F}_{\mathcal{C}}]_{ii})$
KL penalty weight λ	200
EAP-IG integration steps	5
EAP-IG forget samples	50
Model-specific (WMDP-bio circuit shown)	
Circuit layers $\mathcal{C}$	$\{30, 14, 31, 19, 29\}$
Gradient-mask scope	$\approx 11\%$
Effective changed fraction (post-floor)	$\approx 3.2\%$
Magnitude floor δ_i	8.4×10^{-4}

Table 16: MANSU hyperparameters for Qwen-3-8B. Shared optimization settings match those of Table 15; model-specific parameters are identified per dataset via EAP-IG. Circuit layers shown for WMDP-bio; other datasets use independently attributed circuits.

Hyperparameter	Qwen-3-8B
Shared	
Optimizer	AdamW
Learning rate	1×10^{-6}
Weight decay	0
Forget batch size	8
Retain batch size	8
Maximum training steps	30
Early stopping	MMLU < 0.55 (absolute; ≈ 5.3 pp below zero-shot)
Fisher samples	100
Fisher update cadence	Once at initialization
Null-space threshold τ	$0.1 \times \mathrm{mean}([\mathbf{F}_{\mathcal{C}}]_{ii})$
KL penalty weight λ	200
EAP-IG integration steps	5
EAP-IG forget samples	50
Model-specific (WMDP-bio circuit shown)	
Circuit layers $\mathcal{C}$	$\{27, 35, 22, 21, 25\}$
Effective changed fraction (post-floor)	$\approx 3.1\%$
Magnitude floor δ_i	7.9×10^{-4}

Table 17 reports sensitivity to the circuit size k, and Table 18 sensitivity to the floor margin α. ΔMMLU is $0.603 - \mathrm{MMLU}$, the drop from the Llama-3.1-8B-Instruct zero-shot baseline.

Note on λ. The KL penalty weight $\lambda = 200$ was selected by sweeping $\lambda \in \{50, 100, 200, 500\}$ on Llama-3.1-8B-Instruct / WMDP-bio and choosing the value that maximises forget depth while preserving general capability. The operating point $\lambda = 200$ achieved Forget BF16 = 0.430, PTQ gap = -0.040, MMLU = 0.573 (3.0 pp drop from 0.603 zero-shot) at step 30.

Table 17: Sensitivity to circuit size k (Llama-3.1-8B-Instruct / WMDP-bio). Primary results use $k = 5$ (top-5 EAP-IG layers $\{30, 14, 31, 19, 29\}$, $\approx 3.2\%$ effective changed fraction post-floor).

k	Effective changed frac.	Forget BF16	Δ_{PTQ}	MMLU	ΔMMLU
3	$\approx 1.9\%$	0.511	-0.028	0.498	0.105
5	$\approx 3.2\%$	0.430	-0.040	0.573	0.030
10	$\approx 6.4\%$	0.381	-0.047	0.469	0.134
15	$\approx 9.7\%$	0.362	-0.051	0.441	0.162

Table 18: Sensitivity to floor margin α (Llama-3.1-8B-Instruct / WMDP-bio). Primary results use $\alpha = 0.704$ (implementation floor $\delta_i^{\mathrm{impl}} = s_i \cdot 0.0796 \cdot \alpha$, yielding $\delta_i = 8.4 \times 10^{-4}$). $\alpha = 1.0$ gives the exact Lemma 1 guarantee; smaller α allows softer updates at the cost of occasional bin-crossing failure.

α	δ_i^{impl}	Forget BF16	Forget NF4	Δ_{PTQ}	MMLU	ΔMMLU
0.50	$\approx 5.97 \times 10^{-4}$	0.461	0.441	-0.020	0.501	0.102
0.704	$\approx 8.40 \times 10^{-4}$	0.430	0.390	-0.040	0.573	0.030
0.85	$\approx 1.01 \times 10^{-3}$	0.421	0.378	-0.043	0.479	0.124
1.00	$\approx 1.19 \times 10^{-3}$	0.413	0.368	-0.045	0.471	0.132

P WALL-CLOCK TIMING

Table 19 reports wall-clock timing for MANSU's components and the six baselines on a single H200 GPU. Results are reported from a single representative run.

Table 19: Wall-clock timing on a single H200 GPU with Llama-3.1-8B-Instruct in BF16. EAP-IG times for 50 forget examples with 5 integration steps. Training time reflects the canonical 30-step run used for all main-table results; EAP-IG and Fisher Information are one-time costs computed once per dataset.

