EXPOSURE LEAVES A TRACE
A PROVENANCE AUDIT FOR MACHINE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a provenance audit that tests whether an unlearning method removed the *effect* of training on the forget examples, rather than only suppressing their outputs. Behavioral parity with the masked reference cannot distinguish suppression from removal. The audit has two stages. First, the exposure-subspace stage constructs a stream-matched training contrast: two models start from one checkpoint and follow the identical training stream, differing only in the loss weight of the forget examples. We define their activation difference as the exposure effect under this matched training intervention, which controls for shared initialization, data order, and training randomness. Removal and insertion show that a low-rank exposure subspace causally affects target behavior under the specified interventions across Llama-3.2-1B, Qwen2.5-1.5B, and Qwen2.5-7B. In the controlled setting, the exposure subspace does not show a pooled advantage over a target-predictive control, as expected when target-correlated activation can coincide with the exposure effect because the forget examples are the target's only source. On pretrained models, where target-predictive directions can also reflect base knowledge, the exposure subspace transfers more behavior. Across seven core unlearning objectives, all six that reach behavioral parity leave a post-unlearning residual whose transplantation into the masked reference recovers 0.15–0.42 of the suppressed behavior. The residual norms remain comparable in scale to retain-only and replace-relabel controls that do not reach parity. An RMU variant supplied with oracle entanglement labels reduces the measured audit diagnostics to the matched-random level, albeit with a large activation displacement. Behavioral parity and removal of the measured training effect are therefore distinct outcomes.

1 INTRODUCTION

In practical deployments, a model provider that fine-tunes on customer-provided data may later need to respond to a withdrawal or deletion request, creating a need for auditable procedures that assess whether the withdrawn data continues to influence the trained model (Achille et al., 2023; European Parliament and Council of the European Union, 2016; Federal Trade Commission, 2021; Information Commissioner's Office, 2026). The same question arises in an internal audit before release, when a team must confirm that a retracted or mislabeled batch has been removed. Machine unlearning is the tool for both cases, and the benchmarks that certify it judge the model from outside. A method passes when the model stops producing the target answers and keeps its performance on retain data (Li et al., 2024; Maini et al., 2024; Shi et al., 2025). For the provider, the question is different. Did the effect of the data leave the model, or did it leave only the outputs?

Output behavior cannot answer this question. Answers that appeared forgotten return after light relearning or adversarial prompting (Deeb & Roger, 2024; Hu et al., 2024; Lynch et al., 2024). Existing internal tools cannot attribute such effects to the forget data. A probe that finds a representation predicting the target answer cannot say whether that representation came from the forget examples, from the base model, or from retain data that supports the same answer (Chen et al., 2025; Cosma & Finke, 2026; Lee et al., 2026; Yong et al., 2026). Comparing a model trained with the data against one trained without it mixes the effect of the data with optimization drift between two runs (Yang & Yeung, 2026; Zhang et al., 2023). What the provider needs is a measurement of the internal change caused by the forget data.

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.

We build that measurement from the checkpoint, data, and training script. Two copies process the same minibatches in the same order with shared randomness (Figure 1); forget examples retain their positions but receive loss weight one in M_{full} and zero in M_{masked}. Under this matched training intervention, we define their activation difference as the *exposure effect*: the total activation change associated with enabling the forget-example loss, including the later changes it induces in updates on retain data. This definition is specific to the matched training process; it does not posit a universal causal representation of the forget examples. Because the two runs share the starting model and retain stream, the contrast controls for these shared sources rather than selecting directions solely because they predict the target.

The instrument is validated before it is used. A low-rank subspace fitted on one split of prompts captures a component of this intervention-defined exposure effect. Removing this component from M_{full} closes most of the target advantage and inserting it into M_{masked} restores most of it. Matched random directions and directions chosen only to predict the answer transfer far less, a fact's own component transfers while other facts' components do not, and a basis fitted on other seeds transfers into a held-out seed. The result holds on Llama-3.2-1B, Qwen2.5-1.5B, and Qwen2.5-7B, where a second intervention site is needed to recover most of the effect.

The audit can also register a measured reduction. An RMU variant uses ground-truth labels to exclude retain examples entangled with the forget data; it brings the audit diagnostics to the matched-random level, but requires a 28–36$\times$ larger activation displacement and does not preserve every utility measure. An observable substitute does not reproduce this result. Thus the audit distinguishes behavioral matching from a measured diagnostic reduction. It applies wherever a developer can replay the stream-matched training contrast.

Our key contributions are summarized as follows:

- We introduce a provenance audit built on a stream-matched training contrast: two runs start from one checkpoint and use identical minibatches, data order, and training randomness, differing only in the forget-example loss weight. Their activation difference isolates the effect of enabling the forget-example loss under this matched training stream, while controlling for shared initialization, data order, and training randomness.
- We show that a rank-≈ 8 exposure subspace causally affects target behavior under the specified removal and insertion interventions. Removing the exposure component from the exposed model closes 75% of the target-score gap, and inserting it into the masked reference restores 65%, against 36% and 25% for energy-matched random subspaces. These intervention effects replicate on Llama-3.2-1B, Qwen2.5-1.5B, and Qwen2.5-7B (§4, Figure 2).
- We audit seven core unlearning objectives. The six that reach behavioral parity leave post-unlearning residuals with 5.9–12.3$\times$ the energy of a matched random control, and transplanting those residuals into the masked reference recovers 15–42% of the suppressed behavior. Their residual norms remain comparable in scale to retain-only and replace-relabel controls that do not reach parity. An RMU variant supplied with oracle entanglement labels reduces the measured audit diagnostics to the matched-random level, albeit with a large activation displacement (§4.2–§4.3, Table 3).

2 The Provenance Audit

Figure 1 summarizes the stream-matched training intervention, exposure-subspace validation, and residual audit. Table 1 defines the notation.

Let M_0 be the base model and let $D = D_r \cup D_f$, where D_r contains the retain data and D_f contains the forget data. Training on both sets produces M_{full}. A second run, M_{masked}, follows the same training procedure but sets the loss weight of every example in D_f to zero. An unlearning method starts from M_{full} and produces M_u.

The analysis concerns fine-tuning exposure from a shared checkpoint. It does not attribute knowledge acquired during pretraining. Within this scope, behavioral parity asks whether M_u matches M_{masked} on the forget task. The exposure subspace stage identifies the activation change caused by allowing

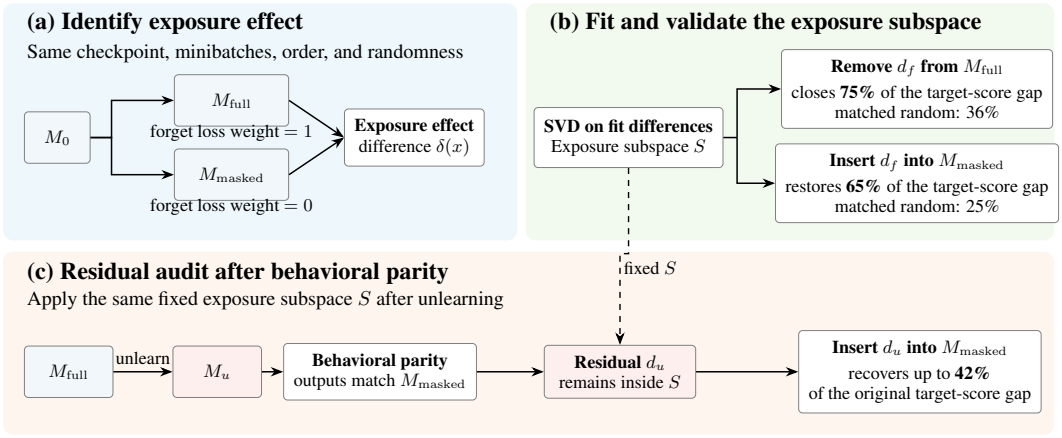

Figure 1: **The provenance audit. (a)** The stream-matched training contrast varies only the forget-example loss weight while controlling for shared initialization, data order, and training randomness; the resulting activation difference is defined as the exposure effect under this intervention. **(b)** A low-rank exposure subspace is fitted to these exposure effects and tested by removal and insertion interventions. **(c)** After the unlearned model reaches behavioral parity with M_{masked}, the same fixed exposure subspace is used to test whether a behaviorally usable residual remains.

Table 1: **Notation.**

M_0; D_r, D_f	checkpoint; retain and forget data	$g(x)$	forget-score gap to M_{masked}
M_{full}, M_{masked}	forget-loss weight 1 / 0	$\text{TE}(x)$	score gap under the matched intervention
M_u	model after unlearning	$T_{\text{rem}}(v;x)$	normalized removal transfer
$h(x)$, $\delta(x)$	activation; exposure effect	$T_{\text{ins}}(v;x)$	normalized insertion transfer
Δ_{fit}; S, P_S	fit contrasts; subspace and projector	$d_u(x)$	projected post-unlearning residual
$d_f(x)$	projected exposure component	$a(x)$, $\varepsilon_\perp(x)$	alignment; orthogonal remainder
$S^{(-r)}$	basis fitted without seed r	$\text{EF}_S(x)$, $I(x)$	energy fraction; interaction

D_f to affect training. The residual audit stage asks whether M_u retains a component of this change and whether that component can still influence the forgotten behavior.

2.1 Stage 1: Exposure Subspace

Let r index a training seed and let ω_r collect the initialization, data order, and training randomness used for that seed. Let $z \in \{0, 1\}$ indicate whether D_f contributes to training. Here, $z = 1$ enables the forget-example loss and $z = 0$ masks it. We write $M^{\omega_r}(z)$ for the model obtained under these choices.

For a minibatch b with $|b| = B$, both runs use the per-example weight $w_i(z)$ and loss

$$w_i(z) = \begin{cases} z, & i \in D_f, \\ 1, & \text{otherwise}, \end{cases} \qquad \mathcal{L}_b(z) = \frac{1}{B} \sum_{i \in b} w_i(z)\, \ell_i.$$

The minibatches, denominator, and random choices are shared. Only the contribution of the forget examples changes.

For layer q and token position p, we define the exposure effect under the stream-matched training intervention for seed r as

$$\delta^{(r)}_{q,p}(x) = h_{q,p}(M^{\omega_r}(1), x) - h_{q,p}(M^{\omega_r}(0), x)\,. \tag{1}$$

Equation (1) compares the two outcomes of the stream-matched training intervention. It captures the total effect of enabling the forget-example loss under this training process, rather than the direct effect of a single update, the effect of training on D_r alone, or a universal representation of the forget examples. In the experiments,

$$M^{\omega_r}(1) = M_{\text{full}}, \qquad M^{\omega_r}(0) = M_{\text{masked}}.$$

The layer and position subscripts are omitted when they are clear from context. Activations are measured at the final prompt token in the controlled setting and immediately before the first divergent answer token in TOFU.

We also use two reference models in sensitivity analyses. M_{retrain} is trained only on D_r. In M_{replace}, each forget row is replaced by a fact with the same relation and sentence template, a new nonce entity, and a different object of the same type. The replacement has the same token length as the original row, preserving its batch position and sequence length. Appendix Table 21 gives an example.

We estimate a shared subspace from the exposure effects on fit prompts. Let $\mathcal{X}_{\text{fit}}$ contain the fit paraphrases for all facts. For seed r, their differences form

$$\Delta_{\text{fit}}^{(r)} = \left[\delta^{(r)}(x)^\top\right]_{x\in\mathcal{X}_{\text{fit}}} \in \mathbb{R}^{|\mathcal{X}_{\text{fit}}|\times d}.$$

The first k right singular vectors of $\Delta_{\text{fit}}^{(r)}$ span the exposure subspace $S^{(r)}$. The decomposition pools all fit prompts, so $S^{(r)}$ is shared across facts rather than fitted separately for each one. For prompt x, the projected exposure component is $d_f^{(r)}(x) = P_{S^{(r)}}\delta^{(r)}(x)$. The layer and rank are selected on validation data and fixed before evaluation on the sealed test split. Appendix F.1 and Table 20 report the selection procedure and hyperparameters.

We test the subspace by removing $d_f^{(r)}(x)$ from M_{full} and inserting it into M_{masked}:

$$\widetilde{h}_{\text{full}}(x) = h_{\text{full}}(x) - d_f^{(r)}(x), \qquad \widetilde{h}_{\text{masked}}(x) = h_{\text{masked}}(x) + d_f^{(r)}(x).$$

Removal asks how much this component supports the target answer in M_{full}. Insertion asks how much of the same behavior it restores in M_{masked}.

Let $\bar{\ell}_M(o \mid x)$ be the mean token log-likelihood of answer o. For target answer o^* and matched distractors $o_1, \ldots, o_m$, define

$$\begin{aligned} \bar{\ell}_M(o \mid x) &= \frac{1}{|o|}\sum_{t=1}^{|o|} \log p_M(o_t \mid x, o_{<t}), \\ s_M(x) &= \bar{\ell}_M(o^* \mid x) - \log\left[\frac{1}{m}\sum_{j=1}^{m}\exp\big(\bar{\ell}_M(o_j \mid x)\big)\right]. \end{aligned} \tag{2}$$

Behavioral parity is measured by

$$g(x) = s_{M_u}(x) - s_{\text{masked}}(x). \tag{3}$$

A value near zero means that the unlearned model matches the masked reference on that prompt. The validation-level parity rule used to select checkpoints is given in Appendix C.

To describe activation interventions, $M \ominus v$ denotes model M with v subtracted at the selected activation site, while $M \oplus v$ denotes the corresponding insertion. The behavioral exposure effect and the two transfer measures are $\text{TE}(x) = s_{\text{full}}(x) - s_{\text{masked}}(x)$ and

$$T_{\text{rem}}(v; x) = \frac{s_{\text{full}}(x) - s_{\text{full}\ominus v}(x)}{\text{TE}(x)}, \qquad T_{\text{ins}}(v; x) = \frac{s_{\text{masked}\oplus v}(x) - s_{\text{masked}}(x)}{\text{TE}(x)}. \tag{4}$$

$T_{\text{rem}}(v; x)$ reports the fraction of the behavioral exposure effect removed from M_{full}. $T_{\text{ins}}(v; x)$ reports the fraction recovered in M_{masked}. A value of zero indicates no transfer, while a value of one indicates complete removal or recovery of the score difference. The main subspace test evaluates these quantities with $v = d_f^{(r)}(x)$. Full-vector insertion, obtained by setting $v = \delta^{(r)}(x)$, gives the empirical amount of transfer available at the selected site.

For the donor test, we extract $d_f(x_{\text{donor}})$ from one paraphrase and insert it while evaluating a held-out paraphrase x_{target} of the same fact. The cross-fact control instead uses a component extracted from a different fact. This test asks whether the intervention transfers across prompt realizations of the same fact, rather than merely replaying the prompt-specific $\delta(x)$.

To test stability across training runs, we also construct a leave-one-seed-out subspace

$$S^{(-r)} = \mathrm{span}\left(V_{1:k}\left(\left[\Delta_{\mathrm{fit}}^{(s)} \right]_{s \neq r} \right) \right),$$

where $V_{1:k}$ denotes the first k right singular vectors. This basis is fitted without seed r and evaluated on that held-out seed. Transfer above a matched random subspace indicates that the estimated structure is reproducible across seeds. Because the runs share M_0, this experiment measures stability under a shared initialization rather than identifiability across independently initialized models.

We compare the exposure subspace with MEANDIFF to test whether target-related activation structure alone can explain the transfer. MEANDIFF is estimated from the mean activation difference between target and matched control prompts in M_{full} and does not use the stream-matched training contrast. It and the random controls use the same layer, rank, and intervention procedure as the exposure subspace. Appendix F.2.3 provides the complete baseline definitions.

2.2 STAGE 2: RESIDUAL AUDIT

The residual audit uses checkpoints selected by their validation forget score. It compares M_u with M_{masked} inside the exposure subspace, then tests whether that difference changes behavior when transplanted. Methods that miss parity are reported separately.

For an unlearned model M_u, define

$$d_f(x) = P_S \delta(x), \qquad d_u(x) = P_S(h_u(x) - h_{\mathrm{masked}}(x)). \tag{5}$$

Seed superscripts are omitted. Define the unprojected difference $\Delta_u(x) = h_u(x) - h_{\mathrm{masked}}(x)$, so $d_u = P_S \Delta_u$. The residual energy fraction is $\mathrm{EF}_S(x) = \|d_u(x)\|_2^2 / \|\Delta_u(x)\|_2^2$; its control uses a random subspace of the same rank. Signed alignment with the original exposure effect is

$$a(x) = \frac{\langle d_u(x), d_f(x) \rangle}{\|d_f(x)\|_2^2}. \tag{6}$$

A positive value means that the residual has a component along the original exposure effect. A negative value indicates an oppositely directed component. If needed, the component orthogonal to the exposure effect within S is $\varepsilon_\perp(x) = d_u(x) - a(x) d_f(x)$.

Alignment does not show whether the residual can affect the model's output. The residual-transplant test therefore inserts $d_u(x)$ into the corresponding activation of M_{masked} and reports $T_{\mathrm{ins}}(d_u(x); x)$. Positive insertion transfer means that the post-unlearning residual can recover part of the behavior associated with the original exposure.

The primary audit is conducted at behavioral parity. Checkpoint selection, trajectory analyses, auxiliary interventions, and collateral measurements are described in Appendices F.1 and C.

3 EXPERIMENTAL SETUP

The controlled setting separates target-answer availability from exposure to the forget examples. We use invented entities to specify the retained evidence for each target. In UNIQUE, only the forget examples state the target. REDUNDANT supplies retained paraphrases, INFERABLE supplies premises that entail it, and SHARED_NEIGHBOR supplies related facts without the target proposition. MULTIFACT assigns separate exposure interventions to facts about the same entity. Both paired runs receive the same retained evidence. Their difference measures the additional effect of enabling the forget loss, even when that evidence already makes the target answer available to the masked model.

4 RESULTS

4.1 DOES THE EXPOSURE SUBSPACE CARRY THE BEHAVIOR?

Takeaway. The exposure subspace carries fact-specific behavior beyond matched random controls. Its advantage over target-predictive controls is supported on pretrained models, but not uniformly in the small controlled model.

Table 2: **Setup at a glance.**

Models	Llama-3.2-1B, Qwen2.5-1.5B, Qwen2.5-7B (full fine-tuning)
Data	Invented entities with constructed facts, so that support in D_r is known exactly. Conditions vary that support: UNIQUE (none), REDUNDANT (paraphrases), INFERABLE (implying premises), SHARED_NEIGHBOR (same entity and relation, different answer), MULTIFACT (two forget facts per entity, separately masked)
Stream-matched runs	Forget loss weight 1 in M_{full} and 0 in M_{masked}
Objectives	Core: GA, GradDiff, NPO, SimNPO, IDK, SCRUB, RMU. Extended: WGA, unlikelihood, IDK-DPO, DareU. Controls: retain-only, replace-relabel
Controls	Matched random subspaces (rank and energy), MEANDIFF (target-predictive), held-out-seed basis, cross-fact donors, and eight negative controls
External data	Controlled exposure injected into TOFU and MUSE

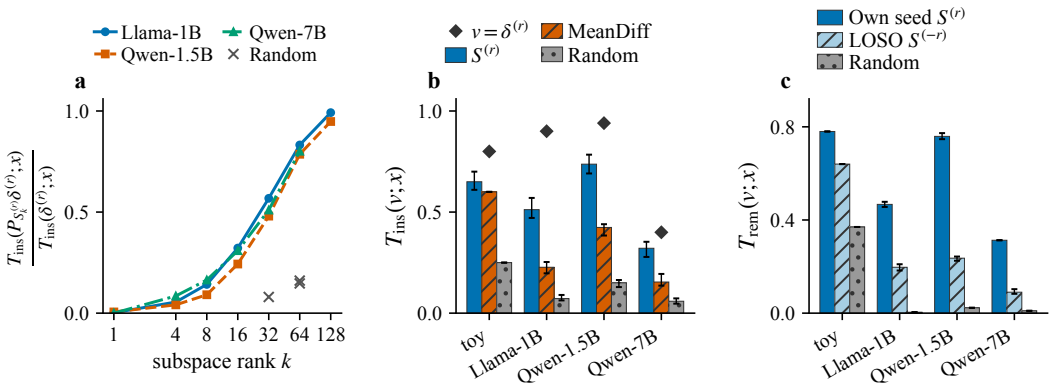

Figure 2: **Validation of the exposure subspace.** (a) Insertion across ranks, normalized by full-vector insertion. (b) Exposure, MEANDIFF, and matched-random insertion across model scales. (c) Removal using own-seed $S^{(r)}$ and leave-one-seed-out $S^{(-r)}$ bases. Patterns distinguish estimators independently of color. Intervals are descriptive 95% fact-bootstrap intervals. Pretrained-model bars show seed ranges.

Removal and insertion validate the measured effect. In the small controlled model, removal closes 75% of the target-score gap and insertion restores 65% at the validation-selected rank of about eight. Both exceed matched random controls (Table 5). Separately selected subspaces transfer behavior on Llama-1B and Qwen-1.5B/7B as well (Figure 2a,b and Table 6). Full-vector replacement provides a site-level reference, not a guarantee that one site contains the entire exposure effect.

Target prediction alone does not identify training exposure. MEANDIFF estimates target-related structure from M_{full} without the stream-matched training contrast. In UNIQUE, the forget examples are the only source of the target proposition, so predictive and exposure-related directions can overlap. REDUNDANT and INFERABLE provide alternative support through retained data. The controlled-scale comparison does not establish a pooled advantage over the target-predictive baselines. The condition-by-estimator interaction is also consistent with zero, and 13 of 16 crossed interactions fall within the stated equivalence bound (Appendix B.2). On the pretrained models, the exposure subspace transfers more behavior than MEANDIFF (Figure 2b). The stream-matched training contrast defines which activation changes result from enabling the forget loss, rather than selecting directions solely because they predict the target.

The component transfers across paraphrases of the same fact and is partly reproducible across runs. A component extracted from one paraphrase transfers to held-out paraphrases of the same fact, whereas components extracted from other facts do not. This cross-paraphrase result argues against the intervention merely replaying a prompt-specific $\delta(x)$. A basis fitted without the test seed also exceeds matched random, although it transfers less than the seed's own basis (Figure 2c). Together, these tests support fact specificity and stability under a shared starting checkpoint. Transfer across independent initializations is weaker (Appendix B). At 7B, the limits of single-site recovery motivate the multi-site analysis in Section 4.3.

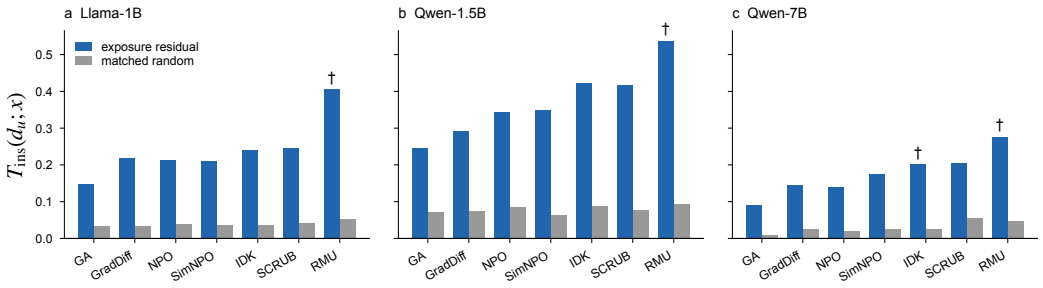

Figure 3: **Post-unlearning residual transfer across model scales.** Insertion transfer $T_{\text{ins}}(d_u(x); x)$ for each unlearned model's projected post-unlearning residual transplanted into M_{masked} (blue), compared with matched-random controls (gray). † marks methods that do not reach behavioral parity. Results aggregate three seeds at 1B and 1.5B and one seed at 7B; corresponding intervals and endpoint diagnostics appear in Tables 3 and 10.

4.2 Does behavioral parity remove it?

Takeaway. All six parity-reaching core objectives at 1B and 1.5B leave aligned, behaviorally usable residuals: insertion recovers 15–42% of the original target-score gap.