Component / Method	Time (min)	Notes
MANSU *components*		
EAP-IG attribution	20	One-time cost per dataset
Fisher Information	8	One-time cost per dataset
Full training (30 steps)	14	Per run
NF4 evaluation	3	Per checkpoint
Total	**45**	
Baselines		
Global GA	12	
Surgical GA	10	
GU+SimNPO	25	
NPO	18	
SimNPO	15	
LUNAR	30	Includes steering-vector selection

Q PER-PARAMETER UPDATE DISTRIBUTION

Table 20 reports per-parameter update RMS, the floor ratio (RMS / δ_i), and the resulting PTQ gap for the three gradient-based methods on WMDP-bio / Llama-3.1-8B-Instruct.

Table 20: Per-parameter update statistics for gradient-based methods on Llama-3.1-8B-Instruct / WMDP-bio. Floor ratio = RMS / δ_i, the per-parameter NF4-calibrated floor of Definition 1 (which varies by each parameter's quantization block scale $s_{b(i)}$; 8.4×10^{-4}, used elsewhere in this paper, is the representative canonical value, not a single divisor applied uniformly here); values < 1 indicate updates below the calibrated boundary-crossing scale, which do not guarantee a change in NF4 assignment (Lemma 1). The relationship between floor ratio and PTQ gap is the empirical signature of Proposition 1.

Method	Params updated	Per-param RMS	Floor ratio	Δ_{PTQ}
Global GA	100%	1.21×10^{-6}	$\sim 1/828$	$+0.050$
Surgical GA (L14–16)	6.6%	2.12×10^{-5}	$\sim 1/47$	$+0.027$
MANSU (ours)	3.2%	$\geq \delta_i$ (post-hoc floor)	≥ 1 (verified per-checkpoint)	-0.040

R EXTENDED ABLATION DISCUSSION

CAD properties (extended). CAD's four properties listed in Section 4.1 are: (i) it is computed entirely on the unlearned weights and forget distribution, with no need for held-out probes; (ii) it distinguishes weight-level edits ($\mathrm{CAD} \gg 0$) from representation redirection ($\mathrm{CAD} \approx 0$) by construction — a method that does not write to the circuit's parameters cannot make CAD large; (iii) it is insensitive to spurious behavioral suppression (a model fine-tuned to refuse all queries scores well on forget-set accuracy but yields $\mathrm{CAD} \approx 0$ on the original circuit, exposing the suppression); (iv) it is not satisfied by indiscriminate weight perturbation — Ablation C(i) re-runs MANSU on a same-size *random* circuit as a non-causal control, and CAD collapses by $\sim 35\%$ relative to the EAP-IG circuit ($1.143 \rightarrow 0.743$ on WMDP-bio). High CAD therefore tracks collapse of a *causally identified* pathway, not simply that some parameters changed — the diagnostic distinction between "circuit attribution reorganized" and "model broken everywhere."

Independent validation of CAD (probes and patching). CAD reuses EAP-IG, the same attribution family used to select $\mathcal{C}$, so a high score could in principle reflect attribution instability rather than genuine structural change (Section 4.1). Two methods sharing no machinery with EAP-IG corroborate

it. *Linear probes* (4-way logistic regression predicting the correct WMDP-bio answer from last-token MLP-output activations, pre- vs. post-unlearning) show answer-label decodability dropping substantially at circuit layers (14: $0.667 \to 0.490$; 29: $0.733 \to 0.500$; 30: $0.733 \to 0.517$; 31: $0.717 \to 0.417$) while matched non-circuit control layers change minimally (layer 10: -0.013; layer 20: -0.007). *Vanilla activation patching* (forward hooks, no gradients) shows normalized causal effects of 0.11–1.05 at circuit layers against -0.22 to 0.02 at non-circuit controls. One complication: patching effects at layer 14 grow modestly after unlearning, and re-attribution on θ' retains four of the original top-5 layers while four new layers enter the top-10 — evidence of structural reorganization of the circuit rather than clean dismantling, which does not affect any reported CAD value (attribution-mass change is large and positive under either interpretation).

AS-C / AS-NC: full diagnostic. CAD measures circuit-level attribution *shift*; the activation-level companions AS-C and AS-NC measure mean activation shift inside $\mathcal{C}$ and on residual-stream positions outside $\mathcal{C}$ respectively (Eq. 12). The diagnostic of localization is the concentration ratio AS-C/AS-NC: a localized structural intervention drives in-circuit activations far more than out-of-circuit activations, while a globally diffuse intervention shifts both by comparable amounts. For global-intervention baselines, AS-C and CAD coincide numerically (Table 21): when every circuit edge is perturbed in proportion to its parameter change, the in-circuit activation shift tracks the edge-attribution shift, so the two metrics quantify the same effect through different observables. The independent diagnostic value of AS-C therefore lies in its *gap* from CAD — present only for localized methods such as MANSU — together with the AS-C/AS-NC ratio. CAD, the AS-C/AS-NC ratio, and the random-circuit control of Ablation C(i) jointly turn "did the model truly forget?" from a question that behavioral evaluation cannot answer into a quantitative test.