Output matching leaves a residual that can restore behavior. Table 3 reports the residual energy fraction and insertion transfer separately. The energy fraction measures how much of the squared activation difference between M_u and M_{masked} lies in S. For the six parity-reaching methods, this fraction is 5.9 to 12.3 times that of rank-matched random subspaces. The difference therefore remains concentrated in the exposure subspace. Its behavioral role is tested separately by transplantation. Inserting $d_u(x)$ into M_{masked} restores 15–42% of the original target-score gap. In the decomposed subset, the exposure-aligned component restores behavior while the orthogonal component does not (Table 12). RMU misses parity in this comparison and is reported separately.

Parity-reaching unlearning leaves residuals on the same broad scale as non-parity controls. Retain-only continuation applies no forget objective, while replace-relabel trains on a matched substitute answer; neither control reaches behavioral parity. Yet their residual insertion transfer is not lower than that of the parity-reaching methods. On Llama-1B, retain-only and replace-relabel yield insertion transfers of 33% and 35%, respectively, compared with 15–25% for the parity-reaching core methods. On Qwen-1.5B, the corresponding controls yield 51% and 56%, compared with 25–42%. These controls show that large residuals can persist without successful behavioral unlearning, but they do not identify how much of the measured residual reflects generic path dependence because neither control reaches behavioral parity.

The pattern extends beyond the non-overlap prediction-loss setting. Residual energy also exceeds matched random when retain data states or implies the target answer (Figure 6). Separately, a DareU-inspired answer-emission objective leaves insertion transfer of 30% on Llama-1B and 46% on Qwen-1.5B in non-overlap conditions (Appendix C.2). This arm changes the optimization objective, but its answer-token proxy does not reproduce DareU's attribution classifier or establish successful removal of attribution.

4.3 When does the residual persist, and what reduces it?

Takeaway. The residual persists beyond the selected checkpoint and appears in the 7B and MUSE audits. Oracle-informed steering weakens the energy and transfer diagnostics, but its large displacement and utility costs prevent interpreting this as deletion.

Persistence is not confined to the stopping checkpoint. Signed alignment remains positive throughout the measured trajectories, and residual energy fractions remain above matched random (Table 13). Among parity-band probes for NPO, IDK, and SCRUB, residual insertion is positive at approximately

Table 3: **Residual audit at 1B and 1.5B.** Residual energy fractions and insertion transfer are shown alongside separate random-control columns. Brackets are 95% hierarchical bootstrap intervals over three training seeds and then facts within each sampled seed. All six unmarked methods reach behavioral parity. †RMU does not. Additional behavioral and amplification diagnostics appear in Table 10.

		Energy fraction [%]		Insertion transfer [%]	
Model	Method	EF_S	Random	$T_{ins}(d_u; x)$	Random
Llama-1B	GA	10.0	1.5	14.7 [6.8, 18.9]	3.3 [2.1, 4.3]
	GradDiff	14.4	1.5	21.9 [17.8, 29.5]	3.4 [2.0, 4.9]
	NPO	14.3	1.6	21.3 [16.4, 24.7]	4.0 [1.8, 5.0]
	SimNPO	13.8	1.6	21.1 [17.4, 26.1]	3.6 [2.8, 5.2]
	IDK	14.9	1.6	24.0 [20.9, 28.4]	3.6 [0.8, 5.3]
	SCRUB	17.2	1.4	24.5 [19.4, 33.6]	4.3 [3.1, 5.4]
	RMU†	30.8	1.5	40.7 [34.6, 46.7]	5.3 [2.7, 7.5]
Qwen-1.5B	GA	20.8	3.5	24.6 [17.9, 28.1]	7.1 [3.5, 11.6]
	GradDiff	28.4	3.8	29.1 [20.1, 36.5]	7.5 [5.3, 10.0]
	NPO	29.1	4.1	34.3 [25.5, 48.9]	8.4 [4.8, 10.7]
	SimNPO	28.0	4.0	34.8 [29.5, 40.1]	6.4 [3.7, 9.5]
	IDK	33.8	4.2	42.2 [37.0, 46.2]	8.9 [5.6, 11.8]
	SCRUB	32.8	4.2	41.7 [35.1, 51.1]	7.8 [5.6, 10.5]
	RMU†	46.6	4.2	53.8 [39.5, 63.0]	9.4 [6.0, 12.7]

98%. GA leaves the band quickly. The trajectories therefore support persistence under the tested schedules, not a claim about every possible optimization path (Figure 8).

Larger models limit single-site access. At 7B, residual energy exceeds random for all seven methods and insertion does so for six; GA shows no clear separation. A separate experiment raises projected insertion to 86–91% with a second site, supporting limited single-site access (Appendices D.2–D.3).

External data tests the controlled-exposure construction. MUSE clears the exposure-adequacy threshold, with positive residual insertion for all seven objectives (Table 15). TOFU exceeds random in 12 of 14 settings but falls below the threshold, so it remains a sensitivity check. Neither experiment audits unknown pretraining exposure.

Oracle-informed steering changes the diagnostics, not necessarily the amount stored. Removing RMU's retain anchor brings the energy fraction near random but severely damages utility. Restricting the anchor with ground-truth entanglement labels also reduces insertion recovery, with roughly 30× the original full-state displacement (Table 16). Qwen stays within the analysis-chosen MMLU margin, but perplexity rises (Table 17). Because $\|d_u\|$ can still exceed $\|d_f\|$, this demonstrates reduced diagnostics in the tested oracle setting, not deletion.

5 Related Work

Counterfactual effects of training data. Counterfactual memorization and distribution restoration compare predictions under data omission or matched training references (Yang & Yeung, 2026; Zhang et al., 2023); influence functions and TRAK estimate training-example effects on predictions (Koh & Liang, 2017; Park et al., 2023). We instead measure an activation effect of enabling the forget loss under a fixed stream. Task arithmetic edits weights between checkpoints, whereas our interventions act on activations from two fine-tuned outcomes (Ilharco et al., 2023).

Audits after unlearning. RULER compares representations with a retrained reference, while UDS replaces activations to measure recoverable knowledge (Cosma & Finke, 2026; Lee et al., 2026). Other work detects residual information or revives forgotten behavior through training and prompting (Chen et al., 2025; Gao et al., 2026; Łucki et al., 2025; Patil et al., 2024; Xu et al., 2026; Yong et al., 2026). We define and intervention-test the exposure subspace before unlearning, then use the same fixed subspace to extract and transplant the post-unlearning residual.

Attribution and data overlap. DareU reduces classifier-estimated attribution; our proxy instead detects answer emission (Lu et al., 2026). BLUR studies forget–retain overlap via combined queries and relearning (Hu et al., 2025); our controlled setting varies retained support while holding the target answer fixed.

Concept erasure and model editing. Concept erasure removes decodable information, while model editing changes fact associations (Belrose et al., 2023; Meng et al., 2022; Ravfogel et al., 2020). Because target-predictive features can reflect pretraining or retained support, we compare against target-predictive and random subspaces under the same patch. The interventions follow causal mediation and abstraction, with controlled subspace comparisons (Geiger et al., 2021; Makelov et al., 2024; Vig et al., 2020).

6 Limitations

Intervention scope. Insertion establishes accessibility under a specified patch, not spontaneous use or a unique storage location. The audit covers selected sites and a linear subspace; energy fraction must be read with residual norm and transfer (Makelov et al., 2024; Zhang & Nanda, 2024).

Training access. The stream-matched training contrast requires a shared checkpoint and replayable stream, so it does not attribute unknown pretraining exposure. Pretrained runs use warn-only determinism.

Independent initializations. Leave-one-seed-out transfer supports stability under a shared checkpoint, but partial transfer across independent initializations does not establish a universal exposure subspace.

Path dependence. Because unlearning starts from M_{full}, the audit does not fully distinguish an exposure-aligned residual from generic path dependence induced by continued optimization. Retain-only and replace-relabel partially probe this alternative, but neither reaches parity.

Data and utility. Invented entities control retained support; MUSE and TOFU use injected exposure, and TOFU falls below the adequacy threshold. The 1B/1.5B audits use three seeds, whereas the primary 7B audit uses one. Oracle-anchor utility is Qwen-specific and uses an analysis-chosen MMLU margin; perplexity worsens. No result certifies deletion.

7 Conclusion

The provenance audit links a stream-matched activation contrast to controlled fine-tuning exposure. All six parity-reaching core objectives leave an exposure-aligned residual whose insertion recovers 15–42% of the original target-score gap; non-parity controls leave residuals of comparable norm and higher transfer, but do not resolve generic path dependence. Behavioral matching can suppress the exposed behavior while leaving a substantial exposure-aligned residual. Oracle-informed RMU reduces the audit diagnostics to matched-random levels only with a large displacement. The audit therefore tests whether a training-intervention-specific effect remains accessible, without claiming spontaneous recall, universal causal representation, or deletion.

Reproducibility statement

All experiments use tracked configurations and an offline-tested pipeline. Selection uses validation data and final evaluation uses a held-out test split; falsification checks were specified in code but not externally preregistered.

AI use statement

Generative AI assisted with implementation and manuscript drafting and editing. The authors verified the code, results, and sources and take responsibility for the manuscript.

REFERENCES

Alessandro Achille, Michael Kearns, Carson Klingenberg, and Stefano Soatto. AI model disgorgement: Methods and choices. *arXiv preprint arXiv:2304.03545*, 2023.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yiwei Chen, Soumyadeep Pal, Yimeng Zhang, Qing Qu, and Sijia Liu. Unlearning isn't invisible: Detecting unlearning traces in LLMs from model outputs. *arXiv preprint arXiv:2506.14003*, 2025.

Georgina Cosma and Axel Finke. RULER: Representation-level verification of machine unlearning. *arXiv preprint arXiv:2605.27569*, 2026.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. General Data Protection Regulation; see Articles 7 and 17.

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 *NeurIPS Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=pVACX02m0p.

Federal Trade Commission. FTC finalizes settlement with photo app developer related to misuse of facial recognition technology. https://www.ftc.gov/news-events/news/press-releases/2021/05/ftc-finalizes-settlement-photo-app-developer-related-misuse-facial-recognition-technology, 2021. Accessed 2026-09-21.

Yichen Gao, Altay Unal, Akshay Rangamani, and Zhihui Zhu. An illusion of unlearning? assessing machine unlearning through internal representations. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026. arXiv:2604.08271.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Jogging the memory of unlearned LLMs through targeted relearning attacks. *arXiv preprint arXiv:2406.13356*, 2024.

Shengyuan Hu, Neil Kale, Pratiksha Thaker, Yiwei Fu, Steven Wu, and Virginia Smith. BLUR: A benchmark for LLM unlearning robust to forget-retain overlap. *arXiv preprint arXiv:2506.15699*, 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Information Commissioner's Office. How do we ensure individual rights in our ai systems? https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-individual-rights-in-our-ai-systems/, 2026. Accessed 2026-09-21.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

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

Jaeung Lee, Dohyun Kim, and Jaemin Jo. Measuring the depth of LLM unlearning via activation patching. *arXiv preprint arXiv:2605.24614*, 2026.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*, 2024.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2022.

Xinyang Lu, Jiabao Pan, Rachel Hwee Ling Sim, See-Kiong Ng, Anthony Kum Hoe Tung, and Bryan Kian Hsiang Low. De-attribute to forget for LLM unlearning. *arXiv preprint arXiv:2605.30919*, 2026.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for AI safety. *Transactions on Machine Learning Research*, 2025.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs. *arXiv preprint arXiv:2402.16835*, 2024.

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

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations (ICLR)*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning (ICML)*, 2023.

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*, 2020.

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 *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.06460.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q. Weinberger. Rethinking LLM unlearning objectives: A gradient perspective and go beyond. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.19301.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*, 2020.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in LLMs. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2505.16831.

Sen Yang and Yuen-Hei Yeung. Unlearning as distribution restoration: A controlled counterfactual study, a validated selective screen, and the limits of oracle-free certification. *arXiv preprint arXiv:2607.19442*, 2026.

Teresa Pui Yee Yong, Win Kent Ong, and Chee Seng Chan. Erased, but not gone: Output forgetting is not true forgetting. *arXiv preprint arXiv:2606.25001*, 2026.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024.

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

A EVIDENCE MAP

What the appendix adds. We give the intervention results behind the exposure subspace in Section B, then examine what remains after unlearning in Section C. Section D follows the residual through training and into larger models and external data. The last sections spell out how to interpret and reproduce these measurements.

Shared evaluation rule. Unless stated otherwise, we fit the subspace on fit prompts, select its site and rank on validation, and evaluate once on the sealed test split. Paired runs share initialization, minibatches, and random choices. Table 4 links each claim to its evidence and limit. Figures 5, 6, 9, and 7 give the separate attribution-proxy, retained-support, 7B, and sensitivity checks.

Table 4: **Claims, evidence, scope, and limits.**

Claim Behaviorally effective exposure component is compact and fact-specific	**Primary evidence** Paired intervention protocol
Scope 2 families, 1B–7B, controlled world	**Does not establish** A shared "forget representation"
Claim Residual persists at behavioral parity	**Primary evidence** Parity-reaching states among ten standard unlearning objectives
Scope Tested models, objectives, and parity rule	**Does not establish** All unlearning methods; the attribution arm is reported separately
Claim Residual persists under a DareU-inspired answer-token objective	**Primary evidence** DareU-style audit in Figure 5
Scope Closed-world answer-emission proxy, 1B and 1.5B	**Does not establish** The original DareU pipeline or an oracle-free result
Claim Residual is accessible under intervention	**Primary evidence** Transplant, amplification, and matched controls
Scope Selected sites and subspaces	**Does not establish** Spontaneous use in ordinary inference
Claim Persistence is not explained by tested stopping points	**Primary evidence** Trajectory audit in Table 13
Scope 4 methods, 3 seeds, 2 families	**Does not establish** All optimization paths
Claim Residual energy fraction and insertion transfer can be reduced	**Primary evidence** Oracle-informed RMU variant
Scope Oracle labels, Qwen MMLU margin, worsened perplexity	**Does not establish** A zero residual, general utility preservation, or deletion
Claim External replication	**Primary evidence** Controlled-exposure TOFU and MUSE
Scope Modified paired settings	**Does not establish** Native benchmark unlearning
Claim 7B scaling	**Primary evidence** Seven-method audit and sealed multi-site intervention
Scope One seed for all seven methods; three for exposure tests and GA/NPO/RMU	**Does not establish** Method-uniform recovery or independent causal shares

B THE EXPOSURE SUBSPACE TRANSFERS BEHAVIOR

The exposure component affects target behavior under intervention. Removal and insertion in Table 5 establish this in the controlled setting. Table 6 repeats the interventions on pretrained models and with altered training setups. The effect is specific to the forget fact and exceeds matched random subspaces. Comparisons with target-informed controls vary by model scale. Full-vector replacement shows what can be reached at the chosen site; it is not a mathematical bound.

Controlled-scale transfer. At the selected rank, about eight, removal and insertion transfer 75% and 65% of the exposure effect. Matched random subspaces transfer 36% and 25%. The difference is significant in each of the three high-TE conditions at $p \leq 0.003$.

Reference and seed sensitivity. Removal transfer is 76%, 79%, and 71% with the masked, replace, and retrain references. Subspaces from runs sharing M_0 overlap by 0.51, against 0.08 for random. A basis fitted without one seed transfers 64% in that seed, compared with 78% for its own basis and 37% for random. Independent initializations reduce insertion transfer to 49%, against 32% for random. Shared-initialization stability does not establish an initialization-independent basis.

Table 5: **Controlled-scale intervention effects.**

Matched activation intervention	T_{rem} [%]	T_{ins} [%]
Total exposure effect TE (margin units)	7.90 [6.18, 9.32], n=117	
Full-vector replacement at the site (empirical reference)	96 [91, 104]	80 [76, 83]
Exposure subspace, rank 8	**75** [71, 80]	**65** [61, 70]
Random subspace, matched rank+energy	36	25
Shuffled-label control (SHUFFLED), matched rank	52	49

The pattern extends to pretrained models. Table 6 gives the selected depth and rank for each model. Exposure-based removal exceeds both matched random and MEANDIFF removal in every row, while insertion is reported on its own. Components transfer across paraphrases of the same fact but not across facts. The extra conditions double the background corpus, merge rank16 LoRA weights, use pure-forget minibatches at matched TE, or lengthen nonce stems. The 7B row covers UNIQUE and SHARED_NEIGHBOR.

Table 6: **Cross-family replication against the masked reference.**

	ℓ^*/L	k^*/d	transfer [%]						
			full vec.		exposure		rand.	MEANDIFF	donor
run			T_{rem}	T_{ins}	T_{rem}	T_{ins}	T_{rem}	T_{rem}	same-f./rel.
Llama-3.2-1B	.94	.016	87	90	43	51	13	12	89 / −13
Qwen2.5-1.5B	.96	.042	94	94	78	74	26	35	90 / −23
Llama-1B, diluted	.94	.016	84	86	42	42	15	13	84 / −10
Llama-1B, LoRA[‡]	.94	.016	88	n/a	63	55	8	22	94 / −16
Llama-1B, grouped batches	.94	.016	87	n/a	43	45	12	14	88 / −12
Llama-1B, 4-syllable nonces	.94	.016	86	88	41	44	14	13	88 / −16
Qwen2.5-7B[§]	.86	.018	41	40	28	32	12	14	38 / +4

B.1 CONTROLLED CONDITIONS

Retained-support conditions. The conditions change the evidence available outside the forget examples. In UNIQUE, the forget rows are the only source of the target. REDUNDANT adds retained paraphrases of the same proposition. INFERABLE adds retained premises that entail it. SHARED_NEIGHBOR preserves the entity and relation structure without stating the target. MULTIFACT places two independently masked facts at the same entity and token position.

Negative controls. Across the experiments, we use eight negative controls: unseen entities, style-matched prompts, unrelated retain facts, shuffled labels, matched random subspaces, permuted

positions, path-broken premises, and cross-fact donors. Each control is used where its corresponding failure mode applies.

Table 7: **Controlled-setting conditions.**

Condition	Construction	What it isolates
UNIQUE	fact appears only in D_f	exposure is the only source of the target, so predictive and exposure directions may overlap
REDUNDANT	paraphrases of the fact also in D_r	answer-source separation for redundant exposure
INFERABLE	premises in D_r entail the fact	retained inference versus direct exposure, with a path-broken control
SHARED_NEIGHBOR	retained facts share entity and relation structure	high-TE separation of the exposure effect from target correlation
MULTIFACT	one entity carries two forget facts with separate masking groups	per-fact counterfactual within one token position

The intervention survives different forms of retained support. The five conditions differ in what retain data can reveal about the target. In the three high-TE conditions, the exposure component beats matched random on removal and insertion at $p \leq 0.003$.

Target-informed comparison. Separation from MEANDIFF is strongest in MULTIFACT and appears on insertion in SHARED_NEIGHBOR; elsewhere it is absent or reversed. The conventional condition-by-estimator interaction intervals include zero: removal -12 $[-36, 13]$ and insertion $+3$ $[-8, 11]$ percentage points. After Benjamini–Hochberg correction across ten contrasts, MULTIFACT on both interventions and SHARED_NEIGHBOR on insertion remain significant. The nominal $p = .044$ reversal does not. When TE is small, insertion is more informative than a removal ratio.

Table 8: **Per-condition specificity in the power run. Entries are raw p-values.**

Condition	TE (margin)	Exposure [%] T_{rem} / T_{ins}	MEANDIFF [%] T_{rem} / T_{ins}	LDA [%] T_{rem} / T_{ins}	GRADPCA [%] T_{rem} / T_{ins}	Random [%] T_{rem} / T_{ins}	exposure vs. MEANDIFF
MULTIFACT	17.2	**76 / 70**	69 / 59	70 / 60	75 / 54	38 / 28	$T_{\text{rem}}p<.001$; $T_{\text{ins}}p<.001$, d=0.26
SHARED_NEIGHBOR	15.5	74 / **70**	69 / 60	68 / 59	59 / 60	30 / 29	$T_{\text{ins}}p=.001$, d=0.82; $T_{\text{rem}}p=.148$
UNIQUE	11.1	72 / 62	57 / 49	54 / 42	68 / 59	30 / 27	$T_{\text{rem}}p=.216$; $T_{\text{ins}}p=.068$ (trend)
REDUNDANT	3.4	94† / 52	73 / 49	74 / 52	74 / 36	59† / -2	$T_{\text{ins}}p=.346$
INFERABLE	6.4	66 / 71	69 / 70	71 / 69	66 / 62	36 / 23	T_{rem}: MEANDIFF ahead, p=.044

B.2 FALSIFICATION TESTS

Falsification result. F6 does not support a pooled controlled-scale advantage over MEANDIFF and GRADPCA. The usual condition-by-estimator interaction is also consistent with zero. A separate equivalence analysis places 13 of 16 crossed interactions within its stated margin, so these tests do not establish an advantage confined to entangled retain conditions. The pretrained-model comparison in Figure 2b shows a narrower scale-specific advantage.

Dose check. F3 finds that behavioral transfer changes with exposure dose while the measured effect norm stays nearly flat. Table 9 reports both quantities.

Table 9: **Pre-specified falsification tests and their outcomes.**

ID	Failure mode	Trigger	Status (toy)
F1	cross-seed instability	subspace overlap at the random-overlap null	pass (identity 0.51 / Procrustes 0.34 vs. 0.08)
F2	control damage	retain/neighbor margin shifts beyond the retrain-noise band under patching	pass
F3	dose insensitivity	TE or residual energy flat in exposure dose	behavioral pass: TE–dose ρ=0.256 (p=.0022), T_{rem}–dose ρ=0.285 (p=.0006); residual *norm* dose-flat (dissociation, discussed in §4)
F4	random equivalence	matched-rank+energy random subspace within CI of exposure	pass (per-condition $p \leq 0.003$ in high-TE cells; pooled ratio permutation not usable because of outliers)
F5	reference fragility	masked/replace/retrain references disagree on $T_{\text{rem}}/T_{\text{ins}}$ beyond CI	pass (T_{rem}76/79/71%)
F6	baseline equivalence	MEANDIFF–GRADPCA match exposure intervention-transfer ratios where the design decouples them	pooled controlled advantage unsupported; conventional interaction interval includes zero; 13/16 crossed interactions equivalence-bounded; pretrained-scale advantage retained
F7	reference disagreement on ratios	validation-selected site and rank selected under one reference fails under others	pass
F8	probe-only identifiability	subspace decodable (probe AUC high) but causally inert	pass (interventional)
F9	training-validity gates	masked reference learned forget facts; base model fails held-out templates; stream mismatch between conditions	pass (gated)

C BEHAVIORAL PARITY LEAVES A USABLE RESIDUAL

A behavioral match can leave the exposure component behind. We test this with the objectives in Table 11 and the intervention in Figure 4. Alignment tells us whether the remaining signal points along the original exposure effect. Insertion tests whether that signal can still change an answer. We also check retained support and sensitivity to the patch.

Parity checkpoint. Every method starts from M_{full}. We select the first checkpoint whose validation forget margin crosses the M_{masked} target. A method that never crosses is reported at its final checkpoint and marked non-parity. The sealed test split is not used to choose the stop.

Residual interventions. The primary test inserts the unlearned model's projected residual $d_u(x)$ into M_{masked}. Amplification adds $\alpha d_u(x)$ inside the unlearned model. Full-activation replacement and a new readout are secondary checks. Endpoint amplification recovery is normalized by the remaining score gap from M_u to M_{full}, rather than by TE. Reinserting $P_S(h_{\text{full}} - h_{\text{masked}})$ supplies the exposure component externally and tests downstream compatibility, not retention. For the endpoint tables and trajectories, collateral is the median absolute retain-score change relative to M_{full}. The masked reference's change relative to M_{full} is reported separately where available.