High CAD without localization: SimNPO on MUSE. SimNPO on MUSE attains CAD $=$ 1.979 together with elevated AS-NC (1.104) and reduced MMLU. High CAD here reflects global representational damage rather than localized erasure, and is correctly flagged by AS-NC and the random-circuit control of Ablation C(i) rather than by CAD alone (cf. Table 2). This is the diagnostic distinction Section 4.1 refers to: high CAD only indicates localized attribution change when paired with low AS-NC.

SSD ablation, cross-domain (extended). The Ablation E pattern (Section 6) holds on chem/cyber: SSD+floor's Δ_{PTQ} is $+0.030/-0.050$ against MANSU's $-0.027/-0.010$, and MMLU trails MANSU's by 0.02–0.14 throughout.

Table 21: **Mechanistic-change metrics** (companion to Table 2). **CAD (↑)** (Eq. 7): relative collapse of EAP-IG attribution mass on the original forget circuit; $\to 1$ = full collapse, $\to 0$ = circuit intact (LUNAR-style redirection: empirically ≈ 0.03–0.05 across all WMDP/MUSE cells; near-zero by construction since LUNAR edits a single MLP projection *outside* the EAP-IG forget circuit). **AS-C / AS-NC (↓)** (Eq. 12): activation shift inside / outside $\mathcal{C}$. Evidence of mechanistic change requires high CAD *and* the gap AS-C $\ll$ CAD (present only for localized methods); for global baselines AS-C $=$ CAD numerically.

	WMDP-bio			*WMDP-chem*			*WMDP-cyber*			*MUSE*		
Method	CAD↑	AS-C	AS-NC↓	CAD↑	AS-C	AS-NC↓	CAD↑	AS-C	AS-NC↓	CAD↑	AS-C	AS-NC↓
Llama-3.1-8B-Instruct												
Zero-shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Global GA	0.523	0.523	0.311	0.400	0.400	0.282	0.440	0.440	0.358	1.660	1.660	1.187
Surgical GA	0.321	0.321	0.173	0.356	0.356	0.172	0.248	0.248	0.185	1.599	1.599	0.469
NPO	0.849	0.849	0.509	0.836	0.836	0.570	0.356	0.356	0.336	1.635	1.635	1.150
SimNPO	1.433	1.433	0.870	1.351	1.351	0.875	1.523	1.523	1.033	1.979	1.979	1.104
GU+SimNPO	1.292	1.292	0.824	1.292	1.292	0.833	1.366	1.366	0.984	1.522	1.522	1.256
LUNAR	0.041	1.187	0.312	0.033	0.974	0.256	0.029	0.897	0.236	0.045	1.248	0.328
MANSU (ours)	**1.143**	**0.412**	**0.138**	**1.097**	**0.398**	**0.141**	**1.118**	**0.387**	**0.143**	**1.671**	**0.318**	**0.097**
Qwen-3-8B												
Zero-shot	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Global GA	1.021	1.021	0.660	1.057	1.057	0.789	0.782	0.782	0.569	1.106	1.106	0.721
Surgical GA	0.663	0.663	0.356	0.612	0.612	0.325	0.262	0.262	0.208	1.090	1.090	0.568
NPO	0.847	0.847	0.550	0.773	0.773	0.551	0.624	0.624	0.500	1.230	1.230	0.770
SimNPO	0.911	0.911	0.609	0.966	0.966	0.724	0.630	0.630	0.385	1.299	1.299	0.828
GU+SimNPO	0.792	0.792	0.544	0.793	0.793	0.543	0.771	0.771	0.544	1.112	1.112	0.757
LUNAR	0.039	1.143	0.301	0.031	0.937	0.247	0.027	0.863	0.227	0.043	1.201	0.316
MANSU (ours)	**1.089**	**0.531**	**0.224**	**1.041**	**0.487**	**0.198**	**1.053**	**0.478**	**0.201**	**1.143**	**0.441**	**0.182**