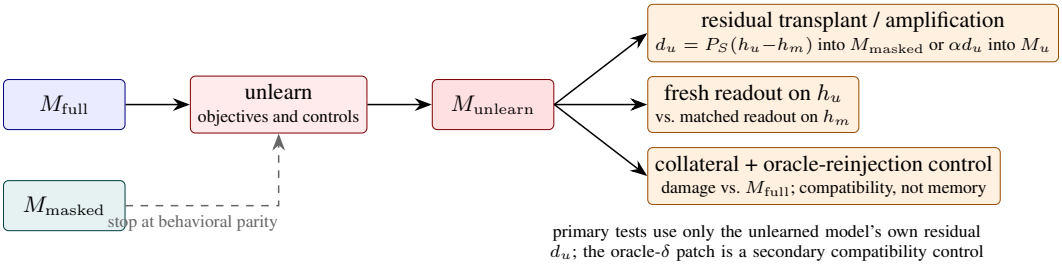

Figure 4: **Design of the parity-stopped, residual-based audit.**

Table 10: **Additional endpoint diagnostics at 1B and 1.5B.** Gap is the target margin relative to M_{masked}. Collateral is the median absolute retain-score change relative to M_{full}. EF_S is the residual energy fraction in S (rank-matched random in parentheses), $T_{\text{ins}}(d_u; x)$ is residual transplantation into M_{masked} (random in parentheses), and amplification is recovery from adding αd_u inside M_{unlearn} at α=2, normalized by its remaining gap to M_{full}. Bold marks the largest recovery among parity-reaching methods. †RMU does not reach parity.

		behavior		residual in S		recovery using d_u	
model	**method**	gap	collateral	EF_S (rand.) [%]	$\|d_u\|/\|d_f\|$	T_{ins} $(d_u; x)$ (rand.) [%]	amplification [%]
Llama-1B	GA	−0.9	16.5	10.0 (1.5)	0.89	14.7 (3.3)	9.7
	GradDiff	−1.7	7.1	14.4 (1.5)	0.85	21.9 (3.4)	17.1
	NPO	−1.9	7.9	14.3 (1.6)	0.86	21.3 (4.0)	13.5
	SimNPO	−1.1	6.1	13.8 (1.6)	0.82	21.1 (3.6)	15.9
	IDK	+1.1	3.6	14.9 (1.6)	0.87	24.0 (3.6)	12.4
	SCRUB	−1.8	8.0	17.2 (1.4)	0.90	**24.5** (4.3)	19.7
	RMU†	+9.8	0.7	30.8 (1.5)	0.96	40.7 (5.3)	−7.8
Qwen-1.5B	GA	−1.5	18.6	20.8 (3.5)	1.11	24.6 (7.1)	18.0
	GradDiff	−3.6	13.0	28.4 (3.8)	0.93	29.1 (7.5)	26.4
	NPO	−2.0	7.4	29.1 (4.1)	0.89	34.3 (8.4)	25.6
	SimNPO	−0.3	6.7	28.0 (4.0)	0.83	34.8 (6.4)	24.2
	IDK	+1.9	3.6	33.8 (4.2)	0.92	**42.2** (8.9)	18.8
	SCRUB	+1.9	5.4	32.8 (4.2)	0.90	41.7 (7.8)	16.5
	RMU†	+9.7	1.2	46.6 (4.2)	0.94	53.8 (9.4)	−20.2

Method settings. Table 11 gives the implemented objectives and their retain terms. GA has no retain term, while RMU uses an activation anchor. The core 1B/1.5B comparisons use a common parity-stop rule and five-epoch cap. The primary 7B GradDiff, SimNPO, and IDK runs receive up to 15 epochs. Their stopping status and additional archived endpoints are reported in Appendix D.2. The DareU-style arm uses an answer-emission proxy instead of DareU's attribution classifier and is evaluated separately on non-overlap conditions. Replace-relabel and retain-only are non-parity controls.

Table 11: **Unlearning-method specification.** The parity column describes the core 1B/1.5B comparisons. The 7B stopping states are reported in Appendix D.2.

method	**source**	**objective as implemented**	**lr**	**key coeff.**	**parity?**
GA	Jang et al. (2023)	$-\text{CE}_f$ (no retain term)	10^{-5}	n/a	yes (overshoots)
GradDiff	Liu et al. (2022)	$-\text{CE}_f + r_w\text{CE}_r$	10^{-5}	n/a	yes
NPO	Zhang et al. (2024)	$-\frac{2}{\beta}\log\sigma(\beta(\text{CE}_f - \text{CE}_f^{\text{ref}})) + r_w\text{CE}_r$	10^{-5}	β=0.1	yes
SimNPO	Fan et al. (2024)	reference-free NPO (β=0.1, γ=0) $+ r_w\text{CE}_r$	10^{-5}	β=0.1	yes
IDK	Maini et al. (2024)	CE of "I don't know"-relabeled answers $+ r_w\text{CE}_r$	10^{-5}	n/a	yes
SCRUB	Kurmanji et al. (2023)	$-\text{KL}_f(\text{ref}\|\theta) + r_w(\text{KL}_r + \text{CE}_r)$	10^{-5}	n/a	yes
RMU	Li et al. (2024)	steer h_f at a 3-layer window to a random control vector; anchor h_r (no CE_r)	5×10^{-5}	c=6.5, α=100	**no**
WGA	Wang et al. (2025)	token-weighted ascent ($w_t = p_t^\gamma$) $+ r_w\text{CE}_r$	10^{-5}	γ=1	partial (1/3, 0/3)
Unlikelihood	Welleck et al. (2020)	$-\log(1-p_t)$ on forget tokens $+ r_w\text{CE}_r$	10^{-5}	n/a	partial (2/3, 1/3)
IDK-DPO	Maini et al. (2024)	DPO, chosen = IDK relabel, rejected = original, $+ r_w\text{CE}_r$	10^{-5}	β=0.1	yes
DareU-style	Lu et al. (2026)	REINFORCE with reward equal to one minus target-answer emission, $+ r_w\text{CE}_r$	10^{-5}	K=8, T=1	separate arm
Replace-relabel	(control)	CE of a matched substitute answer	10^{-5}	n/a	partial (+7–9)
Retain-only	(control)	CE_r only (no forget term)	10^{-5}	n/a	partial (+6–7)

C.1 Component decomposition

Direction matters as well as size. We split the projected residual into a part parallel to the original exposure effect and an orthogonal part within S. A method can move activations far from the exposed model while leaving a smaller component pointing in the exposure direction. Table 12 tests these parts separately for the seven standard methods.

Component definitions. Equation (5) defines d_u, and Equation (6) defines its signed coefficient a along d_f. Thus $d_u(x) = a(x)d_f(x) + \varepsilon_\perp(x)$, with $\langle \varepsilon_\perp(x), d_f(x)\rangle = 0$. The cosine and $E_\parallel$ columns

describe orientation and aligned energy. Each T_{ins} column reports insertion transfer for the indicated component. The interaction is computed per prompt as

$$I(x) = T_{\text{ins}}(d_u(x); x) - T_{\text{ins}}(a(x)d_f(x); x) - T_{\text{ins}}(\varepsilon_\perp(x); x).$$

For this cached component analysis, prompt-level recovery uses a 10^{-6} denominator stabilizer and is clipped to $[-1, 2]$ before averaging within each fact and run. The reported interaction uses these same clipped transfers. This aggregation differs from the unclipped endpoint statistic in Table 3, so their magnitudes are not directly interchangeable. The transfer columns have the same sign across the measured seeds, while the matched-random transplant ranges from 4% to 9%.

Table 12: **Component-transplant decomposition. Intervals resample training runs.**

model	**method**	a	$\cos$ / $E_\parallel$	$T_{\text{ins}}(d_u; x)$ [%]	$T_{\text{ins}}(ad_f; x)$ [%]	$T_{\text{ins}}(\varepsilon_\perp; x)$ [%]	$I(x)$ [%]
Llama-1B	GA	.51 [.44, .55]	.59 / .35	16 [8, 21]	26 [22, 29]	−6 [−10, −3]	−2 [−3, −1]
	GradDiff	.66 [.62, .73]	.78 / .61	24 [18, 30]	34 [29, 40]	−5 [−8, −3]	−2 [−3, −1]
	NPO	.65 [.58, .76]	.77 / .59	22 [17, 27]	33 [29, 38]	−5 [−7, −3]	−1 [−3, 0]
	IDK	.71 [.68, .75]	.82 / .68	25 [22, 29]	36 [32, 41]	−4 [−6, −2]	−3 [−4, −2]
	SCRUB	.72 [.64, .82]	.82 / .68	26 [20, 36]	38 [30, 47]	−5 [−8, −2]	−1 [−2, 1]
	SimNPO	.64 [.55, .76]	.80 / .65	22 [18, 28]	33 [27, 39]	−5 [−8, −3]	−1 [−2, 0]
	RMU	.93 [.89, .95]	.98 / .97	43 [36, 49]	45 [39, 52]	−1 [−2, −1]	0 [−1, 0]
Qwen-1.5B	GA	.54 [.51, .62]	.53 / .28	25 [18, 29]	43 [39, 50]	−16 [−20, −13]	−3 [−5, −2]
	GradDiff	.63 [.56, .68]	.68 / .48	30 [23, 36]	48 [41, 56]	−13 [−16, −8]	−3 [−6, 0]
	NPO	.70 [.55, .79]	.82 / .67	35 [25, 50]	54 [46, 61]	−11 [−13, −8]	−2 [−4, −1]
	IDK	.76 [.73, .80]	.83 / .69	45 [38, 48]	58 [55, 62]	−5 [−7, −3]	−7 [−9, −4]
	SCRUB	.74 [.67, .81]	.82 / .68	43 [36, 53]	56 [50, 67]	−5 [−10, −2]	−5 [−7, −3]
	SimNPO	.62 [.58, .66]	.76 / .59	36 [30, 42]	49 [43, 54]	−6 [−9, −4]	−3 [−5, 0]
	RMU	.87 [.79, .90]	.94 / .88	56 [39, 65]	62 [55, 68]	−1 [−4, 0]	−3 [−5, −1]

C.2 Attribution-based unlearning

Answer-token objective. We test whether the residual also appears under a DareU-inspired policy-gradient objective. It rewards responses that do not emit the target answer token:

$$R(y, x) = 1 - \text{Emit}_f(y, x).$$

Here $\text{Emit}_f(y, x)$ is one when response y emits the target answer token for prompt x. The sampling and optimization settings are listed in Table 11.

Proxy limitation. The reward measures answer emission, not attribution to a training example. A base model or other examples could also supply the answer. We therefore use this arm only where retain data does not support the same answer and interpret it as an answer-targeted check. It is not a reproduction of DareU's learned classifier or full pipeline.

Answer-token result. On non-overlap conditions, residual insertion recovers 30.3% $[25.0, 33.7]$ on Llama-1B and 45.6% $[41.3, 49.0]$ on Qwen-1.5B. The seed-crossed intervals exclude matched random. Residual energy fractions are 14.4% versus 1.6% and 31.4% versus 4.1%. Pooled fact-by-seed permutation tests give descriptive $p = .0005$ for both comparisons. One Llama run has perplexity 218 rather than about 90, so this arm tests for a residual under a different objective without establishing retained utility.

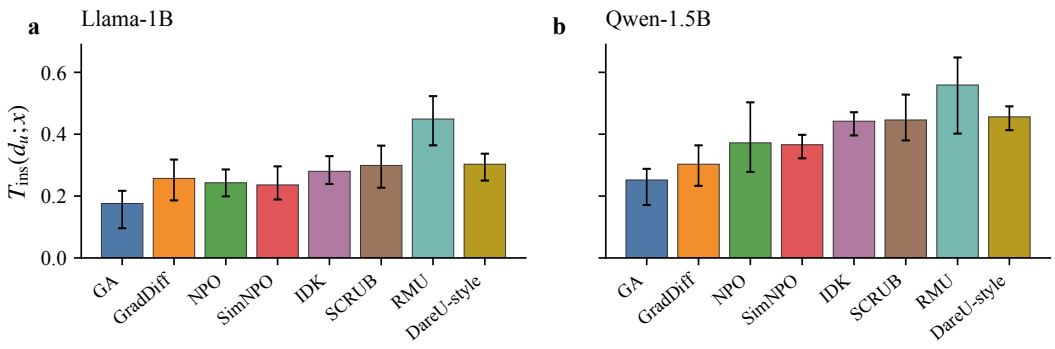

Figure 5: **Residual transplantation after the DareU-style answer-token objective.** Results use the non-overlap conditions.

C.3 Residual energy with retained support

Overlap groups. REDUNDANT and INFERABLE supply the target answer through retain data; UNIQUE and SHARED_NEIGHBOR do not. Figure 6 compares residual energy EF_S across these groups. Normalized recovery is unstable when M_{masked} already scores the target highly, so this test uses energy.

Residual under overlap. In both groups, residual energy inside S exceeds matched random for every method in both model families. The per-seed medians agree in direction. A pooled fact-by-seed sign-flip test gives descriptive $p = .0005$ for each comparison. Retained support for the same answer therefore does not explain away the measured energy, although this test alone does not establish behavioral transfer in those conditions.

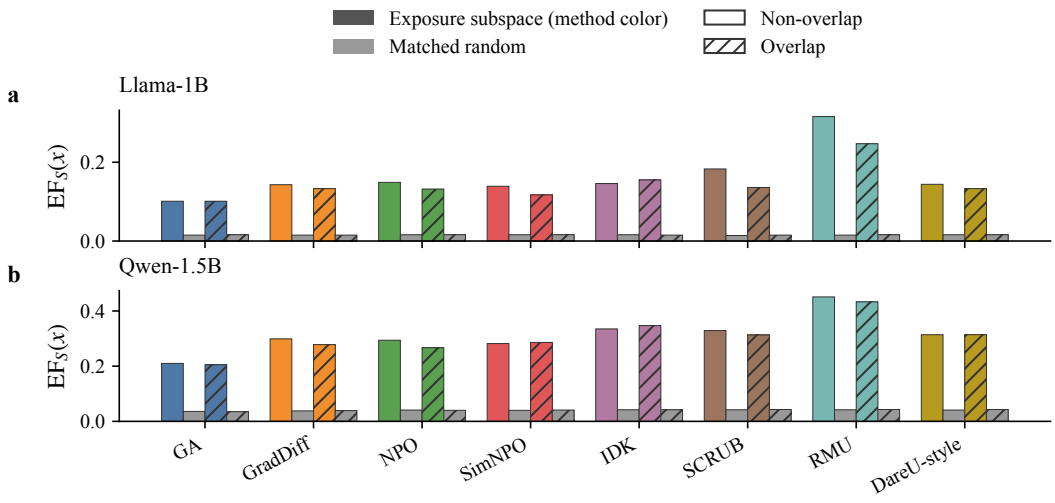

Figure 6: **Residual energy under non-overlap (solid) and overlap (hatched) conditions.**

C.4 Amplification and site sensitivity

Amplification strength. Figure 7a varies α over $\{0.5, 1, 2, 4\}$ while holding the checkpoint and residual fixed. Recovery rises through $\alpha = 2$ for parity-reaching methods and changes little beyond it. SimNPO and SCRUB decline at four. RMU is shown separately because it does not reach parity. Table 10 reports $\alpha = 2$.

Layer and rank sensitivity. Figure 7b sweeps the MUSE intervention site for GA at rank 32. Early layers are near zero; recovery begins at layer 12 and peaks around layers 13 to 15. Across the full

layer-rank sweep, residual energy exceeds matched random in 29 of 30 cells. Ranks 16, 32, and 64 all give positive recovery, so the result does not depend on one selected layer or rank.

In Figure 7, the dotted line marks the reported $\alpha = 2$ setting. RMU, shown with a dashed curve, did not reach parity. The shaded region marks the late layers selected on validation.

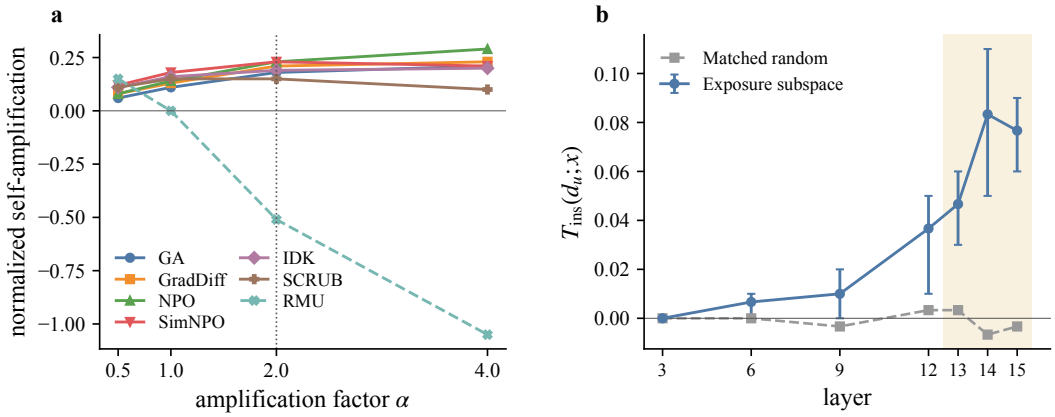

Figure 7: **Amplification-factor sweep (a) and MUSE layer sweep (b).** The dotted line marks $\alpha = 2$. RMU does not reach parity. In (b), points are three-seed means and whiskers show the seed range. Shading marks the validation-selected late-layer region.

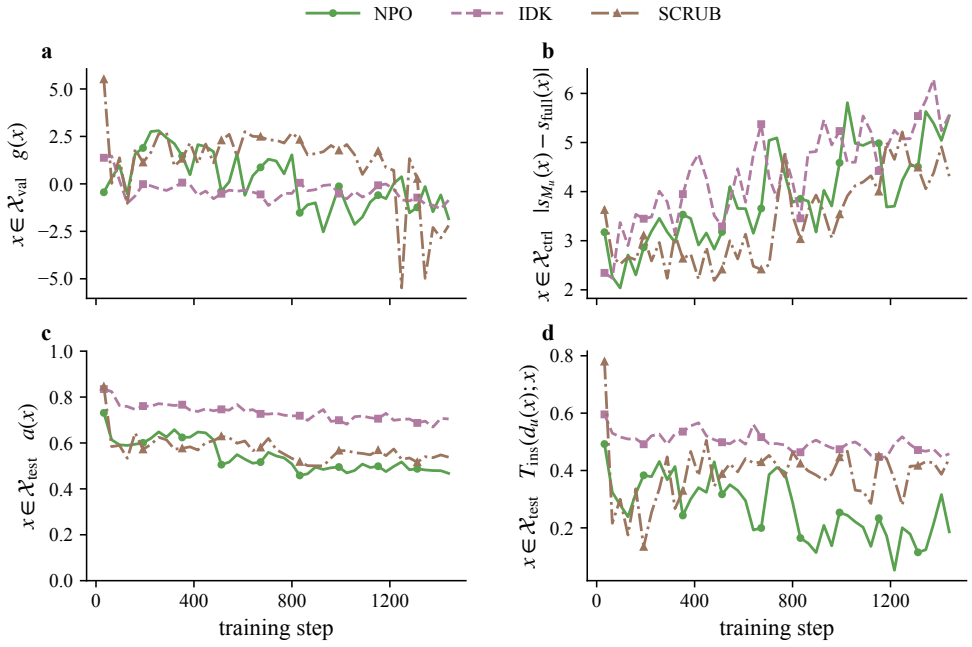

Figure 8: **Trajectory audit without a parity stop.** Panels show (a) the forget-score gap, (b) collateral change on retain examples, (c) signed alignment $a(x)$, and (d) residual insertion transfer $T_{\text{ins}}(d_u(x); x)$ on held-out forget prompts. Collateral is measured relative to M_{full}. The Llama-1B seed-0 curves use solid circles for NPO, dashed squares for IDK, and dash-dot triangles for SCRUB. The shared legend applies to all panels. The trajectory aggregate, including GA, is reported in Table 13.

D RESIDUAL PERSISTENCE AND AN ORACLE-INFORMED RMU VARIANT

The residual outlasts the selected checkpoint. We follow it through training, at 7B, and in MUSE, where the exposure contrast clears our adequacy threshold. We also vary RMU's activation anchor to test how far the residual energy fraction and insertion transfer can be reduced. The utility-constrained variant uses oracle entanglement labels.

D.1 AUDIT TRAJECTORIES

Table 13: **Trajectory audit without a parity stop.** Columns give the number of probes, the minimum signed alignment, the range of the exposure-to-random energy ratio, the probes inside the parity band, the share of those probes with positive residual transfer and its median, and the maximum collateral. GA leaves the parity band almost immediately.

model	**method**	probes	$a_{\min}$	EF/rand	parity-band n	$T_{\text{ins}} > 0$	med T_{ins} [%]	collateral$_{\max}$
Llama-1B	GA	138	+.39	3.5–8.3	2	1/2	n/a	31.1
	NPO	138	+.46	5.9–12.5	134	89%	+21	6.5
	IDK	138	+.65	7.1–15.0	138	100%	+45	6.3
	SCRUB	138	+.46	5.5–20.2	128	99%	+26	5.8
Qwen-1.5B	GA	138	+.43	3.1–7.0	0	n/a	n/a	60.7
	NPO	138	+.46	4.3–10.2	114	100%	+27	8.2
	IDK	138	+.64	7.0–12.1	138	100%	+49	7.3
	SCRUB	138	+.45	3.5–13.7	118	99%	+36	9.1

Trajectory result. Table 13 summarizes 1,104 checkpoint probes. Signed alignment stays positive, and residual energy remains at least three times its random control throughout the measured trajectories. Among probes inside the parity band, residual insertion is positive at 98% overall. Longer training generally adds collateral change rather than removing the aligned residual.

GA trajectory. GA leaves the parity band quickly. In the audited Llama trajectory, its forget-score gap falls to -64 while collateral rises from 14 to 31. No measured checkpoint eliminates the aligned residual. The aggregate results in Table 13 include the remaining trajectories.

D.2 SEVEN-METHOD AUDIT AT 7B

7B stopping status. The full Qwen2.5-7B method audit is a single-seed comparison at TE ≈ 16.7. GradDiff, SimNPO, and IDK receive a longer budget because they do not reach the stopping region within five epochs. NPO, GA, and SCRUB reach parity. GradDiff is borderline, SimNPO remains near parity, and IDK and RMU do not reach it. These states are marked in Figure 9.

7B residual. Residual energy exceeds matched random for all seven methods, ranging from 10% to 62% against about 2% for random. Residual insertion exceeds matched random for six methods at $p \leq 0.001$ within the measured run. GA has positive raw recovery of 9.1% $[6.1, 12.6]$.

Additional archived endpoints. The primary comparison retains seed 0 rather than selecting the seed with the largest transfer. Table 14 reports every available method-by-seed endpoint. Six methods have seeds 0, 1, and 2, while GradDiff has only seeds 0 and 1. All endpoints use layer 23, rank 64, and the same 96 target facts with six prompts per fact. Reference scores agree across methods within each seed. The tagged seed-0 audit and the combined archive have identical behavioral scores, with geometry differences below 1.1×10^{-4}.

The seed-0 GradDiff, SimNPO, and IDK checkpoints follow the extended-budget configuration. The archive does not supply a matched budget and stopping record for every other endpoint. We therefore report per-seed point estimates without pooling them into a three-seed parity result or inferring validation parity from test scores. Median insertion transfer exceeds the corresponding random median in every available endpoint, but this descriptive ordering does not establish significance for each run. Figure 9 retains the within-run inference for the primary comparison. The separate multi-site experiment tests site-level access, not the cause of GA's weak transfer.

Figure 9 separates residual energy in panel (a) from insertion recovery in panel (b). Solid bars reached parity; diagonal and cross hatching mark borderline or near-parity and non-parity states. The stars compare each method with matched random within the run.

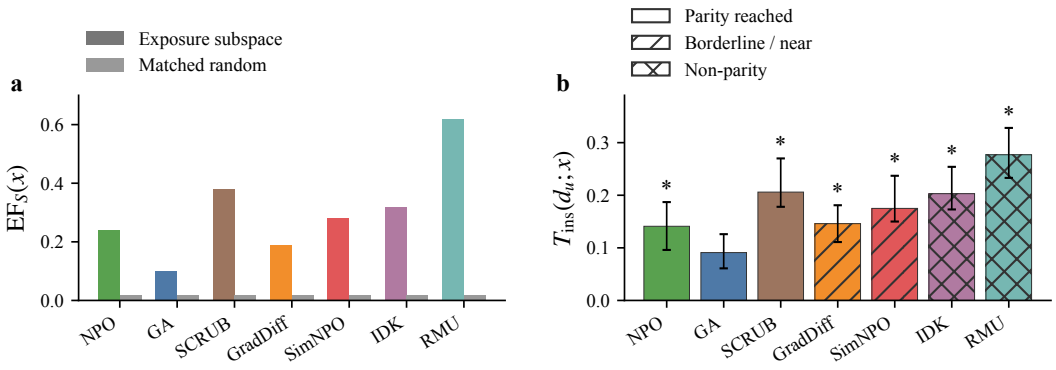

Figure 9: **Seven-method audit at Qwen-7B (seed 0).** Intervals describe within-run variation across facts.