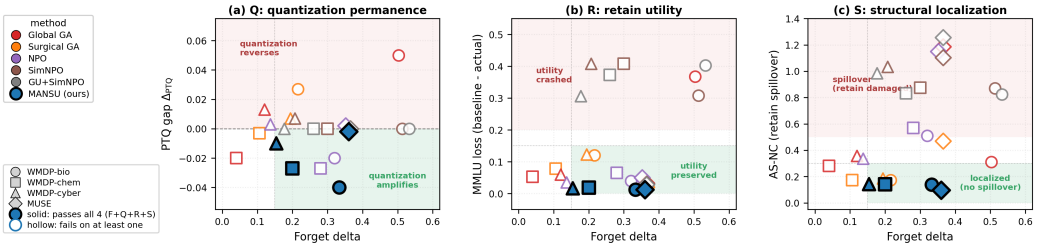


Figure 3: **The four-property scorecard, decomposed.** x: forget delta (all $24 = 6{\times}4$ weight-edit method-dataset cells; LUNAR excluded as representation redirection); y varies per panel: **(a)** Δ_{PTQ}, **(b)** MMLU loss, **(c)** AS-NC. Bottom (green) is desired in all three. Marker colour = method, shape = dataset; **solid** = passes all four thresholds (F $\geq 30\%$, Q ≤ 0, R ≤ 0.15, S ≤ 0.30) – visualization conventions, not statistically calibrated success criteria – **hollow** = fails at least one. MANSU is solid in every panel except Qwen-3-8B WMDP-bio, where forget reduction is 23% (zero-shot $0.803 \rightarrow 0.617$), just below the 30% threshold used here; the joint quantization/utility properties still hold on that cell. No other method is solid across more than a handful of cells.

Scope of MANSU evaluations. MANSU is reported on Llama-3.1-8B-Instruct and Qwen-3-8B for the main table and the companion structural-metrics table, and on six additional model variants in the architecture-independence sweep (Tables 11, 12 and 13).

Ablation A (no magnitude floor). EAP-IG circuit + null-space projection, no floor rescaling. The gradient is correctly projected and concentrated, but sub-floor updates are too small to reliably cross an NF4 boundary. *Confirmed:* BF16 forget rises to 0.513 ($+0.083$ vs. full MANSU) and $\Delta_{\text{PTQ}} = -0.008$, approaching zero — the floor's quantization benefit is clearly visible even at this margin; without it, NF4 robustness degrades by $5\times$. MMLU drop is 0.117 ($\Delta\text{MMLU} = 0.117 - 0.075 = +0.042$ worse), confirming that the floor interacts with the null-space projection to limit retain damage. See Table 3.

Ablation B (no null-space projection). EAP-IG circuit + magnitude floor, raw gradient. Without projection, gradient updates within $\mathcal{C}$ follow the raw forget-set gradient, including components along retain-sensitive directions. *Confirmed:* MMLU drops to 0.449 ($\Delta = -0.154$ vs. baseline; $+0.079$ worse than full MANSU), while $\Delta_{\text{PTQ}} = -0.019$ — the floor remains active but retain damage is severe. This is the direct empirical test of Theorem 1: null-space projection is necessary for acceptable retain-set preservation; the floor alone cannot compensate.

Ablation C (random circuit, same $|\mathcal{C}|$). Replace EAP-IG circuit with a uniformly random selection of 5 MLP layers (seed 42), matching the canonical circuit size ($\mathcal{C} = \{30, 14, 31, 19, 29\}$). Direct test of Lee et al. (2025) and Guo et al. (2025). *Confirmed:* BF16 forget rises to 0.500 ($+0.070$ vs. full), CAD collapses from 1.143 to 0.743 (-35% relative), and $\Delta_{\text{PTQ}} = -0.024$ — narrower than full MANSU's -0.040. Ablation C(ii) with the bottom-k inverse circuit (layers with *lowest* EAP-IG attribution mass) further degrades BF16 forget to 0.551 and flips Δ_{PTQ} positive ($+0.028$), confirming that circuit *identity*, not merely size, drives both forget depth and quantization robustness. Together, C(i) and C(ii) provide a controlled test of localization in our setting: matched random and inverse circuits of identical size perform worse on both axes, indicating that circuit identity contributes materially to MANSU's results here, complementing rather than settling the broader question raised by Lee et al. (2025).

Ablation D (global null-space + floor). Apply null-space projection and floor globally rather than within $\mathcal{C}$. *Confirmed:* BF16 forget accuracy rises to 0.697 — global projection disperses the forget gradient over all parameters, reducing per-parameter update magnitude below δ_i for most parameters and undermining both forget depth and quantization robustness ($\Delta_{\text{PTQ}} = +0.013$, positive). This isolates the contribution of circuit localization to the overall result and shows that the floor constraint alone, without concentrated circuit-level application, is insufficient to achieve negative PTQ gap.