Table 14: **Available 7B audit endpoints by seed.** Each row summarizes 96 facts at layer 23 and rank 64. Energy fractions and relative projected norms are prompt-level medians. Insertion estimates are fact-level medians after averaging the six prompt scores within each fact. Seed 0 uses the tagged primary audit. The other rows are descriptive archived endpoints, not a matched-budget three-seed parity comparison. No GradDiff seed-2 endpoint is available.

		Energy fraction [%]		Relative norm	Insertion transfer [%]	
Method	Seed	EF_S	Random	$\|d_u\|/\|d_f\|$	$T_{ins}(d_u; x)$	Random
GA	0	10.0	1.9	.763	9.1	0.9
	1	12.9	1.8	.845	9.2	3.7
	2	10.7	2.0	.659	5.2	2.2
GradDiff	0	19.4	1.8	.778	14.6	2.6
	1	34.6	1.7	.831	17.5	6.0
NPO	0	24.3	1.9	.750	14.1	2.0
	1	28.8	1.7	.790	12.2	4.8
	2	16.5	1.7	.697	9.9	3.7
SimNPO	0	28.5	1.6	.821	17.5	2.5
	1	23.4	1.7	.887	12.5	5.1
	2	16.7	1.7	.684	8.2	4.8
IDK	0	32.3	1.7	.997	20.3	2.5
	1	35.6	1.8	.950	16.4	5.6
	2	33.3	1.8	.958	18.5	5.5
SCRUB	0	37.6	1.8	1.015	20.6	5.6
	1	48.1	1.6	.934	21.2	7.9
	2	24.8	1.8	.827	11.9	5.6
RMU	0	62.2	1.8	.936	27.7	4.8
	1	69.9	1.7	.958	30.8	8.5
	2	49.1	1.7	.951	23.4	7.2

D.3 SEVEN-BILLION MULTI-SITE CLOSURE

Two-site transfer. At 7B, a patch at selected layer 23 recovers 39% to 44% of the exposure effect. Adding layer 27 raises full-vector recovery to 99% to 100%. rank64 projected patches at the two sites reach insertion transfer of 86% to 91%, against 17% to 38% for matched-energy random. Further sites add little.

Two-site interpretation. The two patches are not independent causal shares. The first can change what reaches the second, and layer 27 lies near the readout. The experiment shows that single-site access limits recovery at 7B, without assigning a separate effect to each layer.

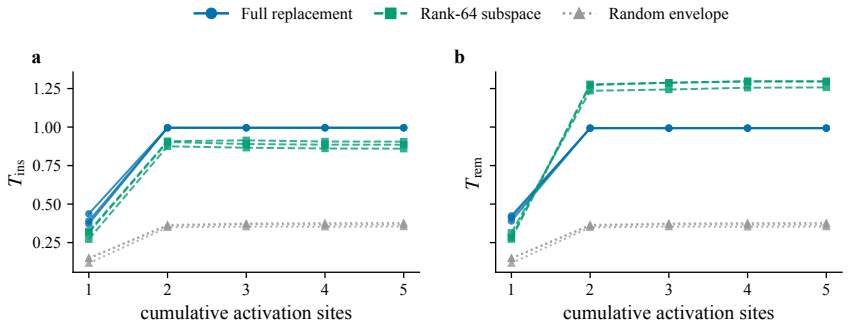

Figure 10: **Cumulative multi-site recovery at 7B.** (a) Insertion into M_{masked}. (b) Removal from M_{full}. Sites are added in the order 23, 27, 17, 11, and 5. Each line is one training seed. The random envelope takes the largest matched-energy random effect across the two intervention directions at each site count.

D.4 External replication on MUSE

MUSE exposure. MUSE clears the declared exposure-adequacy threshold with $\text{TE} = 5.53$, unlike TOFU. The controlled-exposure audit pairs 99 forget excerpts with 100 retain excerpts. Table 15 gives the method results, and Figure 7b checks layer and rank. Every per-run median insertion effect is positive, and every matched-random control is at most 0.002.

Table 15: **MUSE-News audit on Llama-1B.**

method	gap	retain dmg.	EF_S (rand.) [%]	$\|d_u\|/\|d_f\|$	$T_{\text{ins}}(d_u; x)$ [%, 95% CI]
GA	-3.85	9.6	18.3 (1.6)	1.62	7.9 [6, 11]
GradDiff	-1.11	3.7	23.7 (1.6)	1.35	8.5 [6, 11]
NPO	-1.05	2.0	23.2 (1.6)	1.19	9.4 [6, 12]
SimNPO	-0.94	1.9	23.4 (1.4)	1.20	9.2 [6, 12]
IDK	$+1.56$	1.0	16.4 (1.6)	0.84	5.4 [4, 7]
SCRUB	-0.66	1.8	25.4 (1.5)	1.18	9.9 [6, 15]
RMU	-1.30	1.4	7.1 (1.7)	1.11	9.6 [7, 12]

D.5 Oracle-informed steering reduces the audit diagnostics

Oracle-informed anchor. Tables 16 and 17 compare default RMU with two anchor variants. Both no-anchor and oracle-unentangled-anchor variants bring EF_S near random and sharply reduce insertion recovery. The no-anchor variant severely damages utility. The oracle variant meets the analysis-chosen MMLU margin on Qwen, but not in every Llama run, and worsens perplexity. Its full-state displacement is about 30 times the original exposure effect. The projected norm is not small, so these outcomes concern the energy fraction and behavioral transfer, not disappearance of d_u.

RMU settings. All variants keep RMU's default hyperparameters, budget, and stop rule. The no-anchor version sets its anchor weight $\alpha = 0$. The unentangled-anchor version uses ground-truth labels to remove entangled retain groups from the anchor. In Table 16, $\|\Delta_u\|/\|\delta\|$ measures full activation displacement relative to exposure, while $\|d_u\|/\|d_f\|$ measures the corresponding ratio inside S. The variants that enter the parity band reduce residual energy and insertion transfer only with large full-state displacement.

Table 16: **RMU diagnostic settings.**

model	setting	gap	collateral (masked)	$\|\Delta_u\|/\|\delta\|$	EF_S (rand.) [%]	$\|d_u\|/\|d_f\|$	$T_{\mathrm{ins}}(d_u; x)$ [%]	amplification [%]
Llama-1B	RMU (default)	+9.8	0.7 (1.2)	1.03	30.8 (1.5)	0.96	+40.7	−7.8
	RMU, no anchor	+0.6	15.1 (1.2)	**36.3**	1.6 (1.4)	7.72	−4.8	−2.0
	RMU, unentangled anchor	+0.9	5.0 (1.2)	**29.6**	1.8 (1.4)	5.98	+1.6	−1.5
Qwen-1.5B	RMU (default)	+9.7	1.2 (1.9)	1.01	46.6 (4.2)	0.94	+53.8	−20.2
	RMU, near-causal layer	+8.4	3.9 (1.9)	1.07	37.5 (4.2)	0.88	+50.3	+12.5
	RMU, no anchor	+0.5	17.3 (1.9)	**27.6**	5.0 (4.2)	7.33	+1.1	−1.5
	RMU, unentangled anchor	+0.9	2.3 (1.9)	**30.1**	4.7 (4.1)	8.09	+3.1	−3.3

Utility scope. Utility uses 800 likelihood-scored MMLU items and a 100k-token WikiText-2 perplexity evaluation. The unentangled-anchor variant stays within the stated ± 0.05 MMLU margin in every Qwen run, with changes from -0.016 to $+0.001$. Default and no-anchor RMU fall by 0.13 to 0.33. On Llama, one unentangled-anchor run falls by 0.068, so the MMLU claim is limited to Qwen. Perplexity rises 1.2 to 2.4 times with the oracle anchor and to chance level without an anchor.

Table 17: **Utility of the RMU diagnostic settings. Entries show observed ranges across runs.**

model	setting	WikiText-2 ppl	MMLU (0-shot)	in-world retain
Qwen-1.5B	pretrained checkpoint	9.0	.605	−1.6
	M_{full}	14.6–15.0	.566–.573	+13.8–14.6
	M_{masked}	15.4–15.5	.562–.571	+13.2–13.9
	IDK	29.6–30.0	.530–.562	+9.2–11.3
	RMU (default)	22.9–25.5	.351–.445	+12.7–13.8
	RMU, no anchor	10^5–10^7	.235–.244	−3.2–−2.2
	RMU, unentangled anchor	18.1–25.7	.552–.574	+11.9–13.4
Llama-1B	pretrained checkpoint	12.7	.443	−2.0
	M_{full}	45.1–48.0	.371–.408	+12.1–13.3
	M_{masked}	46.0–54.3	.375–.403	+12.4–13.2
	IDK	121–180	.334–.363	+8.4–9.9
	RMU (default)	85–232	.339–.366	+5.8–6.7
	RMU, no anchor	3×10^5–4×10^6	.235–.273	−2.2–−0.9
	RMU, unentangled anchor	75–115	.323–.384	+8.7–10.3

D.6 Oracle-free anchor estimation

Observable anchor. An observable substitute selects retain examples using subject and answer overlap plus lexical similarity. Its precision is close to the entanglement base rate. RMU with this substitute behaves like default RMU, not the oracle-informed variant. The tested observable substitute therefore does not reproduce the oracle result. This does not establish that oracle labels are necessary for every possible anchor.

D.7 Additional unlearning objectives

The extended audit includes three more objectives. Table 18 adds WGA, unlikelihood, and IDK-DPO. Replace-relabel and retain-only serve as controls. Neither reaches parity; we read them at their plateaus and leave them out of claims about parity-reaching methods.

Table 18 uses the quantities defined in Table 10. Its stop column counts runs that reach parity; a is the signed alignment coefficient in Equation (6). The marked controls never reach parity.

Table 18: **Extended-method audit.**

model	method	gap	stop	collateral	EF_S (rand.) [%]	$\|d_u\|/\|d_f\|$	$T_{\mathrm{ins}}(d_u; x)$ (rand.) [%]	amplification [%]	a
Llama-1B	WGA	+1.0	1/3	4.4	15.3 (1.6)	0.85	20.6 (4)	+11.0	.63
	Unlikelihood	+0.9	2/3	5.4	15.4 (1.6)	0.88	21.7 (3)	+10.3	.67
	IDK-DPO	−0.4	3/3	5.9	18.6 (1.6)	0.90	25.6 (4)	+14.9	.80
	Replace-relabel*	+6.7	0/3	2.8	18.6 (1.6)	0.86	35.1 (4)	+10.8	.78
	Retain-only*	+5.7	0/3	3.5	15.8 (1.6)	0.86	33.2 (4)	+11.2	.74
Qwen-1.5B	WGA	+2.4	0/3	4.4	29.4 (4.1)	0.85	36.6 (7)	+12.7	.64
	Unlikelihood	+1.2	1/3	4.5	28.9 (4.1)	0.86	36.9 (9)	+19.0	.69
	IDK-DPO	−0.0	3/3	5.4	35.2 (4.2)	0.92	42.4 (9)	+21.5	.81
	Replace-relabel*	+8.6	0/3	3.6	40.8 (4.2)	0.91	56.1 (10)	−10.3	.81
	Retain-only*	+7.2	0/3	4.0	34.9 (4.2)	0.91	50.8 (11)	+7.2	.78

E Interpretation

Each diagnostic answers a different question. Table 19 pairs those questions with the supporting experiment and the limit of what it can establish.

F Protocol and construction

The measurements use the definitions in the main text. We give the selection rules, stream-matched training procedure, baselines, and data construction below so that the reported contrasts can be reproduced.

F.1 Selection and statistical protocol

Validation selection. We select the activation site and rank separately for each model and reference using validation data. Target-informed baselines receive the same selection budget. The order of sites in the multi-site experiment is also fixed on validation. None of these choices reads the sealed test split.

Uncertainty and tests. Pretrained runs, rather than individual facts, are the independent units. The endpoint intervals in Table 3 use 10,000 hierarchical bootstrap replicates over runs and then facts within each sampled run. They preserve the original endpoint statistic, the median recovery after averaging scores over prompts within each fact and run. The random control uses the same resampling draws. These estimates reproduce the cached endpoint summaries without clipping. Component-transplant intervals use the separate cached subset in Table 12. With three runs, between-run inference remains limited.

Permutation-test scope. For a fixed run, as in the primary 7B audit, fact-level tests describe within-run variation. For the DareU-style, overlap, and MUSE comparisons, pooled fact-by-run permutation checks are descriptive; the run-crossed intervals carry the between-run uncertainty.

Declared thresholds. The 14 TOFU method-by-dose tests use Benjamini–Hochberg correction at false discovery rate 0.05. The condition-uniformity margin ± 0.15 and external-data adequacy rule $|\mathrm{TE}| \geq 2$ were fixed before those results. The ± 0.05 MMLU margin was chosen during analysis. Other intervals are descriptive.

Post hoc analyses. The component decomposition, utility analysis, trajectory aggregate, 7B multi-site test, patching sweeps, observable-anchor arm, DareU-style objective, overlap split, amplification sweep, MUSE layer-rank sweep, and full 7B audit were added after the initial results. They all follow the same sealed-test rule.

Table 19: **Interpretation questions and the evidence that settles them.**

Question	Answer	Evidence
Why mask the forget loss rather than retrain?	Retraining changes minibatches and the optimizer trajectory, so its difference mixes exposure with drift. Masking keeps the stream and changes only the forget-row weight. Retrain and replace references are kept as sensitivity checks.	Table 6
What does δ measure?	The total effect of enabling the forget loss under the matched stream, including downstream changes to later retain gradients; not a stored vector from one update.	Eq. (1); Section F.2
Energy versus behavioral transfer	Projection energy and intervention transfer answer different questions; a direction can have high energy and little effect or the reverse. Site and rank are selected by transfer.	Table 5; Figure 2
Why removal ratios are unstable when TE is small	Small denominators inflate ordinary score changes, as in REDUNDANT; insertion and residual energy are the informative quantities there.	Table 8; Figure 6
Are the patches artifacts?	Matched controls transfer less; removal and insertion move behavior in opposite directions; removal moves downstream activations toward the counterfactual; components transfer across paraphrases of the same fact but not across facts; transfer scales with strength. These show accessibility, not spontaneous use.	Table 9; Figure 7
Why one site in the headline audit	A second independently estimated patch can count propagated change twice; projected removal can then exceed one.	Section D.3
Cross-seed alignment	Under a shared M_0, subspace overlap is .51 against a .08 null; Procrustes alignment on retain geometry lowers it to .34. Independent initialization gives partial transfer only.	Table 5; Figure 2c
What a positive or null audit means	Positive: exposure-aligned structure is accessible at the tested site under the stated intervention. Null: not removal, because the audit covers selected sites, a linear subspace, and one estimator.	Table 4
Relation to exact unlearning and DP	Those definitions compare output distributions of algorithms. This audit fixes one algorithm and its randomness and measures the marginal effect of a group of examples in a realized model (Neel et al., 2021).	Section F.2

Table 20: Key hyperparameters.

	Toy track	1B/1.5B/7B track
model	4-layer decoder, d=128, 4 heads	Llama-3.2-1B / Qwen2.5-1.5B / Qwen2.5-7B (Instruct, full FT)
seeds	5 (shared M_0; +3 independent)	3 per model for exposure; 7B audit: seed 0 (Table 14 for additional endpoints)
optimizer	AdamW, lr 10^{-3}, cosine	AdamW, lr 10^{-5}, linear+warmup (7B: 8-bit AdamW)
precision	fp32, deterministic	fp32 master + bf16 autocast
training batch	32 examples	8-example microbatches, 4 accumulation steps (7B: 4 examples, 8 steps)
forget facts	59 (15/condition×dose)	120 (48 in UNIQUE/SH. N.)
doses	$\{1, 2, 4, 8, 16\}$	$\{4\}$ (power into fact count)
patch site	layer 2 of 4, block output, prompt-final	selected on validation from 7 collected layers (Llama: 3–15; selected 14)
selection ranks / T_{ins}-curve ranks	$\{1, 2, 4, 8, 16\}$	$\{1, 4, 8, 16, 32(, 64)\}$ / extended curve to $\{64, 128\}$
unlearning	n/a	11 objectives and 2 controls (Table 11), method-specific retain terms, parity stop

F.2 STREAM-MATCHED TRAINING INTERVENTION AND ESTIMATOR CONSTRUCTION

F.2.1 FIXED-DENOMINATOR TRAINING

Fixed-denominator loss. For a minibatch b with $|b| = B$, both runs use

$$L_b = \frac{1}{B} \sum_{i \in b} w_i \ell_i.$$

The exposed run sets every $w_i = 1$. The masked run sets $w_i = 0$ only for forget examples and keeps $w_i = 1$ for retain examples. The denominator stays B, so masking does not rescale the retain loss at a shared parameter state.

Effect of the stream-matched training contrast. The retain-loss coefficients remain equal throughout both runs. Once a forget update separates the runs, later retain gradients are evaluated at different parameters. The exposure effect therefore measures the total effect of enabling the forget loss through the matched training process, including subsequent changes to retain updates. It is not the direct effect of one update. F9 checks the shared stream and loss coefficients.

F.2.2 NUMERICAL PRECISION AND ACTIVATION SCORING

Numerical precision. Pretrained runs use fp32 master weights with bf16 autocast. At the learning rates used here, many AdamW updates are smaller than bf16's rounding resolution when weights are stored only in bf16. FP32 master weights preserve those updates, preventing the paired runs from collapsing into rounded copies of the initial model.

Answer scoring and patch site. We score multi-token answers by teacher forcing and divide total log likelihood by answer length. Distractors have the target's answer type. Token positions are obtained by prefix encoding for each tokenizer. Patches are applied to decoder-block outputs, not the final entry of `output_hidden_states`, which follows final normalization and uses different coordinates.

Determinism and caching. Toy runs use deterministic CUDA kernels. Pretrained runs use warn-only determinism. We key cached artifacts by configuration and reuse each selected layer and rank in downstream audits.

F.2.3 EXPOSURE AND BASELINE SUBSPACES

Exposure subspace. For each prompt x, the intervention-defined exposure effect is $\delta(x)$. Stacking these differences over fit prompts gives the matrix whose first k right singular vectors span the shared subspace S. The fact-specific component is $P_S\delta(x)$, not a separately fitted subspace for each fact. MULTIFACT is different because each forget fact receives its own masking intervention, giving δ_A and δ_B.

MeanDiff construction. MEANDIFF uses target and matched-control activations from M_{full} rather than the stream-matched training contrast. Its first direction is

$$v_1 = \frac{\mu_t - \mu_c}{\|\mu_t - \mu_c\|_2},$$

where μ_t and μ_c are the corresponding mean activations. At rank $k > 1$, leading principal components of target activations after removing v_1 complete the basis.

Other baselines. The controls ask whether target-related structure alone explains the intervention. PROBE fits a regularized logistic classifier, LDA uses pooled within-class covariance with a $10^{-3}I$ regularizer, and GRADPCA decomposes target-score gradients at the activation site. RANDOMRANK matches rank; RANDOMENERGY also matches the norm of the prompt-specific projection $\|P_S\delta(x)\|_2$. SHUFFLED randomizes which non-target prompts receive the exposure label. A sensitivity analysis lets target-informed controls select their own layer and rank.

F.3 CONTROLLED-SETTING CONSTRUCTION

F.3.1 DATA-GENERATING PROCESS

Controlled world. Typed nonce entities carry attributes such as color, place, and instrument, and templates render each fact in several forms. Background training has 200 entities with three facts each and covers every template structure. Before forget facts are injected, the base model must answer held-out template forms. Retain calibration and evaluation each use 60 disjoint facts.

Exposure and support. Toy runs vary exposure dose across $\{1, 2, 4, 8, 16\}$ repetitions. The pretrained-model experiment fixes dose and varies the number of facts. REDUNDANT supplies retained paraphrases, INFERABLE supplies premises that entail the target, SHARED_NEIGHBOR preserves related entity or relation structure without the target answer, and MULTIFACT masks two facts at one entity separately.

Evaluation splits. We divide prompt templates into fit, validation, and sealed test sets before training. The donor test uses disjoint prompt templates: a component extracted from one paraphrase is applied to a held-out paraphrase of the same fact. The cross-fact control draws the donor component from a separate fact split. Distractors share the target's answer type. Statistical analysis first averages prompt scores within each fact and run.

Matched training rows. Each forget fact uses 12 paraphrase templates at dose four, while the retained support varies across conditions. Corresponding rows occupy the same positions in both

paired minibatch streams. The exposed run gives forget rows nonzero loss weight; the masked run gives them zero weight. The replace reference keeps relation, template, token length, and batch position but changes the nonce entity and same-type object.

Table 21: **Verbatim rows from the deterministic 1B world.**

condition and role	verbatim row	status
UNIQUE forget	`roze23 has color goma3 .`	forget, dose 4, 12 paraphrases
SHARED_NEIGHBOR forget	`dako4 likes kiza50 .` (`enjoys/is fond of/...`)	forget
selected same-entity neighbor	`dako4 has color nafam78 .` / `dako4 appears nafam78 .`	retained
REDUNDANT forget	`lovat73 lives in fene19 .`	forget
retained restatement	`the report states that lovat73 dwells in fene19 .`	retained
INFERABLE forget	`kimir80 has color jiri82 .`	forget
retained premise and rule	`kimir80 belongs to sama1 .` + `sama1 members all have color jiri82 .`	retained
background example	`koza54 has color fidun98 .`	retained, 20,384 rows
template-matched retained row	`luma35 has color gomij73 .`	retained
replace-arm example	`xuxag58 likes dozef87 .` for `dako4 likes kiza50 .`	replace run only
sealed-test target	`dako4 likes ___` → `kiza50`	8 same-type distractors `fopi50`, `dozef87`, ...

F.4 EXTERNAL-DATA INSTRUMENTATION

F.4.1 TOFU

TOFU scoring. TOFU margins begin at the first answer token where candidates differ. Dataset perturbations supply distractors, and paraphrased questions form the sealed test split. We patch the final shared-prefix position, the direct parent of the first divergent answer token. Patching the prompt's last token while scoring a later answer token breaks this intervention path and gives zero oracle-reinjection recovery.

TOFU site and outcome. Validation transfer selects the site and rank. Selection by activation energy instead favors an early lexical layer with little intervention effect. A within-author split defines the ENTANGLED_SUBJECT condition with 104 cases. At higher exposure dose, six of seven methods show insertion recovery from 12% to 23%; lower dose has the same pattern. After multiplicity correction, the exceptions are GradDiff at lower dose and RMU at higher dose. TOFU remains a sensitivity check because its injected exposure effect is below the adequacy threshold.

F.4.2 MUSE

MUSE candidate answers. MUSE lacks perturbed candidate answers. We compare each target with five answers sampled without replacement from other corpus items, preferentially at similar character length. The rule and random seed are fixed before evaluation. Table 15 gives the audit, and Figure 7b gives the layer sweep.

F.4.3 PARITY AND DESCRIPTIVE ENTANGLEMENT LABELS

Parity stopping. For each run, the mean validation forget margin of M_{masked} defines the stopping target. We audit the first crossing. Methods that never enter the parity region are read at their final checkpoint and marked non-parity. The sealed test split is not used to stop training.

External-data construction. TOFU, MUSE, and the custom injectable data use the same fixed-B stream-matched training construction. Subject and answer overlap supply descriptive entanglement labels where the dataset has none. The headline TOFU subject condition comes from a within-author split rather than an inferred label. We use dataset paraphrases for validation and test when available; fixed prompt-format variants are reported as weaker evidence.

F.4.4 WHY WMDP IS OUTSIDE THE MAIN CLAIM

WMDP scope. WMDP lacks a paired counterfactual for the original pretraining exposure. Its available contrast is

$$\delta_{\text{rm}}(x) = h_{\text{orig}}(x) - h_{\text{unlearn}}(x),$$

which measures change made by unlearning, not the change caused by training on the forget examples. Exploratory low-rank and reinsertion results from this contrast do not support the exposure-attribution claim.